Resources for Text Retrieval and Search Engines MOOC

1. Useful reference textbooks (sorted by time; most recent first)

- *Text Data Analysis and Management: A Practical Introduction to Text Mining and Information Retrieval*, by ChengXiang Zhai and Sean Massung, ACM and Morgan & Claypool Publishers, forthcoming
- *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor, Springer 2011.
- *Modern Information Retrieval: the concepts and technology behind search*, by R. Baeza-Yates and B. Ribeiro-Neto, Addison-Wesley Professional, 2011.
- *Information Retrieval: Implementing and Evaluating Search Engines*, Stefan Buttcher, Charlie Clarke, Gordon Cormack, MIT Press, 2010.
- *Search Engines: Information Retrieval in Practice*, by Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.
- *Introduction to Information Retrieval*, by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.
- *Information Retrieval: Algorithms And Heuristics*, by David A. Grossman, Ophir Frieder), 2nd edition, 2004, Springer.
- *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*, by Richard K. Belew, Cambridge University Press, 2001.
- *Managing Gigabytes:compressing and indexing documents and images*(MG), I. Witten, A. Moffat, and T. Bell (1999), Morgan Kaufmann, 1999.
- *Foundations of Statistical Natural Language Processing* (SNLP), C. Manning and H. Schutze (1999), MIT Press, 1999.
- *Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer* , Gerard Salton, Addison-Wesley Pub (Sd). Aug. 1988.
- *Information Retrieval* C. J. van Rijsbergen, 2nd Edition, Butterworth-Heinemann, Newton, MA, USA, 1979

2. Useful Tutorial Books (sorted by time; most recent first)

- *Information Retrieval Evaluation*, Donna Harman, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers 2011
- *Test Collection Based Evaluation of Information Retrieval Systems, Mark Sanderson*, Foundations and Trends in Information Retrieval, 4(4): 247-375 (2010)
- *Methods for Evaluating Interactive Information Retrieval Systems with Users*, Diane Kelly, Foundations and Trends in Information Retrieval, 3(1-2): 1-224 (2009)
- *Statistical Language Models for Information Retrieval* , by ChengXiang Zhai, Morgan & Claypool Publishers, 2008.

3. Related Courses

- INST 734 Information Retrieval Systems (**an online course with lots of useful stuff!**), University of Maryland
- CS446 Search Engines , University of Massachusetts, Amherst.
- 11-741 Information Retrieval , Carnegie Mellon University.
- CS410 Text Information Systems, University of Illinois at Urbana-Champaign
- CS276/LING 286 Information Retrieval and Web Search, Stanford University.
- **More to be added later**

4. Related MOOCs: Under Construction

5. Suggested Readings: Under Construction

6. Additional Resources

Introduction to IR Research (by ChengXiang Zhai)

IR Resources by the Stanford NLP group

Please send any suggestions for additional useful resources to ChengXiang Zhai at czhai AT illinois.edu.

MORGAN & CLAYPOOL PUBLISHERS

Information Retrieval Evaluation

Donna Harman

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES

Gary Marchionini, *Series Editor*

Information Retrieval Evaluation

Synthesis Lectures on Information Concepts, Retrieval, and Services

Synthesis Lectures on Information Concepts, Retrieval, and Services is edited by Gary Marchionini of the University of North Carolina. The series will publish 50- to 100-page publications on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. The scope will largely follow the purview of premier information and computer science conferences, such as ASIST, ACM SIGIR, ACM/IEEE JCDL, and ACM CIKM. Potential topics include, but not are limited to: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Estimating the Query Difficulty for Information Retrieval
David Carmel, Elad Yom-Tov
2010

iRODS Primer: Integrated Rule-Oriented Data System
Arcot Rajasekar, Reagan Moore, Chien-Yi Hou, Christopher A. Lee, Richard Marciano, Antoine de Torcy, Michael Wan, Wayne Schroeder, Sheau-Yen Chen, Lucas Gilbert, Paul Tooby, Bing Zhu
2010

Collaborative Web Search: Who, What, Where, When, and Why
Meredith Ringel Morris, Jaime Teevan
2009

Multimedia Information Retrieval
Stefan Rüger
2009

Online Multiplayer Games
William Sims Bainbridge
2009

Information Architecture: The Design and Integration of Information Spaces
Wei Ding, Xia Lin
2009

Reading and Writing the Electronic Book
Catherine C. Marshall
2009

Hypermedia Genes: An Evolutionary Perspective on Concepts, Models, and Architectures
Nuno M. Guimarães, Luís M. Carrico
2009

Understanding User-Web Interactions via Web Analytics
Bernard J. (Jim) Jansen
2009

XML Retrieval
Mounia Lalmas
2009

Faceted Search
Daniel Tunkelang
2009

Introduction to Webometrics: Quantitative Web Research for the Social Sciences
Michael Thelwall
2009

Information Retrieval Evaluation

Donna Harman

www.morganclaypool.com

ISBN: 9781598299717 paperback
ISBN: 9781598299724 ebook

DOI 10.2200/S00368ED1V01Y201105ICR019

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES

Lecture #19
Series Editor: Gary Marchionini, *University of North Carolina, Chapel Hill*
Series ISSN
Synthesis Lectures on Information Concepts, Retrieval, and Services
Print 1947-945X Electronic 1947-9468

Information Retrieval Evaluation

Donna Harman
National Institute of Standards and Technology

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #19

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Evaluation has always played a major role in information retrieval, with the early pioneers such as Cyril Cleverdon and Gerard Salton laying the foundations for most of the evaluation methodologies in use today. The retrieval community has been extremely fortunate to have such a well-grounded evaluation paradigm during a period when most of the human language technologies were just developing. This lecture has the goal of explaining where these evaluation methodologies came from and how they have continued to adapt to the vastly changed environment in the search engine world today. The lecture starts with a discussion of the early evaluation of information retrieval systems, starting with the Cranfield testing in the early 1960s, continuing with the Lancaster "user" study for MEDLARS, and presenting the various test collection investigations by the SMART project and by groups in Britain.

The emphasis in this chapter is on the how and the why of the various methodologies developed. The second chapter covers the more recent "batch" evaluations, examining the methodologies used in the various open evaluation campaigns such as TREC, NTCIR (emphasis on Asian languages), CLEF (emphasis on European languages), INEX (emphasis on semi-structured data), etc. Here again the focus is on the how and why, and in particular on the evolving of the older evaluation methodologies to handle new information access techniques. This includes how the test collection techniques were modified and how the metrics were changed to better reflect operational environments. The final chapters look at evaluation issues in user studies–the interactive part of information retrieval, including a look at the search log studies mainly done by the commercial search engines. Here the goal is to show, via case studies, how the high-level issues of experimental design affect the final evaluations.

KEYWORDS

evaluation, test collections, information retrieval, Cranfield paradigm, TREC

Contents

Acknowledgments

No book is written by one person, and I am grateful to all the writers that came before me for their meticulous attention to detail and careful documentation of their methodologies and results. I am also very grateful for the many discussions with members of the information retrieval community over the years as we jointly debated evaluation. The Retrieval Group at NIST (especially Ian Soboroff and Paul Over) and Chris Buckley patiently listened to me and then told me where I was wrong, and for this I am very beholden for getting the details correct in this lecture. Emily Morse of NIST did a preliminary review of Chapter 3 and assured me that I had properly documented an area that is not my strength.

A very special thanks to Stephen Robertson for several critical items. He helped to gather the older material from the British Library and got permission for me to scan this for public consumption. Having access to all these older reports, plus some that Doug Oard was able to borrow from the University of Maryland gave me the material I needed. Stephen was also a preliminary reviewer of Chapter 1 on the early history and I am grateful for his helpful comments.

A second very very special thanks to Ellen Voorhees, not only for many good discussions, but also for an extremely careful review of Chapter 2, making sure that I did get it right and that the (often) subtle issues in TREC were properly recorded.

And finally I am very grateful to the reviewers of the entire lecture, Mark Smucker and John Tait, who not only caught many little errors, but highlighted areas that needed more work or better explanations. I appreciate the time they took to read it, to write helpful comments, and to help me think further about some important issues.

Donna Harman
May 2011

CHAPTER 1

Introduction and Early History

1.1 INTRODUCTION

Information retrieval systems can be evaluated for many reasons, such as which search engine to use personally, what commercial system to buy, or how to improve the user-friendliness of a system interface. This lecture is not about any of these types of evaluation, but rather about the measurement of how well a system is doing in retrieval, and about how to develop testing that will enable system researchers to better understand what is happening inside the system. The emphasis in the entire lecture is on making sure that the methodologies used for testing actually measure what the researchers *think* is being measured, and that any biases in evaluation can be recognized.

The first chapter discusses the early evaluation of information retrieval systems, starting with the Cranfield testing in the early 1960s, continuing with the Lancaster "user" study for MEDLARS, and presenting the various test collection investigations by the SMART project and by groups in Britain. The emphasis in this chapter is on the how and the why of the various methodologies developed. One component of the methodologies is the measures of effectiveness, i.e., the metrics that were used, and whereas there is some discussion of these metrics, readers are generally referred to more complete presentations elsewhere. It should be noted that many of the older references for this chapter are now available online. Readers particularly interested in the early history should also see [134, 147, 167].

The second chapter covers the more recent "batch" evaluations, examining the methodologies used in the various open evaluation campaigns such as TREC, NTCIR (emphasis on Asian languages), CLEF (emphasis on European languages), INEX (emphasis on semi-structured data), etc. Here again the focus is on the how and why, and in particular on the evolving of the older evaluation methodologies to handle new information access techniques. This includes how the test collection techniques were modified and how the metrics were changed to better reflect operational environments. The chapter also contains some advice on how to build a test collection. Most of the references from this chapter are also available online.

The third chapter looks at evaluation issues in user studies–the interactive part of information retrieval. The chapter starts with a short review of evaluation in early user studies, such as those of indexers and search intermediaries, followed by a discussion of the evaluation experiences in the interactive track in TREC. The final section is a look at recent user studies, including the search log studies mainly done by the commercial search engines. Here the goal is to show, via case studies, how the high-level issues of experimental design affect the final evaluations.

The fourth and final chapter presents some thoughts on how to actually do an experiment, pulling together some of the ideas from earlier chapters. Additionally there is discussion of some very recent issues in evaluation, both in methodology and in metrics, and a look ahead at some future challenges.

1.2 THE CRANFIELD TESTS

The Cranfield paradigm (developed by Cyril Cleverdon) is often cited as a "standard" for information retrieval evaluation. But what exactly is this paradigm, where did it come from, and what are the critical components of it? (To read Cleverdon's own account of the Cranfield tests and their background, see his acceptance speech for the 1991 SIGIR award [53].)

There were actually two separate tests conducted by Cleverdon, Librarian of the College of Aeronautics, Cranfield, England and his staff. The first, Cranfield I running from 1958 to 1962 [48, 49], was specifically designed to test four manual indexing (classification) methods. It is hard today to imagine the information access methods that existed at that time–text was not available electronically, and information could only be found by "word-of-mouth" or specialist librarians, who mainly used manually produced, massive indexes to publications. Examples of these systems still exist today, such as the Medical Subject Index, or the Engineering Index.

There had been a huge increase in the volume of scientific papers after World War II and scientists were forced to rely on these indexes to keep current. But these indexes were very expensive to create and there was much contention as to which type of indexing system to use. An editorial [1] called for serious evaluation of these various competing indexing (classification) systems: "Cautious and searching evaluation of all experimental results is essential in rating the efficiency of documentation systems. May the age old controversies that arose from the conventional concepts of classification not be reborn in the mechanized searching systems of the future."

Cleverdon took this editorial as encouragement, and after presenting a paper in Detroit in June 1955, was invited to submit a proposal to the National Science Foundation to create such an evaluation. This proposal was funded and work started in April of 1958. The proposal summarized the various factors that needed to be considered, including the number and type of documents which were to be indexed, the indexing systems, the indexer's subject knowledge of the documents and familiarity with the indexing system, and the type of question to be searched. It also proposed to examine the overall efficiency, such as the time cost to prepare the index and to locate required information, and the probability of producing the required answer while minimizing the irrelevant answers ("noise").

The 18,000 papers and reports to be indexed were from the field of aerodynamics (obviously readily available). The four indexing systems tested were an alphabetical subject catalogue, a faceted classification scheme, the Universal Decimal Classification and the Uniterm system of co-ordinate indexing, representing the dominant types of manual indexing schemes in vogue. There were 3 indexers with different levels of experience in the subject matter. Experience using the different indexing systems was controlled by careful rotation among the various systems throughout the

testing (similar to the Latin Square designs used today). Indexing was done in rotation of blocks of 100 documents, with one system used as the primary index (to be created in an average of 16 minutes per document) and then the indexing for the other three systems finished in much shorter times (the time allowed for indexing was an additional control). It took two years of work to finish this indexing project, with many difficulties encountered along the way (such as major indexer fatigue).

Whereas this first stage of Cranfield 1 was working within a somewhat familiar paradigm, there was little to guide decisions for the second stage (the searching). One previous test, known as the ASTIA-Uniterm test, had been done in 1953, but never fully documented [76]. In this case there had been two teams, each using one of two indexing methods for 15,000 documents. About 93 real user questions were then searched by each team using their index. The end result was that there was no agreement between the teams as to the relevance of the documents and each team generated their own report based on their results!

Cleverdon wanted to avoid this relevance trap, but also wanted to make sure the results would pass significance testing. He estimated that as many as 1,600 questions would be needed, with full searching then done on the four different indexes. This seemed impossible and therefore he decided on what is known today as *known-item searching,* i.e., finding the one document (which he called the "source document") that was guaranteed to be relevant for a given question.

It was critical that the evaluation mirror a true operational setting, in particular that of a researcher searching for a given document in the 18,000 indexed documents. Cleverdon carefully constructed the questions by asking authors of documents in his indexed collection to select some of their documents and then "frame a question that could be satisfactorily answered by that document". In all he received 1500 such questions, which were then subsetted into random batches for various stages of testing. As a final check, he submitted the first batch of 400 questions to a panel who verified that these were indeed typical user questions (only one question was discarded).

The searching process required using each index to manually search for the documents, recording the search time and the success (or failure) of the search. The results–the search failed an average of 35% of the time, with no significant differences among the indexing systems. All of the failures were due to "human indexing error", which did not significantly differ across the indexers.

Whereas the results seemed inconclusive on the surface, Cleverdon was able to discover the real problem simply because of the huge amount of data that was examined in the failure analysis. The issue was not the specific indexing system used, but rather the actual content descriptors that were used for each document. Were the descriptors multiple terms or single terms and how many descriptors were included? More descriptors (exhaustive indexing) lead to better recall (generally), but at the expense of precision (called "relevance" up to 1964). What about weighting the indexing terms, or what he called syntactic indexing involving a method of showing relationships between terms? The problem of how to select content descriptors for indexing, and an increasing interest in evaluation issues, led Cleverdon to continue his investigations in the Cranfield 2 (1962-1966) project [51, 52].

What was learned in Cranfield 1 about the evaluation methodology? Because of the multiple partitionings of the documents for indexing and searching, and the fact that results were similar across these batches, Cleverdon realized that far fewer documents were needed for testing. This was verified by a small experiment by Cleverdon and Jean Aitchinson at Western Reserve University [3] where only 1000 documents were used in similar testing. The creation of the test questions by the use of source documents seemed to work well, avoiding the need for relevance judging (and its inherent instability).

Cleverdon applied these ideas in his design for Cranfield 2, deciding to aim for about 1200 documents and 300 questions. He felt that it was critical to first build the test collection (documents, questions, and relevance judgments), and then do the experiments on the indexing and searching. Additionally it was critical to carefully model his users in building this test collection. The documents needed to be ones they would naturally search, the questions needed to reflect ones they might ask, and the relevance judgments needed to mirror the type of judgments researchers would make for documents they examined in the search process.

He again used the source document method to gather questions, but modified it in order to locate all the relevant documents for a given question rather than just the one source document. The titles of 271 papers published in 1962 on the subject of high speed aerodynamics and the theory of aircraft structures were sent to their authors, along with a listing of up to 10 papers that were cited by these papers. The following instructions were also sent to these authors.

1. State the basic problem, in the form of a search question, which was the reason for the research being undertaken leading to the paper, and also give not more than three supplemental questions that arose in the course of the work, and which were, or might have been, put to an information service.

2. Assess the relevance of each of the submitted list of papers which had been cited as references, in relation to each of the questions given. The assessment is to be based on the following scale of five definitions:

   (a) References which are a complete answer to the question.

   (b) References of a high degree of relevance, the lack of which either would have made the research impractical or would have resulted in a considerable amount of extra work.

   (c) References which were useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.

   (d) References of minimal interest, for example, those that have been included from an historical viewpoint

   (e) References of no interest

There were 173 useful forms returned, with an average of 3.5 questions per form. The document collection was built by merging the 173 source documents with their cited documents (those

that had been previously sent to the authors for judgments), and 209 similar documents, for a total of 1400 documents.

The next stage involved getting relevance assessments for all of the 1400 documents. A smaller set of 361 questions were selected based on their grammatical correctness and on a minimal number of known relevant from the authors. Five graduate students spent the summer of 1963 making preliminary (and liberal) judgments for these 361 questions against all 1400 documents. These judgments were then conveyed to the question authors for a final decision, based on the five graded levels of judging used previously. Judgments were returned for 279 of the questions, although for various reasons usually only 221 of them were used in testing (compound questions were removed for example).

At this point, the test collection was built and experimentation could begin. Note that the goal of Cranfield 2 was to examine in more depth the various properties of index methodologies. Rather than selecting specific indexing systems as in Cranfield 1, Cleverdon and his team wanted to build a "complete" set of index variations for each document and then perform manual experiments using these variations, in this case 33 different index types (see Figure 1.1 for the complete set). The primary performance scores, however, would come from the results of searching using these various index types.

The experimental setup was as follows: each experiment was defined by a series of rules that specified which of the many possible combinations of variables would be used. The searchers then manually followed these rules using coded cards stored in what was known as the "Beehive". So for example, an experiment could involve one specific type of index and a series of *precision device* "runs" using different levels of co-ordination (the Boolean "anding" of terms or concepts during the search process) on a per question basis. For "run 1" of that experiment, each question was indexed by the specific index type being investigated, all index terms from that question were "anded", and the documents meeting that criterion were manually retrieved. This resulted in a single score. The experiment continued with "run 2" which had one less index term used, and so on until only one index term was used for searching.

Cranfield 2 did not emphasize the searching; in general only different co-ordination levels were used. The main focus was the indexing, which was done manually using three basic types of indexing languages: single terms, simple concepts, and controlled terms. On top of this basic structure there were what Cleverdon called *recall devices*, such as the use of the (stemmed) *word forms*, and the use of synonyms and/or hierarchies from a thesaurus. There were also precision devices such as weighting in addition to the co-ordination level searching.

The documents were first indexed at the simple concept level, i.e., "terms which in isolation are weak and virtually useless as retrieval handles were given the necessary context; such terms as 'high', 'number', 'coefficient', etc.". These simple concepts were manually assigned weights based on their importance to document: 9/10 for the main general theme of the document, 7/8 for a major subsidiary theme, and 5/6 for a minor subsidiary theme. The simple concepts could then be broken into their single terms, with weights assigned to these terms based on the concept weight

ORDER	NORMALISED RECALL			INDEXING LANGUAGE
1	65.82	I-3	Single terms.	Word forms
2	65.23	I-2	Single terms.	Synonyms
3	65.00	I-1	Single terms.	Natural Language
4	64.47	I-6	Single terms.	Synonyms, word forms, quasi-synonyms
5	64.41	I-8	Single terms.	Hierarchy second stage
6	64.05	I-7	Single terms.	Hierarchy first stage
7=	63.05	I-5	Single terms.	Synonyms. Quasy-synonyms
7=	63.05	II-11	Simple concepts.	Hierarchical and alphabetical selection
9	62.88	II-10	Simple concepts.	Alphabetical second second stage selection
10=	61.76	III-1	Controlled terms.	Basic terms
10=	61.76	III-2	Controlled terms.	Narrower terms
12	61.17	I-9	Single terms.	Hierarchy third stage
13	60.94	IV-3	Abstracts.	Natural language
14	60.82	IV-4	Abstracts.	Word forms
15	60.11	III-3	Controlled terms.	Broader terms
16	59.76	IV-2	Titles.	Word forms
17	59.70	III-4	Controlled terms.	Related terms
18	59.58	III-5	Controlled terms.	Narrower and broader and terms
19	59.17	III-6	Controlled terms.	Narrower, broader and related terms
20	58.94	IV-1	Titles.	Natural language
21	57.41	II-15	Simple concepts.	Complete combination
22	57.11	II-9	Simple concepts.	Alphabetical first stage selection
23	55.88	II-13	Simple concepts.	Complete species and superordinate
24	55.76	II-8	Somple concepts.	Hierarchical selection
25	55.41	II-12	Simple concepts.	Complete species
26	55.05	II-5	Simple concepts.	Selected species and superordinate
27	53.88	II-7	Simple concepts.	Selected coordinate and collateral
28	53.52	II-3	Simple concepts.	Selected species
29	52.47	II-14	Simple concepts.	Complete collateral
30	52.05	II-4	Simple concepts.	Superordinate
31	51.82	II-6	Simple concepts.	Selected coordinate
32	47.41	II-2	Simple concepts.	Synonyms
33	44.64	II-1	Simple concepts.	Natural language

FIGURE 8.1 T ORDER OF EFFECTIVENESS BASED ON NORMALISED RECALL FOR 33 CRANFIELD INDEX LANGUAGES (AVERAGE OF NUMBERS)

Figure 1.1: The 33 variations of indexing used in Cranfield 2.

and the indexer's view of the term's concreteness and potency. The controlled terms were created by translating the simple concepts into the vocabulary of the Thesaurus of Engineering Terms of the Engineers Joint Council.

Chapter 5 of [52] gives the details of this indexing, including tables showing that there were 3094 unique single terms in the 1400 documents, with an average postings per document of 31.3 single terms (with any weights), 25.2 single terms with weights 7/10 and 12.9 single terms with weights 9/10, giving three levels of exhaustivity of indexing. The manual document indexing was done on the full documents, but as a contrast, and as a way of creating two more levels of exhaustivity, the titles only and the titles plus the abstracts were "automatically" indexed. The details of this are sketchy in the report; however, it is likely that any of the terms that had been declared single terms (3094 of them) and were contained in the abstracts/titles were considered to the automatic indexes of these, including multiple occurrences of the same term. This gave an average of 7 single terms to the titles and 60 single terms to the abstracts plus titles. Figure 1.1 gives the 33 types of index schemes, including the performance ranking of the various schemes using types of scoring that will be described next.

There had been a lot of discussion previously about metrics, centering around the well-known categories shown in Table 1.1. Cleverdon decided to use the "Recall Ratio" defined as $a/(a+c)$, and the "Precision Ratio" $a/(a+b)$. These had been used by Perry [128] and called the Recall Factor and the Pertinency Factor, respectively. Other names previously used for the recall ratio were the sensitivity or the hit ratio, with the precision ratio known as the relevance ratio, the pertinency factor or the acceptance rate (see [147] for more on the history of metrics). Cleverdon liked both the simplicity of recall and precision and the fact that they directly described a user's experience and therefore chose these over more complex formulas such as those suggested by Bourne, Farradane, Vickery, etc.

It should be noted that for each of the 221 questions the recall and precision ratios measured a *single* point for each run in an experiment. Using the example experiment described earlier, each of the co-ordinate levels would generate a single recall and a precision point, e.g., co-ordinating 5 terms yields 28% recall at 29% precision, 4 terms gives 40% recall at 12% precision, and using only one term gives 95% recall at 1% precision. These could be plotted on a recall/precision curve looking much like today's curves, but with each point representing a single experiment as opposed to one curve for each experiment.

There were also issues about how to average these points across the full set of questions. The Cranfield experiments usually worked with the grand total figures of the relevant and retrieved across all of the questions, i.e., sum up the total number of relevant retrieved and the total number of documents retrieved for all questions and then divide by the number of questions. This today is called the micro-averaging method and was the simplest to calculate (remember they were not using computers). Cleverdon was aware of the problems with this method; in that, questions with many relevant documents skewed the results, and therefore he did some experimentation with per question ratios (known as macro-averaging).

Table 1.1: Possible categories of documents in searching.

	Relevant	Non-relevant	
Retrieved	a	b	a + b
Not Retrieved	c	d	c + d
	a + c	b + d	a + b + c + d = N

It soon became apparent that there was too much to do; several subsets of the collection were then used. In particular, the Cranfield 200 was created using 42 questions on aerodynamics, along with 198 of their relevant documents (but not the source documents for these questions). This then created a new problem because experiments done on different subsets could not be directly compared; there were radically different ratios of relevant/non-relevant documents for the 42 questions in the 200 vs. the 1400. The "generality" measure, the ratio of number of relevant documents to the total number of documents in the collection, was defined as $(a + c)/(a + b + c + d)$ and used, along with the "fallout" ratio $b/(b + d)$ which measured the experiments ability to avoid retrieving non-relevant documents.

So what were the results of this huge set of experiments? Figure 1.1 copied from Figure 8.1T in [51], lists the order of effectiveness of 33 different "index languages". The scoring is based on a simulated ranking method (see Chapter 5 in [51] for details) coming from the SMART project and using the normalized recall metric (see Section 1.4). The top seven indexing languages used only single terms, with the very best results found using the word forms (stems) of these single terms. Cleverdon summarized his reaction to these results on the first page of his conclusions [51].

"Quite the most astonishing and seemingly inexplicable conclusion that arises from the project is that the single term index languages are superior to any other type. ....This conclusion is so controversial and so unexpected that it is bound to throw considerable doubt on the methods which have been used to obtain these results, and our first reaction was to doubt the evidence. A complete recheck has failed to reveal any discrepancies, and unless one is prepared to say that the whole test conception is so much at fault that the results are completely distorted, then there is no other course except to attempt to explain the results which seem to offend against every canon on which we were trained as librarians."

Of course there was a great furor from the community and arguments over the Cranfield methodology were fierce [84, 130, 171]. These mostly centered on the use of source documents to generate questions (as opposed to real questions) and on the definitions of relevancy. Whereas some of these came from community rejection of the experimental conclusions, many were reasonable objections for the Cranfield paradigm (although the general consensus was that the experimental results were valid).

There were two very important outcomes from Cranfield 2 for the field of information retrieval. First it had been shown *conclusively* that using the actual terms in a document, as opposed to any type of pre-determined combination or alteration of these terms (such as the simple concepts), resulted in the best searching performance. Whereas the single terms at the top seven ranks were the result

of manual indexing (specific terms being picked), even the "automatic" indexing of abstracts and titles at rank 13 beat out the simple concepts. This result was not only astonishing to the library community, but gave the new computer science community, such as Gerard Salton and the SMART project, the justification to move from experiments in indexing to more detailed experiments in searching.

The second important outcome was the Cranfield paradigm for evaluation. Today this is generally taken to mean the use of a static test collection of documents, questions, and relevance judgments, often with standard recall and precision metrics. But there are two other subtle components that Cleverdon was insistent on. The first was the careful modeling of the task being tested by the test collection. So his collection of scientific documents, his selection of "users" (via the source document method) that would heavily use this collection, and his careful definition of relevance based on how these *particular* users might judge documents were critical pieces of the Cranfield paradigm. The other component was his strict separation of the building of the test collection from the experimentation itself. This was done to avoid problems in earlier experiments, and one can only wonder if his results would have been so clear if he had not followed this principle. Both of these latter components are largely forgotten today, but need to be further examined in light of the various current projects based on Cranfield paradigm.

1.3 THE MEDLARS EVALUATION

It is interesting to contrast the Cranfield work with the other major study that took place during 1966 and 1967–the MEDLARS (Medical Literature Analysis and Retrieval System) evaluation [115]. By way of background, there were approximately 700,000 indexed citations online for medicine by 1966, with 2400 scientific journals being indexed using a huge (7000 categories at that time) controlled vocabulary thesaurus (the Medical Subject Headings, or MeSH). The citations were growing at around 200,000 per year, and the National Library of Medicine offered a search service using search intermediaries for these citations.

The Library asked F.W. Lancaster to do an evaluation of this search service, with the following goals: to study the user requirements, to find out how well the service was meeting those requirements, to identify factors affecting performance, and to suggest improvements. The evaluation was also required to create a test collection, with documents, requests/questions, indexing, search formulations and relevance assessments, and Cleverdon was an advisor.

The specific factors to be measured were the coverage of MEDLARS, the recall and precision of the searches, the response times, the format of the results, and the amount of effort needed by users. It should be noted that these factors depended on many of the variables that Cleverdon had already been investigating, such as inter-indexer performance, the necessary level of exhaustivity of indexing, and the adequacy of the indexing language (the MeSH thesaurus). But because this was an operational test, there was specific interest in the users, such as what were their requirements with respect to recall and precision, what were the best modes of interaction between the users and search intermediaries, and what was the effect of the response times?

This emphasis on the users meant that the user "population" had to be picked carefully. Rather than using a random sample of requests, the study identified a number of specific user groups who agreed to cooperate during the evaluation program. These were groups that were likely to submit enough requests during the testing period, to put in the types of requests that would be representative of the whole user population, to come from different types of organizations, and to have close enough interaction with the search services so that interactions could be carefully studied. Twenty-one groups were selected, including seven academic organizations, five U.S. government health organizations, two pharmaceutical companies, five clinical organizations such as hospitals, and two U.S. regulatory agencies.

These groups agreed to participate and 410 requests were collected between August 1966 and July 1967. Note that the individual requesters (users) did not know about the testing until their requests were submitted, at which time they were asked if they were willing to be part of this test. So the requests were real information needs. These requests were then searched in a normal manner (although the searchers did know these were test requests) and the citations that were found were returned to the user (copies of the entire documents).

In addition to this normal procedure, these users were asked to fill in two types of forms. The first type asked for "relevance" assessment of a subset of the documents that had been returned. Between 25 and 30 documents were in this subset, randomly selected from the full set of documents that had been returned to the user. For each document the user was asked first if they knew of this document's existence, and then to pick one of three relevance categories: "of major value to me in relation to my information need", "of minor value", and "of no value". They were also asked to explain why.

These assessment forms allowed the computation of a precision ratio, in which the number of documents marked relevant was divided by the number in the judged subset (not the number in the full MEDLARS collection). Because these documents had been picked at random from the full returned list, it was assumed that this precision ratio could be extrapolated as the precision for this request if the whole list had been examined. There was also a novelty ratio, calculated as the percentage of the relevant returned that had NOT been previously known to the user.

But this did not address recall, and it would have been impossible to judge all of the MEDLARS citations for each request. The recall was estimated by building what was known as the "recall" base. Each of the users was asked to fill in a second type of form after they submitted their request. In this form they were asked to list all known relevant documents that had been published since July 1963 (when the MEDLARS citation service had started), and to list relevancy (presumably major or minor) for each of those documents. A quick glance at the full results shown in Appendix 4 of [115] shows that most forms listed well less that 10 known documents, with a likely median of around 5 (it is not clear from the report what this number actually is). For about 80% of the searches, additional documents were found using manual searching (usually by NLM staff) using non-NLM tools to avoid test bias. These additional documents were also submitted to the user for relevance assessment (mixed in with the other citations that had been returned). The recall ratio could then be

calculated based on the number of relevant (either known beforehand or marked relevant) divided by the number of documents in the recall base for that request.

In the end, 303 requests were used for testing, with the results showing that the MEDLARS system was operating on average at about 58% recall and 50% precision. These averages cover a huge variation in performance across the requests. Figure 1.2 reproduces page 129 in the Lancaster report and shows a scatterplot of the results, broken down by the number of documents in the recall base. The results look almost random, due to the huge variation in performance across the requests. (The three curves show what Lancaster called the performance guarantees, i.e., what MEDLARS could guarantee their users: a 90% guarantee of performance for curve A, 80% for curve B, and 75% for curve C.)

The rest of his report contains detailed failure analysis for these requests. For each request Lancaster looked at the relevant documents that had been missed by the system (recall failures), and the documents that had been found by the system but judged non-relevant (the precision failures). This involved looking at the full text of each document, the indexing record for that document, user request statement, the manual search formulation, and the reasons the user had given for his judgments. This time-consuming examination led to detailed conclusions for each request, and his final chapter attempted to generalize based on these conclusions.

One generalization involved the performance levels, which were lower than expected. MEDLARS was retrieving an average of 175 citations per search at 58% recall and 50% precision. To operate at an average recall of 85-90% recall (and 20-25% precision), Lancaster estimated that 500 to 600 citations would have to be found and examined by the users. He polled a small sample of users and found that 5 out of those 8 users were happy with less than the maximum recall; his recommendation to MEDLARS was therefore to allow users to specify a high recall search rather than to try to boost their recall by always getting more citations.

The Lancaster study was a test of an operational system, with the goals of finding the problems in that one particular system. So whereas a test collection was built, and recall/precision was the metric for performance, the emphasis was on understanding the nature of the problems rather than experimenting with various parameters. His definition of relevance was more similar to utility, i.e., was this document of major or minor *value* to the user. Lancaster's method of measuring recall within a huge document base was unique; essentially a variation of today's known item retrieval methods. His careful selection of appropriate user groups, and getting "natural" requests from the users was critical to the study in terms of its results being accepted. And finally the methods of failure analysis, and the tremendous attention to detail in these analyses is a lesson for today's researchers.

1.4 THE SMART SYSTEM AND EARLY TEST COLLECTIONS

Gerard Salton started the SMART system at Harvard University in 1961, moving with it to Cornell University in 1965 [116, 146]. His initial interests were in the indexing structures used by manual indexers, but H.P. Luhn's suggestions [119] that simply using the words of a document for indexing might work was intriguing. The Harvard SMART framework was built to allow insertion of different

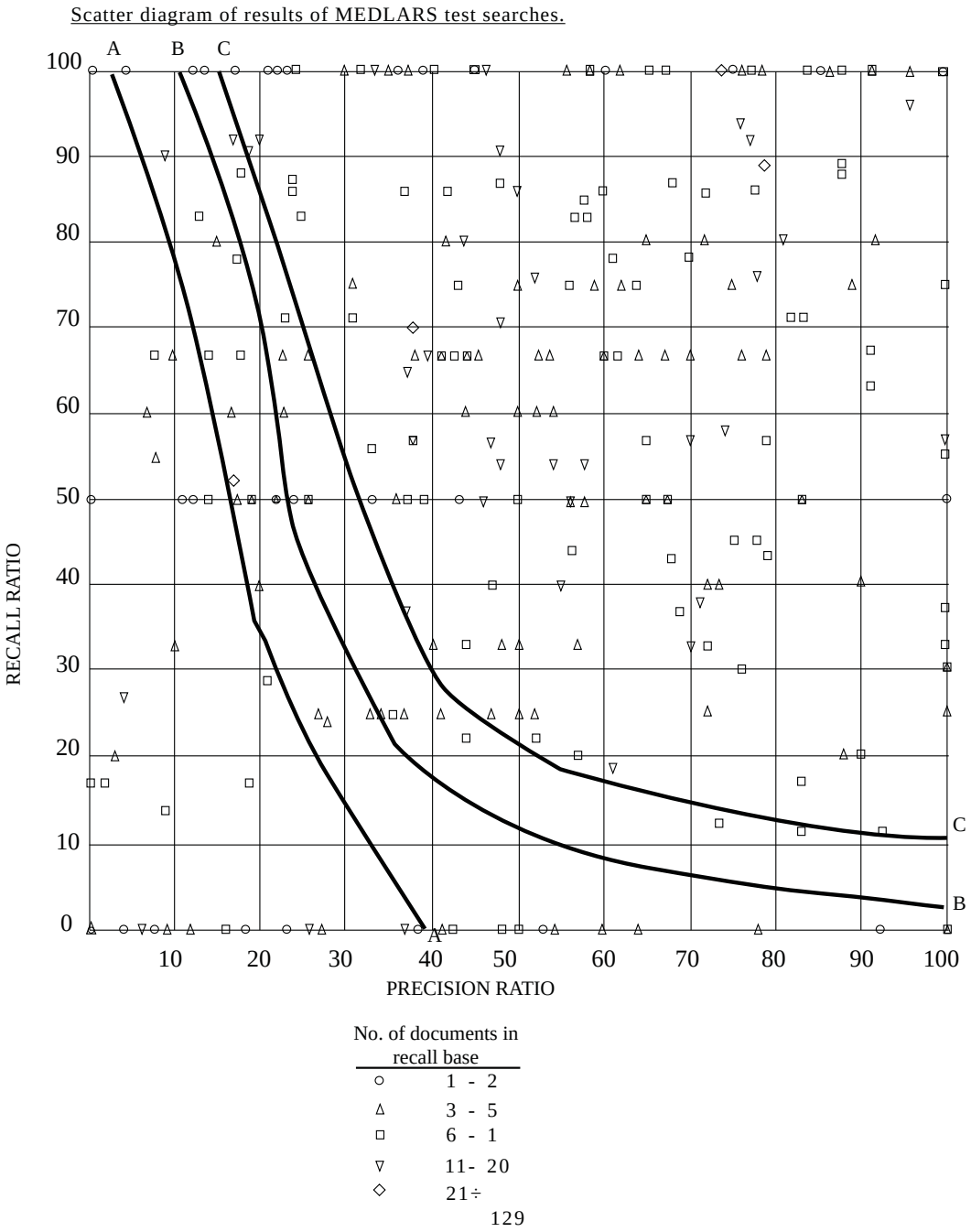


Figure 1.2: Scatter diagram of MEDLARS test searches.

software modules for experimentation with various indexing methods, such as citation indexing, thesaurus coding and the use of simple words. The SMART project continued at Cornell until Salton's death in 1995, producing enormous amounts of research that is documented in journals and proceedings and in the 22 "ISR" reports to NSF. A discussion of the research is beyond the scope of this lecture, but Salton's contributions to evaluation, especially in metrics and new test collections, are critical to the continuing saga of evaluation in information retrieval.

A major SMART evaluation contribution was in metrics, first started at Harvard by Joe Rocchio and continued at Cornell by Mike Keen (borrowed from the Cranfield project in 1966 and 1967). It should be remembered that the Cranfield and MEDLARS experiments were all working with single recall/precision points, e.g., each experiment in Cranfield or each request in Medlars generated a single recall/precision number. But the SMART experiments produced *ranked* or semi-ranked lists of documents, making it impossible to use the single-point recall/precision metrics. Salton and Rocchio proposed two sets of metrics called *rank recall* and *log precision* and *normalized recall* and *normalized precision* [102], both of which measured the differences between the actual ranked positions of the relevant documents and their "ideal" positions. Keen [107, 146] discussed and used the normalized ones for experiments, with the following definitions.

$$\text{Normalized recall} = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)}$$

$$\text{Normalized precision} = 1 - \frac{\sum_{i=1}^{n} \log r_i - \sum_{i=1}^{n} \log i}{\frac{\log N!}{(N-n)!n!}}$$

n = number of relevant documents,
N = number of documents in collection,
r_i = rank of ith relevant document,
i = ideal rank position for the ith relevant item.

Keen also worked on further methodology for the recall/precision curves since he agreed with Cleverdon and Lancaster that it was important to provide separate recall and precision performances. The ranked results led naturally to recall/precision curves, but with problems as to where to make the actual measurements, i.e., the *cutoff points*. So for example one could plot the actual recall and precision for a set of queries at a fixed set of document cutoffs, e.g., 1,2,5,10,15 …documents retrieved; Keen called this the "Pseudo-Cranfield" method. The problem with this method is that queries had different numbers of relevant documents so these individual results did not average properly. Alternatively one could pick a fixed recall, e.g., 10%, 20%, … 100% and plot the precision at this point. However few queries have actual precision measurements at these exact recall points so interpolation was required, especially in the days of small collections (with small numbers of relevant documents). There were several types of interpolations proposed, including using the precision of the document with the highest precision at that recall point, the one with the lowest precision at that point, etc. The decision was to use the highest precision, the "Semi-Cranfield" method illustrated in the top graph in Figure 1.3. This figure (copied from Figure 17 in [107]) also shows the averages

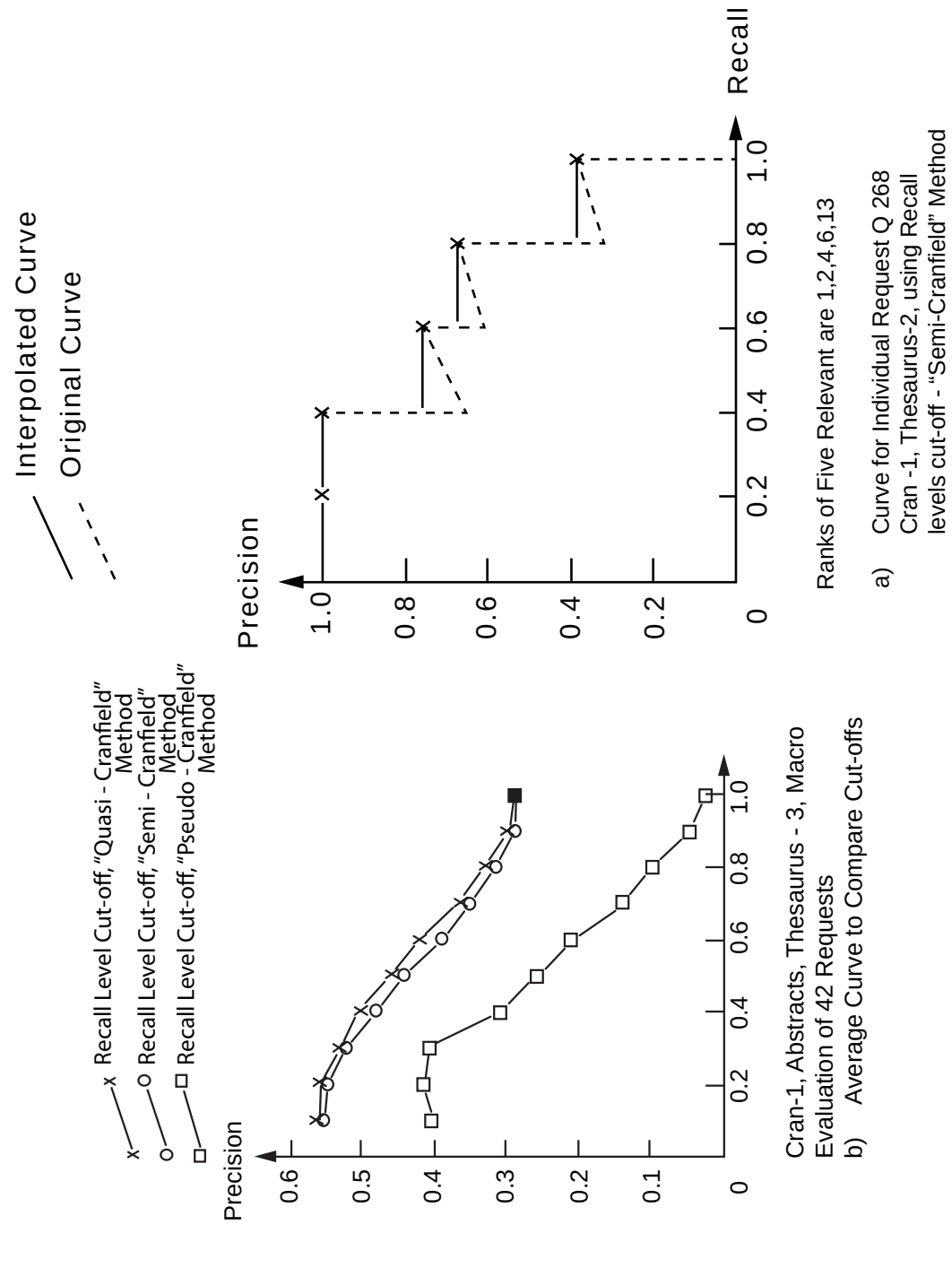


Figure 1.3: Illustrations of different methods of recall/precision interpolation

using three cutoff/interpolation methods in the bottom graph. It can be seen that the results from the document level method are very different from those using the recall-level methods. The final method that was used, called then the "Semi-Cranfield" method, is a variation of the quasi method that deals more appropriately with the frequent situation of two relevant documents adjacent in the ranking. This is the interpolated method used today, where the interpolated precision at a given recall-level is the maximum precision obtained at any recall point greater than or equal to that recall-level. Note that the performance differences for the various recall-level methods were slight even in those days of tiny collections.

There was continued discussion at Cornell in the mid 1960s about the micro vs. macro averaging issue (generally resolved in favor of using macro averaging now that a computer could do the calculations). Additionally there was work [39, 138, 139] with the generality measure (the percentage of documents in the collection that were relevant) since SMART had three versions of the Cranfield collection. There was the full collection of 1400 documents (Cran-1), the subcollection of 200 built by Cleverdon (Cran-2), and a second subcollection of 425 documents built at Cornell. A run on 1400 abstracts cost about $104 in 1965 (about $600 in today's dollar); very few runs were made on the full collection and therefore generality was an issue. As a final note on SMART's contribution to the metrics, it is important to realize that whereas there was discussion of these metrics with other IR researchers, and certainly some disagreements, once these were resolved, the field was able to move forward without many metrics "fights" and this has been a blessing for the whole IR field.

Another critical contribution was the idea of building a framework that allows for "easy" experimentation. The initial one at Harvard was updated to one at Cornell [146, 200] that allowed students to do one-semester projects with minimal recoding simply by changing of the parameter settings. This idea seems obvious today, but the framework and the availability of several test collections made information retrieval experimentation a routine process at a time when experimentation in most human language technology fields was still a difficult endeavor.

The final contribution of the SMART group to evaluation was a very large set of new test collections. All of these collections were built according to the Cranfield paradigm, with each collection designed and built using rigorous specifications. Usually these specifications were dictated by a specific goal for experimentation. Tables 1.2 and 1.3 lists the collections used by SMART, with all of these built in-house except those marked with an asterisk. The dates are approximate, based mostly on when they were first used in an experiment, and the various counts, etc. are taken from different sources so there will be small discrepancies. The number of queries shown is the number of queries with relevant documents, with the number of relevant being an average per query. The lengths of the documents and the queries are (mostly) after stopword removal and stemming, and any large inconsistencies are likely to come from counting unique terms vs. counting all terms, not removing stopwords, or using different stemmers (this was unfortunately seldom documented). What follows is a short description of how each collection was built, both for historical purposes and to illustrate various issues in test collection construction.

The first collection built was the IRE-1 collection of 405 abstracts from computer literature published in the 1959 IRE Transactions of Electronic Computers (and keypunched at Harvard). Salton was interested in how document analysis processes affected results, and had over 35 different document processing options in 1965. For example he used a "null thesaurus" consisting of automatically-created word stems, a manually-constructed thesaurus of 600 concepts in computer literature corresponding to 3000 word stems, hierarchical arrangements of those concepts, statistical phrases, syntactic analysis, etc. [108, 142, 145]. Note that these experiments were investigating similar themes to the Cranfield experiments, with the major exceptions being that these searches were carried out automatically once the document analysis was done, and that the returned results were ranked lists of documents, not single points per experiment. The 17 initial requests (queries) for the IRE-1 collection were built by three project staff, two of which had extensive knowledge of the system (but no knowledge of how it would actually perform). These three people also made the relevance assessments by looking at *all* the abstracts. The collection was extended using similar abstracts and 17 requests written by one non-staff person to become the IRE-3 collection.

The second early collection, the ADI collection, was a set of short papers from the 1963 Annual Meeting of the American Documentation Institute, keypunched at Harvard. The requests this time were created by two technical staff not familiar with the system, and once again these people made relevance judgments against the full collection. The reasons for building the ADI collection were similar to the earlier IRE one, but in addition there were titles, abstracts, and full texts, and this allowed experimentation with document lengths [106, 146]. Note that unlike the Cranfield collection, the first two SMART collections made no effort to get "real" user requests; however, Salton was careful to specify that the requests were not derived from the documents and therefore avoided the issue of the source documents that had caused Cleverdon so much pain [108].

The full Cranfield collection of abstracts was also keypunched at Harvard, allowing comparisons to Cleverdon's results, and the use of a very large collection compared to the two earlier ones. SMART mostly used the Cran-2 collection (200 documents), with stemmers and using the thesaurus previously built at Cranfield. As mentioned before, the cost of the runs allowed few runs with the full 1400 collection; eventually in 1969/1970 a "slimmed down" version of Cranfield with 424 documents and 155 queries was built by first eliminating the source documents, then the documents that were not relevant to any query, and finally creatively finding ways of reducing the documents to 424 while keeping as many queries as possible [116].

Having set up the basic experiments with these three collections, Salton moved on to tackle two thorny issues. The first issue was the continued strong criticisms of these evaluations based on the lack of realistic queries and on the unreliability of the relevance judgments. These issues had plagued the Cranfield experiments and a new test collection, the ISPRA collection, was built specifically to address this problem [117, 146]. The ISPRA documents were 1268 abstracts in the field of documentation and library science from American Documentation and several other journals (all again keypunched at Harvard). The queries this time were very carefully built by 8 library science students or professionals. Each person was asked to construct 6 requests that might actually be asked

Table 1.2: SMART test collections

Collection	Year	#Docs	#Questions	why built
Cran-2*	1964	1398	225	compare indexing methods
Cran-1*	1964	200	42	Cranfield subset of 42 questions
Cran424	1970	424	155	Cornell "reduced" subset
IRE-3	1965	780	34	indexing/dictionary experiments
ADI	1965	82	35	doc length experiments
ISPRA	1967	1268	48	multiple relevance judgements
ISPRA	1968	1095/468	48	CLIR English/German
MED273	1967	273	18	comparison to Lancaster
MED450	1970	450	30	"corrected" Medlars
MEDLARS	1970	1033	30	larger medical collection
OPHTH.	1970	853	30	specific medical domain
TIME	1970	425	83	full text articles
NPL*	1970	11429	93	indexing experiments
INSPEC*	1982	12684	84	indexing, Boolean
CACM	1982	3204	52	additional metadata
ISI/CISI	1982	1460	76	co-citations

Table 1.3: SMART test collections

Collection	Doc.length	Quest.length	#relevant	comments
Cran-2*	53.1	9.2	7.2	#unique terms in abstract
Cran-1*	91	8	4.7	no source docs
Cran424	83.4	8	6.4	fixed # of docs, no source docs
IRE-3	49	12	17.4	
ADI	710/35/7	8	4.9	text, abstracts, title lengths
ISPRA	abstracts	longer	17.8	relevant by author
ISPRA	abstracts	longer	14.2/13.6	relevant English/German
MED273	60?	9.3	4.8/11.1	precision/recall bases
MED450	64.8	10.1	9.2	
MEDLARS	51.6	10.1	23.2	
OPHTH.	60?	10?	30	
TIME	263.8	16.0	8.7	
NPL*	20.0	7.2	22.4	
INSPEC*	32.5	15.6	33.0	
CACM	24.5	10.8	15.3	
ISI/CISI	46.5	28.3	49.8	

by library science students. There was a detailed set of instructions containing criteria such as "use from 50 to 100 words and up to 3 sentences", "do not submit queries corresponding to the contents of a specific document", "do not rephrase specific document contents", etc. The relevance criteria was also strict: an abstract was considered relevant only "if it is directly stated in the abstract as printed, or can be directly deduced from the printed abstract, that the document contains information on the topic asked for in the query". Relevance judgments were made by the original query author *and* by one of the other professionals. The average agreement between these two was found to be only 30%, however similar to later experiments in TREC, it was shown that the performance ranking of the various processing methods, such as words vs. stems and thesaurus vs. stems *did not change* depending on the relevance judgment set that was used [117, 146].

The ISPRA collection was also used to test cross-language retrieval [137, 146]. The English abstracts were 1095 taken from the initial ISPRA collection; additionally there were 468 German abstracts for the collection. The same 48 queries were used, with translations to German done by a native German speaker; the relevance assessments against the German abstracts were done by a different native German speaker.

Salton's second issue was the need to show that the SMART system, e.g., a ranking system using the naturally-occurring terms in the documents and queries, was equivalent to the manually-indexed Boolean retrieval systems [116]. Proving this required showing that the SMART system could scale to operational-size collections; this issue drove many of the clustering theses during those days. But it also caused the creation of the first Medlars collection [143, 146]. The collection was based on the Lancaster study, with 18 of the queries that had been used in that study, along with 273 of the 518 abstracts (the others were not in English or not easily available). The problem came in the comparison of the results. Lancaster had measured his recall and precision by using two *separate* collections: the recall base and the precision base. Additionally he had computed a single recall and precision for each query. The comparison of a ranked list to a single point was one problem, but this was compounded by many other problems and the final solutions (see [143, 146]) show both the creativity and the dedication to proper testing methods possessed by Salton and Keen.

A second Medlars collection was built in 1970 [140, 141] in order to do a better comparison. Twelve more of the original Lancaster requests were added to the original 18 queries, and documents that had been identified as relevant from all of the 30 requests were checked against the Science Citation Index to find a total of 15 additional citations (per request) that were in MEDLARS. These citations were added to the SMART collection and a medical student reviewed them for relevance. This student first tried 6 other MEDLARS requests and had a 69% agreement with the original users and this was felt to be sufficient based on the earlier ISPRA study. The third and final version of the Medlars collection was an expansion by adding a total of 583 relevant documents from the recall bases for the 30 queries.

Around the same time, an ophthalmology collection was built from "scratch" using 30 real user requests and the document set that was retrieved for these requests. The relevance assessments were (mostly) all done by a medical student. The goal of this second medical collection was to allow

comparison between a "general" medical collection (the large Medlars one) and one in a specific domain.

The final test collection of this busy period was a long-planned collection from TIME magazine. Here the goal was to build a collection of longer documents that were not in a scientific domain. The 425 documents were taken from an old print tape [116] and converted to SMART format. The problem was finding a useful query set. After several different tries, 83 questions were taken from four different sources: the New York Times News of the Week news questions (44), the New York Times Current Affairs Test for Colleges (27), the TIME Current Affair Test (7) and Senior Scholastic columns (5). These questions all involved news issues in 1963 and were selected from a total of 182 questions from these sources. Once again, one person made all of the relevance judgments (all questions for all documents).

Two other collections from this period were imported. The NPL collection came from Britain and is discussed in the next section on the British collections. The INSPEC collection came from Syracuse University, where the topics had been built in several versions by information science students as part of a project; the documents were abstracts in electrical engineering from the INSPEC database.

This initial set of test collections was heavily used, particularly the Cranfield424, TIME, and MED450 set (for example, see [144]). There were no more collections built until a very energetic graduate student (Ed Fox) built two much more complex collections in 1982 [69]. These collections, the CACM and ISI ones, were constructed specifically to test the use of multiple concept types as input to the system, such as bibliographic citations/co-citations.

The CACM documents covered all articles in issues of the CACM from 1958 to 1979. The data contained the titles and abstracts for automatic indexing, but also the authors, the computing reviews categories, the citations, the co-citations, and the links (references to or citations between articles). This complex set of information was used in different types of Boolean queries for Fox's thesis. The queries for this collection were true user queries, gathered from faculty, staff and students from various U.S. computer science departments. The relevance judgments were performed by these users, but only on specially selected documents that were likely to be relevant (selected in a manner similar to the pooling operation done in TREC but using only various SMART runs).

The ISI collection was specifically selected because of the co-citation data involved. The 1460 documents were based on a set of information science documents published between 1969 and 1977 that had been collected by Dr. Henry Small. The abstracts, titles and author list were used, along with information allowing co-citation vectors to be produced. The queries were the same queries as the ADI collection (35) plus 41 queries from the ISPRA collection that had relevant documents in this collection and some more modern queries constructed from the abstracts section of the SIGIR Forum. Four members of the SMART project did exhaustive relevance judgments.

Most of these collections became the basis for information retrieval evaluation for over 20 years and illustrate the evolving science of test collection construction. First there was the insistence on realistic test quest ons, and then the fuller understanding of the effects of relevance judging that

came from the ISPRA experiment, which showed that single non-user judges could be employed. The test collections are all excellent examples of the Cranfield paradigm, but in hindsight they exhibit unexpected problems (in addition to being very small). For example the Medlars collections have an extremely high ratio of relevant documents to non-relevant documents because of the particular construction of the document set. Additionally the queries have words that are very specific, making it easy to retrieve the relevant documents at a high rank and therefore "everything works on Medlars". Note that each of these collections were built correctly for their initial goals; the problem is that later experimenters used the collections for *different* goals and were unaware of these biases.

Chris Buckley specifically looked at the 7 still existing collections in his series of weighting runs in the mid 1980s, and reached the following conclusions [116].

- CACM: Avoid document length normalization experiments (strong bias towards long relevant documents)
- NPL: Avoid idf experiments (Queries are short, and all terms are equally good)
- CRAN: Be careful of tf experiments. Title is simply first sentence, so all terms duplicated.
- MEDLARS: Avoid relevance feedback or expansion experiments (documents cluster around queries)
- TIME: Too small, but otherwise good.
- CISI: Poorness of results means random factors can dominate. Bimodal distribution of query lengths a problem.
- INSPEC: Relevance judgements are suspect, but otherwise OK

1.5 THE COMPARATIVE SYSTEMS LABORATORY AT CASE WESTERN UNIVERSITY

In about the same timeframe as the Medlars study and the SMART system, the Comparative Systems Laboratory (CSL) [152] was created by Tefko Saracevic at the Case Western University in Cleveland. This project was notable not only for further innovations in evaluation but also for a "parallel" course [151] linked to the laboratory that allowed students to share in the excitement of live research (similar to students working with the SMART system). The CSL work was done from 1963 to 1968, resulting in a massive report which was summarized in [152]. Saracevic was interested in testing various indexing and (manual) search variables, but additionally had the goal of gaining "a further understanding of the variables and processes operating within retrieval systems and of methodologies needed for experimentation with such systems". His test collection was built in a similar manner to the Cranfield 2 work, starting with a document collection of 600 full-text articles on tropical diseases (which represented about half of the open literature at that time), and questions gathered from 25 user volunteers from 13 organizations who were specialists in their field

and who were asked to "provide questions related to their current research interest". The documents were then indexed (by trained indexers using a manual) with five different indexing languages: telegraphic abstract (index terms assigned without constraints including syntactic roles and links), keywords selected by indexers, keywords selected by a computer program, a metalanguage and the Tropical Disease Bulletin index terms. The indexing was done on titles and abstracts and full-text for the first three languages, resulting in 5 to 8 terms for titles, 23 to 30 terms for abstracts, and 36 to 40 terms for full-text.

There were 124 questions gathered from the users that were then used for searching with five types of question analysis done on each question (unit concepts (A), expansion of A using a thesaurus (B), expansion of A using other tools (C), further expansion of C using thesaurus (D), and E which was a user-verified version of D). Note that these question analysis types were mirroring the types of "natural" steps that a search intermediary might make, as opposed to the more constrained approach to searching that was used in Cranfield. However there was extensive checking of the searches for consistency with over half being re-run with errors fixed. Additionally all of the searches were done in both a narrow method based strictly on the question analysis and a broader search statement looking at a more general subject aspect.

The relevance judgments were created using a type of pooling method, i.e., a *universal set* was created by the union of all sets of outputs from all the indexing strategies the question analysis strategies and all searching strategies. These universal sets were then judged by the 25 users on a three-point scale: relevant ("any document which on the basis of the information it conveys, is considered to be related to your question even if the information is outdated or familiar"), partially-relevant ("any document which on the basis of the information it conveys is considered only somewhat or in some part related to your question or to any part of your question"), and nonrelevant. The judgments made were clearly stringent, with over half of the questions having no relevant documents, and 80% of the remaining questions having only one to five relevant.

The metrics used were *sensitivity* (defined the same as recall) and *specificity* which is the number of nonrelevant documents NOT retrieved divided by the total number of nonrelevant documents in the universal set. Both of these metrics were calculated over all queries, i.e., the micro-averaging method used in Cranfield, along with a combined metric *effectiveness* which was defined as sensitivity plus specificity minus one.

The paper [152] contains a thorough analysis of the indexing schemes, the question-analysis schemes, and the searching strategies, coming to many of the same conclusions as both Cleverdon and Lancaster. This project, however, was the first to employ real user questions and relevance judgments, and also the first to try pooling as a method of lowering the effort for relevance judging. An interesting note on the pooling is that only three of the five indexing schemes were operational in time for the initial pooling (retrieving a total of 1518 answers); however, when the other index files were added, there were 1108 additional answers almost all of which were nonrelevant. Saracevic noted that "this finding is taken to demonstrate that the files responded equally well in retrieving

relevant answers, but that there were large, symmetric differences in the retrieval of non-relevant answers", something that is still seen today in batch evaluations.

1.6 CAMBRIDGE AND THE "IDEAL" TEST COLLECTION

In the mid 1960s Roger Needham (eventually joined by Karen Spärck Jones) was working on a theory of clumping of terms using automatic classification techniques [125]. This was applied to the Cranfield 200 collection with the goal of finding if these classifications were useful for information retrieval [165]. Whereas only the manually indexed version of the collection was used (rather than the actual words in the abstract and text), this work could be nicely compared with both Cleverdon's and Salton's results. However Spärck Jones was not satisfied with simply using the Cranfield collection, but went on to use two other collections, an INSPEC collection of 541 documents (different than the one used by SMART), and a new collection by Mike Keen, the ISILT collection [110]. Unfortunately the results were different for each collection [163] and this led to a serious interest in the properties of test collections.

It also led to a proposal in 1975 by Spärck Jones and Keith van Rijsbergen [166] for the creation of large test collection(s). The proposal had two overarching criteria: first, that it would allow for commonality of testing across retrieval researchers, and second that it would be adequate for many various projects. It should be noted that in 1975 the SMART collections (other than the Cranfield ones) were not only small in size, but lacked manual indexing and other "hooks" for document analysis. The group at Cambridge and other British researchers continued to be more interested in the document analysis part of information retrieval, such as finding better ways of indexing the documents, rather than the searching stage of retrieval.

This is why Spärck Jones had chosen to use the British collections for her work in 1973, and her proposal presented details of British collections as background material and as a basis for the new collection(s). Four of these are minimally described here to show the contrast in goals and construction with the SMART collections, and to motivate some of the issues in the proposal.

The ISILT collection was from Mike Keen, then at the University College of Wales in Aberwystwyth. There were 800 documents from the area of documentation, with 63 real requests and exhaustive relevance judgments by subject experts. The purpose of this collection was to continue to investigate manual indexing using five kinds of manual index terms. Three additional British collections also involved indexing studies. The INSPEC collection [4] used 542 documents in the sciences. The queries were based loosely on SDI profiles (Selective Dissemination of Information, where a library would select a limited set of citations to match an "SDI" profile), with the users asking a question that was within the scope of their SDI profile and then making relevance judgments only on those documents previously deemed relevant to their profile. Once again there were five types of manual indexing created. The UKCIS collection [19] was for Chemical Abstracts, testing the effectiveness of manually indexing titles, titles plus keywords, and titles plus digests. There were many digests/abstracts in several different subsets and 193 queries, again loosely based on SDI profiles and with user relevance judgments. Finally the NPL collection [181] had 11571 documents (which

is why it was heavily used in the United States) from a published scientific abstract journal, with 93 requests based on source abstracts (similar to Cranfield). The documents were indexed semi-automatically using 1000 index terms and the purpose of the collection was to look at association between word pairs and clustering.

Note that all of these collections (with the exception of ISILT) used data from commercial search services. By using this type of data, the collections were guaranteed real user requests and also several types of manual indexing. This fitted in well with the kinds of experiments done in Cambridge, but also may have simply been what was available. The SMART collections were mainly built for the testing of the searching stage of retrieval and were funded by NSF. This allowed more freedom in creating collections, but also mostly negated the use of commercial search services.

Given this background, the proposal for the "ideal" test collection continued primarily in the British tradition of building test collections. The proposal was presented at a workshop and a six-month project at Cambridge worked on further details [164]. The following outline specification for the collection is quoted directly from this report.

> 1) Documents
>
> Size: a main set of 30,000 documents broadly representative of service data bases in size and subject composition.
>
> One or more other sets of 3000 documents complementing the main set in subject, etc. Thus, for example, if the main set was in a scientific area, one other set would in social science.
>
> These sets would cover short time periods and English language material; they would have core characterizations. A random subset of the main set, containing 3000 documents, would be established, with enriched characterizations.
>
> Properties: the main set, and the complementary other sets, would be heterogeneous on identified collection variables such as subject solidity, document type, author type, etc.
>
> The size of the main set should permit the identification of subsets, say containing 3000 documents, which would be homogeneous on such variables.
>
> In addition, one or more other sets would be required for time and language contrasts, and possibly for gross contrasts on other variables, for example covering monographs as opposed to articles. These would have core or enriched characterizations as appropriate or available.
>
> 2) Requests
>
> Size: a primary set of 700-1000 requests would accompany the main set of documents.
>
> Secondary sets of 150-250 requests would accompany other sets of documents.
>
> As the primary sets would be of one form, envisaged as retrospective off-line queries, alternative sets representing different forms, e.g., SDI is one such method, and containing

150-250 requests were proposed for the main set. (At least some overlap of the primary and alternative sets through derivation from common need statements was suggested; this overlap would define a base set of 30 requests). These sets would have core characterizations. A random subset of the primary set, containing 150-250 requests, would be established, with enriched characterizations.

For document sets representing time and language contrasts, subsets of the primary set would be appropriate as requests.

Properties: the primary and alternative sets would be heterogeneous on such variables as topic type, user type, etc.

The size of the primary set should permit the selection of homogeneous subsets of perhaps 150 requests.

The request sets should represent many users, as well as many requests.

3) Relevance judgments

The proposals are not fully worked out and are further developed in (this) report. For reference, the main points were: default judgments by the users of their own search output; exhaustive judgments of the random subset of documents; pooled judgments on variant strategy search output; these would all use abstracts; checking judgments for the base set of requests, e.g., against the texts of the random subset, against another random subset, etc.

The data characterizations are:

a) core: for documents, all regular bibliographic information, abstracts, citations, natural language indexing, controlled language indexing, (using thesaurus terms or subject headings) and high level subject class codes, and an about sentence;

for requests, a verbal need statement, lists of free and controlled terms, and a Boolean specification.

b) enriched: for documents, a more exhaustive indexing, indexing from different sources, indexing by different people, PRECIS, etc; for requests, term weights, indexing by different people, etc.

Note that the variants for pooled judgments would constitute further request formulations.

Queries in the form of source documents should also be obtained.

It was proposed the full "sociological" background information relative to requests should be obtained.

c) relevance judgments; essential information would consist of two relevance grades and also a novelty indication; known relevant documents would be recorded. Judgments by different people would be covered by the basic design.

Appendix 7 of the report provided a theoretical basis for the construction of the relevance judgments. The approach was to create a "pool" of likely relevant documents and then obtain a limited number of relevance judgments for that pool (either by users or non-users). The pool was to be constructed by using different strategies (probably different document and request analysis strategies rather than searching strategies); then the total number of required assessments (number of requests multiplied by number of assessments per request) was to be estimated based on the requirement for significant differences between two potential strategies. A statistician was funded to further investigate this methodology, and a third report [74] provided detailed analysis of the proposed method and other possible methods. Further discussion of this report is beyond the scope of this lecture, but interested readers are strongly encouraged to invest some time here since the pooling issues are still difficult in today's testing methodology.

The 1977 report included information on possible document collections (very few machine-readable databases were available at that time), sources for requests (usually from commercial search services), and how the collections might be managed. This included cost estimates and also results of a survey on British groups that might use this collection. The final sentence of the report is as follows: "Thus if the BLR&DD can satisfy itself that, say 7 good projects will be forthcoming, "ideal" collection version D at £85K or even £100K is a good buy".

The "ideal" collection specifications have been provided here in such detail because it was not funded. The CACM collection built at Cornell had some of these characteristics, and a version of pooling was used in the relevance assessment, but it was not large, and it was built more for combining different search strategies than for document analysis. The large test collections that were dreamed of in 1975 were not built until 1992, and the TREC collections were not built based on these specifications but on other requirements (which are detailed in the next chapter).

A final result of the Cambridge investigation was a book "Information Retrieval Experiment" by Spärck Jones [167]. Published in 1981, it became a major influence in information retrieval evaluation, presenting chapters discussing different aspects of evaluation by most of the leading researchers at that time.

1.7 ADDITIONAL WORK IN METRICS UP TO 1992

The metrics used by Cleverdon and further developed by Salton and Keen for SMART were generally heavily used by all researchers. Several problems existed with these measures; most importantly that there was no easily-interpretable single measure of performance. By the late 1960s the normalized recall and precision had been replaced by several methods of averaging the recall-level values, such as averaging across 3 recall-levels, or 10 recall-levels, or 11 (including the two end points). But two other metrics were developed that have had significant use in the research community.

The first of these was William Cooper's expected search length [55]. His paper argued that not only was there no single measure, but that the various averages did not take into account how many relevant documents the user *actually* wanted. So his idea of the expected search length was to measure the "cost" to the user to find their desired number of relevant documents, which in the

case of a ranked list is the number of *non-relevant* documents that they need to examine in order to satisfy their needs. The expected search length measure has often been used for the cases when only one relevant document, or a single "correct" document is the object of the search (known item searching).

The second metric to combine recall and precision was van Rijsbergen's E measure [180]. The E measure was defined for set-based retrieval, such as the Cranfield or Lancaster experiments, but could be extended to ranked list retrieval by calculating recall and precision at document cutoff levels, such as recall and precision at 20 documents.

$$E = 1 - \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

where α is a variable to control how much emphasis to put on precision versus recall, and P and R are the precision and recall values for the set being examined. Again this measure allows user requirements to be a part of the test by the use of the α parameter. However it works best for set retrieval rather than ranked list retrieval since use of the document level cutoff creates problems with averaging across sets of requests. A similar measure called the F measure (1 - E) is heavily used in classification experiments and natural language experiments where set retrieval is the norm.

The van Rijsbergen book [180]and the papers by Robertson [133] and Sanderson [147] are excellent reviews of the metrics in this period and are strongly recommended reading for both derivations of the metrics and comparisons across the various metrics that were proposed (including many not discussed here). Additionally the various chapters in the Spärck Jones book [167] discuss metrics, including ones by van Rijsbergen and Cooper.

CHAPTER 2

"Batch" Evaluation Since 1992

2.1 INTRODUCTION

In 1992 a new large test collection became available to the information retrieval community. The TREC (Text REtrieval Evaluation Conference) collection and its sister evaluation efforts built on the Cranfield methods, extending the methodology where appropriate. This chapter elaborates on how these new test collections were created, how the various evaluations were designed, and what changes to the Cranfield methods had to be made. The later sections of the chapter summarize TREC's sister evaluations and discuss general lessons that can be drawn from all these evaluations.

2.2 THE TREC EVALUATIONS

The National Institutes of Standards (NIST) was asked in 1991 to build a test collection to evaluate the results of the DARPA (Defense Advanced Research Projects Agency) TIPSTER project [121] . The goal of this project was to significantly improve retrieval from large, real-world data collections, and whereas only four DARPA contractors were involved, the TREC initiative opened the evaluation to the wider information retrieval research community, with 25 additional groups taking part in 1992. TREC has now been running for nearly 20 years. Full coverage of the research is clearly beyond the scope of this lecture and readers are referred to [184] as a general reference and to the full series of online proceedings at `http://trec.nist.gov/` for details relating to each year. However the details of the test collection methodology and the metrics presented in this chapter show the further evolution of evaluation for information retrieval

2.3 THE TREC AD HOC TESTS (1992-1999)

The term "ad hoc" here refers to the classic information retrieval user model used in the Cranfield method where ad hoc requests are searched against a fixed document collection. This section discusses the TREC ad hoc test methodology (test collection and metrics) in detail both because this methodology was extended for later TREC tasks (and other evaluation efforts) and because the ad hoc test collections are still heavily used in research today and it is important to understand how and why they were built.

2.3.1 BUILDING THE AD HOC COLLECTIONS

The TIPSTER/TREC test design was based squarely on the Cranfield paradigm, with a test collection of documents, user requests (called topics in TREC), and relevance judgments. Like Cranfield,

it was important to create all parts of the collection based on a realistic user model, in this case the TIPSTER application. The TIPSTER users were presumed to be intelligence analysts, but could also be other types of users that work with information intensively, such as journalists, medical researchers, or legal staff.

The document collection needed to have a very large number of full-text documents (2 gigabytes of text was generally used each year), which needed to cover different timeframes and subject areas. They also had to be of varied length, writing style, level of editing and vocabulary. Table 2.1 lists the document sources used during the initial eight years of the ad hoc evaluations; these were selected based on availability and suitability to the TIPSTER task. Articles from newspapers and newswires covered all domains and contrasted in their format, style, and level of editing. Documents from *Computer Selects* were from different initial sources, but dealt with the single domain of computer technology. Finally there were documents selected less for their content than for the length of articles: the *Federal Register* ones were especially long, and of non-uniform length, and the DOE abstracts were very short. All documents were converted to a SGML-like format with enough uniformity to allow easy processing by the systems. Note that at 2 gigabtyes these collections were beyond the ability of most research systems in 1992 to handle, mainly because the storage to index them (say a total of 4 gigabytes) cost around $10,000 at that time.

Earlier test collections had typically provided only sentence-length requests; however, the TIPSTER/TREC topics contained multiple fields, including a user need statement that is a clear statement of what criteria make a document relevant. Having these multiple fields allowed for a wide range of query construction methods, and having clear statements about relevancy improved the consistency of relevance judgments. All topics were designed to mimic a real user's need, although the actual topic writers, the topic format and the method of construction evolved over time. The first two TRECs (topics 1-150) involved actual users of a TIPSTER-like search system and had very elaborate topics. By TREC-3 the topics were reduced to three fields and were written by the same group of "stand-in" users who did the relevance assessments. Figure 2.1 shows a sample topic from TREC-3. Each topic contains a number and title, followed by a one-sentence description of the information need. The final section is the narrative section, meant to be a full description of the information need in terms of what separates a relevant document from a nonrelevant document.

The definition of relevance has always been problematic in building information retrieval test collections [26, 36, 56, 85, 103]. The TIPSTER task was defined to be a high-recall task where it was important not to miss information. Therefore the assessors were instructed to judge a document relevant if information from that document would be used in some manner for the writing of a report on the subject of the topic, even if it was just one relevant sentence or if that information had already been seen in another document. This also implies the use of binary relevance judgments; that is, a document either contains useful information and is therefore relevant, or it does not. Documents retrieved for each topic were judged by a single assessor so that all documents screened would reflect the same user's interpretation of topic.

Table 2.1: Document collection statistics.

	Size: MB	# Docs	Median # Words	Mean # Words
Disk 1				
Wall Street Journal 1987–1989	267	98,732	245	434.0
Associated Press 1989	254	84,678	446	473.9
Computer Selects	242	75,180	200	473.0
Federal Register 1989	260	25,960	391	1315.9
abstracts from DOE	184	226,087	111	120.4
Disk 2				
Wall Street Journal 1990–1992	242	74,520	301	508.4
Associated Press 1988	237	79,919	438	468.7
Computer Selects	175	56,920	182	451.9
Federal Register 1988	209	19,860	396	1378.1
Disk 3				
San Jose Mercury News 1991	287	90,257	379	453.0
Associated Press 1990	237	78,321	451	478.4
Computer Selects	345	161,021	122	295.4
U.S. patents, 1993	243	6,711	4445	5391.0
Disk 4				
Financial Times 1991–1994	564	210,158	316	412.7
Federal Register 1994	395	55,630	588	644.7
Congressional Record 1993	235	27,922	288	1373.5
Disk 5				
Foreign Broadcast Information Service	470	130,471	322	543.6
Los Angeles Times 1989-1990	475	131,896	351	526.5

There was the additional requirement that the relevance assessments be as complete as possible. This became a critical piece of both the implementation of TREC and the later analysis of the collections. Three possible methods for finding the relevant documents could have been used. In the first method, full relevance judgments could have been made on over a million documents for each topic, resulting in over 100 million judgments (clearly impossible). The second approach, a true non-

```
<num> Number: 168
<title> Topic: Financing AMTRAK

<desc> Description:
A document will address the role of the Federal Government in
financing the operation of the National Railroad Transportation
Corporation (AMTRAK).

<narr> Narrative:
A relevant document must provide information on the government's
responsibility to make AMTRAK an economically viable entity. It
could also discuss the privatization of AMTRAK as an alternative
to continuing government subsidies. Documents comparing
government subsidies given to air and bus transportation with
those provided to AMTRAK would also be relevant.
```

Figure 2.1: Sample topic statement from TREC-3

biased random sample of the documents, would have been prohibitively expensive for acceptable completeness levels. Therefore a biased sampling method called "pooling" was adopted from the 1977 proposal to the British Library for building an "ideal" test collection [164]. To construct the pools for TREC, the following was done. Given a ranked list of results from a single system, for each topic select the top X ranked documents for input to the pool. Then merge this set with sets from all systems, sort the final list based on the document identifiers, and remove duplicates (identical documents found by multiple systems in the pool). This created the pooled list for each topic that was then judged by the assessors.

2.3.2 ANALYSIS OF THE AD HOC COLLECTIONS

Since the ad hoc evaluations were run for 8 years (see Table 2.2 for details of the eight collections), it was possible to analyze how well the various evaluation decisions were working and to modify them as necessary. This analysis is presented in detail here because these large collections are still in heavy use and it is critical to know their strengths and weaknesses in order to avoid experimental bias. It also provides guidance for some of the types of issues that need to be investigated in building future test collections.

The document selection and semi-standardized formatting worked very well. Groups had little trouble indexing the documents (other than scaling issues), and had no problems with domain or time elements. Analysis showed that by far the largest number of relevant documents came from the document sources covering all domains: *Wall Street Journal* (WSJ) and *Associated Press* (AP). In

Table 2.2: Document and topics sets for the first 8 TRECs.

TREC ad hoc	Document sets	Topic Numbers
TREC-1	disks 1 & 2	51-100
TREC-2	disks 1 &2	101-150
TREC-3	disks 1 & 2	151-200
TREC-4	disks 2 &3	201-250
TREC-5	disks 2 & 4	251-300
TREC-6	disks 4 &5	301-350
TREC-7	disks 4 & 5 (minus Congressional Record)	351-400
TREC-8	disks 4 &5 (minus Congressional Record)	401-450

contrast, the very long *Federal Register* (FR) documents had few relevant documents, but in TREC-2 most retrieval systems had difficulty in screening out these long documents. By TREC-3 this effect had disappeared as most of the systems made major corrections in their term weighting algorithms between TRECs 2 and 3 and thus could cope with any length document.

The ad hoc topics built for TREC underwent major evolution across the first five TRECs. Part of this evolution came as a result of changes in the personnel constructing the topics, but most was the result of deliberate changing of the topic specifications. The elaborate topics in the first two TRECs contained a field with manually-selected keywords (the concepts field) and this was removed in TREC-3 because it was felt that real user questions would not contain this field, and because inclusion of the field discouraged research into techniques for expansion of "too short" user need expressions. The TREC-4 topics were made even shorter, with removal of the title and the narrative field, however this turned out to be too short, especially for groups building manual queries, so the TREC-3 format became standard.

There was also a change in how the topics were constructed. In TREC-3 the assessors brought in "seeds" of topics, i.e., ideas of issues on which to build a topic. These seeds were then expanded by the assessor, based on looking at the items that were retrieved. To avoid this tuning to the data, starting in TREC-4 the assessors were asked to bring in a one-sentence description that was used for the initial searching to estimate the number of relevant documents that are likely to be found. Topics with "reasonable" numbers of relevant documents were then kept for further development into the final TREC ad hoc topics.

Another issue about the topics relates to measuring the difficulty of a given topic. There has been no attempt in TREC to build topics to match any particular characteristics, partly because the emphasis was on real user topics, but also because it is not clear what particular characteristics would be appropriate. A measure called topic "hardness" was defined for each topic as the average over a given set of runs of the precision at R (where R is the number of relevant documents for that topic) *OR* the precision at 100 if there were more than 100 relevant documents. This measure is therefore oriented towards high recall performance and how well systems do at finding all the relevant documents. In TREC-5 an attempt was made to correlate topic characteristics with this

hardness measure, but neither topic length nor the number of relevant documents were found to be correlated [191], and it is still unclear what topic characteristics make a topic harder. Further work on topic characteristics has been carried out in an extended workshop, the Reliable Information Access (RIA) workshop in 2004 [82].

A related issue concerns the "required" number of topics for a test collection, i.e., how many topics are needed in order for the performance averages to be stable, much less show significant differences between systems or techniques. There has always been a huge variability in the performance across topics, as seen in the Lancaster experiments described earlier or in the selection of the large number (221) of requests in the Cranfield collection. TREC was no exception here, with a huge variability in the "hardness" of the topics, in the system performance on each topic, and in the performance of different techniques, such as relevance feedback on each topic. However it is critical that the average performance measure truly reflect differences rather than just random performance points. It was "folklore", at least in the SMART project, that a minimum of 25 topics were needed. Although TREC's 50 topic sets have been shown to produce stable averages [34, 189], the measurement of significant differences is still a problem in information retrieval, with some TREC-specific work starting in TREC-3 [173], and much more work since then (see Chapter 5 in [147]).

The TREC relevance judgments were specifically designed to model users interested in high recall tasks and therefore the more complete the relevance judgments are, the better the test collection models the high-recall needs of these users. Additionally, the more complete the test collection, the more likely that future systems using the collection for evaluation can trust that all/most of the relevant documents in the collection have been identified. Note that the pooling methodology *assumes* that all documents that have not been judged can be considered non-relevant.

A test of the relevance judgment completeness assumption was made using TREC-2 results, and again during the TREC-3 evaluation. In both cases, a second set of 100 documents was examined from each system, using only a sample of topics and systems in TREC-2, and using all topics and systems in TREC-3 [78, 79]. The more complete TREC-3 testing found well less than one new relevant document per run. These levels of completeness are quite acceptable; furthermore the number of new relevant documents found was shown to be more strongly correlated with the original number of relevant documents, i.e., topics with many relevant documents are more likely to have additional ones, than with the number of documents judged.

These findings were independently verified by Justin Zobel at the Royal Melbourne Institute of Technology (RMIT) [204]. Additionally Zobel found that lack of completeness did not bias the results of particular systems and that systems that did not contribute documents to the pool can still be evaluated fairly using the pooled judgments. Since the goal of the TREC collections is to allow comparisons of multiple runs, either across systems or within systems, having the exact number of relevant documents, or having an exact recall number is not as important as knowing that the judgments are complete enough to insure that comparisons of two methods using the test collections will be accurate.

A second issue important to any set of relevance judgments is their consistency, i.e., how stable are the judgments and how does their stability or lack thereof affect comparison of performance of systems using that test collection. Salton investigated this during the ISPRA experiments [117, 146] and Cleverdon also did a small experiment [50] showing minimal effects on system comparison. For TREC each topic was judged by a single assessor to ensure the best consistency of judgment and testing of this consistency was done after TREC-2, and more completely for TREC-4 [78, 80]. All the ad hoc topics had samples rejudged by two additional assessors, with the results being about 72% agreement (using the overlap measure of the intersection over the union) among all three judges, and 88% agreement between the initial judge and either one of the two additional judges. This remarkably high level of agreement is probably due to the similar background and training of the judges, and a general lack of ambiguity in the topics as represented by the narrative section.

Unfortunately, most of this agreement was for the large numbers of documents that were clearly nonrelevant. Whereas less than 3% of the initial nonrelevant documents were marked as relevant by secondary judges, 30% of the documents judged relevant by the initial judge were marked as nonrelevant by both the additional judges. This average hides a high variability across topics; for 12 of the 50 topics the disagreement on relevant documents was higher than 50%.

While some of these disagreements were likely caused by mistakes, most of them were caused by human variation in judgment, often magnified by a mismatch between the topic statement, the task, and the document collection. For example, topic 234 is "What progress has been made in fuel cell technology?". A lenient interpretation might declare relevant most documents that discuss fuel cells. A strict judge could require that relevant documents literally present a progress report on fuel cell technology. Additionally some of the more problematic topics were either very open to different interpretations (topic 245: "What are the trends and developments in retirement communities?") or so badly mismatched to the document collection that the initial assessor made extremely lenient relevance judgments (topic 249: "How has the depletion or destruction of the rain forest effected the worlds weather?").

Note this topic and user variation is very realistic and must be accepted as part of any testing. Users come to retrieval systems with different expectations, and most of these expectations are unstated. If test collections do not reflect this noisy situation, then the systems that are built using these collections to test their algorithms will not work well in operational settings.

A critical question is how all this variation affects system comparisons. Voorhees [185] investigated this by using different subsets of the relevance judgments from TREC-4. As her most stringent test, she used the intersection of the relevant document sets (where all judges had agreed), and the union of these judgements (where any judge had marked a document relevant). She found that although the mean average precision of a given set of system results did change, the changes were highly correlated across systems and the relative ranking of different system runs did not significantly change. Even when the two runs were from the same organization (and therefore are more likely to be similar), the two systems were ranked in the same order by all subsets of relevance judgments. This clearly demonstrates the stability of the TREC ad hoc relevance judgments in the

sense that groups can test two different algorithms and be reasonably assured that results reflect a true difference between those algorithms.

These results were independently verified as a result of the University of Waterloo's work in TREC-6 [57]. Waterloo personnel judged over 13,000 documents for relevance, and these judgments were used by Voorhees in a similar manner as the TREC-4 multiple judgments. Even though there was even less agreement between the NIST assessors and the Waterloo assessors (very different backgrounds and training), the changes in system rankings were still not significant. The one exception to this was the comparison between two same-system runs in which one run had used manual relevance feedback. For this reason, comparison between automatic runs and runs with manual intervention, particularly manual relevance feedback which basically adds a third relevance judge, should be more carefully analyzed as they are the comparisons most likely to be affected by variations in relevance judgments.

2.3.3 THE TREC AD HOC METRICS

TREC followed the Cranfield metrics, essentially using the metrics developed by Mike Keen and Gerard Salton discussed earlier. Starting in TREC-1 Chris Buckley made available the evaluation program used by SMART called trec_eval. Researchers in the field at this time were using various metrics from SMART, but with different implementations and with different choices of which metrics to report. This made comparison across systems difficult and the availability of a standard package at least made for a common implementation of these metrics.

Figure 2.2 shows the set of metrics provided for a run in the TREC-8 ad hoc track. The recall level and document level precision averages across the 50 topics are shown, in addition to a new non-interpolated average precision, defined as "the precision at each relevant document, averaged over all relevant documents for a topic" [33]. The non-interpolated average precision is then averaged over all the topics to produce the "mean average precision" or MAP, which has been used as the main measure in TREC. Other new metrics include the R-Precision which was proposed by Buckley to better measure the high-recall task being modeled in TREC. For more details on these metrics, including a discussion of their relative strengths and weaknesses, see [33, 147]. Note that this result page also includes a histogram showing the results for all of the 50 topics so that groups could easily spot how their systems had performed with respect to the median system performance per topic.

2.4 OTHER TREC RETRIEVAL TASKS

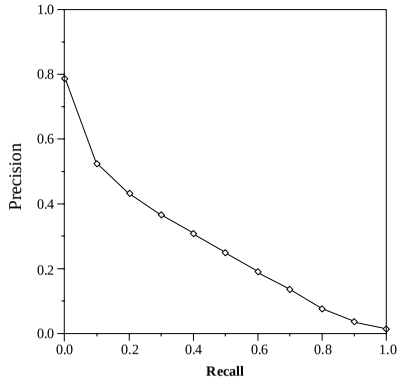
New tasks called tracks were added in TREC-4, and led to the design and building of many specialized test collections. None of these test collections were as extensive as nor as heavily used as the ad hoc collections described earlier, but the necessary changes in the design criteria provide useful case studies in building test collections. Note that these changes were required either because of the specific data characteristics, or because the track research goals dictated modifications to the standard Cranfield implementation. The track descriptions that follow are ordered by the amount

Ad hoc results — SabIR Research/Cornell University

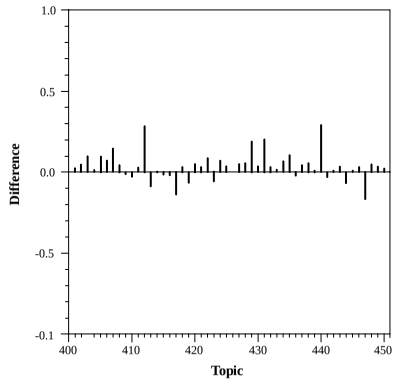
Summary Statistics	
Run Number	Sab8A4
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel-ret:	2986

Recall Level Precision Averages	
Recall	Precision
0.00	0.7860
0.10	0.5229
0.20	0.4324
0.30	0.3644
0.40	0.3084
0.50	0.2498
0.60	0.1912
0.70	0.1360
0.80	0.0776
0.90	0.0362
1.00	0.0133
Average precision over all relevant docs	
non-interpolated	0.2608

Document Level Averages	
	Precision
At 5 docs	0.5200
At 10 docs	0.4800
At 15 docs	0.4413
At 20 docs	0.4090
At 30 docs	0.3733
At 100 docs	0.2384
At 200 docs	0.1702
At 500 docs	0.0985
At 1000 docs	0.0597
R-Precision (precision after R does retrieved (where R is the number of relevant documents))	
Exact	0.3021



Recall-Precision Curve



Difference from Median in Average Precision per Topic

Figure 2.2: Sample evaluation report from TREC-8

of change that was necessary to the TREC ad hoc design, but in all cases there was a user model to guide the task definition (and hence the test collection), and to define the correct metrics.

2.4.1 RETRIEVAL FROM "NOISY" TEXT

Here the user model is still the ad hoc task, but the documents are "noisy" text, such as that produced by OCR scans or speech transcription, and the goal is to understand how retrieval performance is affected by this noise. In TREC-4 the ad hoc topics were used against artificially degraded *Wall Street Journal* data that reflected error rates of 10 and 20% character errors. For TREC-5 [191], a more sophisticated test used actual OCR data from the 1994 *Federal Register*, with comparison of results using the correct copy (electronic version), plus two scanned copies at 5% error rates and 20% error rates. The topics were changed to known item topics, i.e., each topic was created to uniquely retrieve one document, a type of search that is particularly useful for testing in "noisy" data, where a single corrupted term might cause a system to miss that document. The metric used to measure performance for this task was a variation of Cooper's expected search length [55], and was based on the rank where the item was retrieved. In this case however a measure called *mean-reciprocal-rank, MRR* was used to allow proper averaging. MRR is the mean of the reciprocal of the rank where the known item was found, averaged across all the topics, and is mathematically equivalent to the mean average precision when there is only one target document [183].

A second type of noisy text comes from speech recognition systems. Two groups at NIST, the Speech group and the Retrieval group, collaborated to implement test collections for broadcast news for TRECs 6-8 [73]. Note that for speech processing there is no clear definition of a document; documents were defined to be specific stories, usually separated by a change in speaker and/or a change in topic. Additionally the "documents" were provided both as text transcriptions of the data (manually or automatically transcribed) or as recorded waveforms. TREC-6 used a 50 hour set of news; TREC-7 used 87 hours, and there were 557 hours for TREC-8. In addition to the normal information retrieval performance metrics, the speech metric word error rate (WER) was also reported.

In 2001 TREC started a video retrieval track, working with 11 hours of video, and 74 known item searches contributed by the participants. This increased the next year to 40 hours of video from the Open Video Project and the Internet Archive, with 25 specially created topics from NIST. The topics were expressed in a multi-media way, with text supplemented by video clips or bits of speech in order to mimic possible types of requests from likely users. The definition of documents became more fuzzy as the extent of the video needed to respond to the topic had to be measured at shot boundaries. By 2003, the video retrieval task had become much more concerned with the image part of the video and the task was split into its own evaluation, TRECvid (`http://www-nlpir.nist.gov/projects/trecvid/`). For more on video retrieval, see both the TRECvid site and references from the video retrieval community such as [160].

2.4.2 RETRIEVAL OF NON-ENGLISH DOCUMENTS

TREC-3 began work with Spanish and Chinese to investigate how the retrieval techniques would work outside of English. The user model was still the ad hoc model, but with Spanish and Chinese topics and documents respectively. The only change necessary to the ad hoc model was the issue of character sets, including how to deal with the accents in Spanish and the Chinese character sets. One of the important lessons learned in this first evaluation was the importance of using native speakers in both topic generation and relevance judgments to reflect a real user model and to allow speedy relevance judgments!

But what if the user is trying to retrieve documents in a language that is not their native language? The user modeled for cross-language retrieval (CLIR) would be creating topics in their native language, which would then be searched across the set of other languages, since it is assumed that they have more proficiency in reading another language than in creating topics in that language. This implies that it is critical that the topics be created by native speakers in each language to insure that they reflect how a person would actually express their information need in that language.

The major impetus for the first CLIR track (TREC-6) was the creation of a document collection [156] that had "parallel" documents in three languages (French, German, and Italian) from the Swiss newswire *Schweizerische Depeschen Agentur*. English documents from the *AP* newswire in the same year were added, and all documents were put in the SGML-like format with a new field added to indicate the (major) language of that document. The 25 topics used the first year were written in three languages at NIST, however by the second year this became a co-operative effort across groups in Switzerland, Germany and Italy, with 56 topics written in four languages (English, French, German, and Italian) for TRECs 7 and 8 [29].

In order to balance input from the initial topic languages, equal numbers of topics were chosen from each site for the final topic set. The full set of topics were then translated to all four languages. Relevance judgments were made at all four sites for all topics, with each site examining only the pool of documents in their native language. This distributed scenario for building topics and making relevance judgments was a necessary but major departure from the Cranfield model.

Note that the effect of the distributed method of relevance judgments on results is probably small since the major distribution was across languages, not topics. As long as results are compared within the same language, i.e., pairs of results on German documents, and not across languages, i.e., results on English documents vs. German documents, there are no problems. However comparing results across different languages is both comparing across completely different document collections and across two human relevance judges and is not valid.

The European CLIR track moved to Europe after TREC-8 to become a separate evaluation (CLEF: `http://www.clef-campaign.org/`) and information about further work in CLEF is in a later section. TREC continued CLIR for three more years, first with English and Chinese, and finally for two years using English and Arabic.

2.4.3 VERY LARGE CORPUS, WEB RETRIEVAL, AND ENTERPRISE SEARCHING

Another issue in retrieval is the effect of much larger collections of documents, both on the retrieval results and on efficiency. David Hawking and his colleagues at CSIRO put together five large collections to test efficiency and later to test web applications [87]. This was piloted in TREC-6 with 20 gigabtyes of assorted text (called VLC1), but scaled up to 100 gigabytes of web data from the Internet archive (VLC2) the next year [86, 88]. The TREC-7 ad hoc topics were used against VLC2, but with very shallow pools (20 documents deep vs. 100), and whereas this test collection would not meet a completeness test, it was appropriate for normal high-precision web testing. In TREC-8, the "large" web track used the VLC2 data for 10,000 web queries from a search log (only a random sample of 50 were judged). Additionally the VLC2 data was appropriately down-sampled to create a "small" web environment of 2 gigabytes (WT2g) and the TREC-8 ad hoc topics were used, with normal ad hoc relevance assessment done for both the text and web collections. This allowed for "complete" judgment of the smaller web track against the ad hoc topics, but with more realistic web queries being used for the larger web collection [89].

Whereas this testing was initially only for efficiency (and various efficiency metrics were used in all three years), the use of web data as documents was also of interest. Like the speech data, web documents do not have a clear definition: in this case a document was defined as a single web document, including all of its parts but not including other documents linked to it. TREC-9 saw the introduction of the WT10g collection, a 10 gigabyte subset of the VLC2 corpus specifically constructed to mirror the web rather than simply being a large chunk of documents. There was a high degree of inter-server connectivity, along with other desirable corpus properties that were carefully preserved, creating a collection that was a proper testbed for web experiments [15]. Two years later a new collection (gov1) at 18 gigabytes was created from material in the *.gov domain [60].

Other than the vast scaling of the task, the web track up until 2000 (TREC-9) was still much in the Cranfield model. The initial topics were the ad hoc topics because a major goal of these experiments was to see how the scale and the structural information for the web documents affected "normal" ad hoc retrieval (it didn't much). However discussion with various web groups at this point made it very clear that the ad hoc topic style was only a small part of the web activity and if TREC wanted to reflect more realistic user models, the topics needed to change. So the TREC topics in 2000 looked for online services (such as ordering flowers), with short topics reverse-engineered from web query logs. The assessors selected topics from the search log and created a "full" TREC topic with description and narrative that matched *their* interpretation of that topic before examining the document set. The original log entry was included as the title of the new topic (with any original misspellings). This worked well in that the topic was well-enough defined to have consistent relevance judgments, but of a length that matched real web logs.

This reverse-engineering method continued to be used in the web track, but topics evolved into more specialized (and realistic) tasks in 2002, such as finding homepage locations and topic distillation (finding home-pages of relevant sites). The metrics also reflected this realistic model,

with an emphasis on early success: success@1 (proportion of the queries with a good answer at rank 1), success@5 and success@10, and of course mean average precision just for comparison [61]. Starting in 2000 three-level judgments were used, again to reflect the web user model where web sites were likely to be highly relevant, relevant or non-relevant. The three-level judgment evaluation was effective, and in general the NIST assessors liked this mode of judgment because it made the decision process easier (personal communication from Ellen Voorhees). In 2004 there was a second web track called the Terabyte track, which used a much larger web collection, the 426 gigabyte"gov2" collection. This track is further discussed in Section 2.4.5.4.

In 2005 the original web track became the enterprise track, with the emphasis on looking at tasks more specific to intranet searching. Here the user model is someone within an organization searching in-house information. The data for 2005 and 2006 was a crawl of the W3C site, including email discussion lists, web pages and text in various formats. Topics were constructed for three search tasks: a known-item task (email message), an ad hoc search for email on a specific topic, and search for experts in a specific area [59]. Two of these tasks continued in 2006, but similar to earlier web track issues, there was concern that the topics/tasks being modeled were not realistic because they were being built by people outside of the organization. So in 2007 and 2008 the data *and* topics came from within one organization, CSIRO (Australian Commonwealth Scientific and Industrial Research Organization). The data was the CSIRO public web site, and the task that was modeled was that of creating a new "overview" page for a specific topic, which would include key links from the existing CSIRO web site along with key experts [17]. The CSIRO staff generated 50 such topics, with the track participants doing most of the relevance judging. For a discussion of how well this distributed judging worked, see [16].

2.4.4 DOMAIN-SPECIFIC RETRIEVAL TASKS

The tasks described in previous sections have all dealt with domain independent tasks: the data came from general news-type sources or from the web. With the possible exception of the enterprise track, where the user model required task-specific data, the domain of the documents did not influence the task. However some tracks in TREC have worked with domain-specific data, where the tasks and user models were driven by this data. One of the critical issues in dealing with domain-specific data is to determine the true information needs (and appropriate relevance criteria) for tasks within that domain. Note that in a minor way the Cranfield model was domain-specific in that the user model was defined as a scientist or engineer that might be searching in the domain of aerodynamics.

2.4.4.1 Genomics

Bill Hersh from Oregon Health Sciences Institute ran the genomics track for retrieval from medical information starting in TREC 2004. Part of the goal of this track was to see how well the ad hoc methods worked for the medical domain, but an additional part was to develop test collections and methodologies specifically suited to the retrieval needs of the genomics community. The document set for the first two years was 4 million MEDLINE records, about one-third the size of the full

MEDLINE database. Since these are basically abstracts but with many types of added metadata, the last two years of the track used 160,000 full-text papers from 49 genomics-related journals.

Whereas the data here was similar to the news-type sources, the selection of tasks and the building of the topic sets needed to closely mirror the requirements of the genomics community. In the first year, volunteers interviewed 43 genomic researchers, gathering 74 different information needs [93]. These were used to create 50 topics in a standard format similar to the ad hoc model. The task that year was a straight ad hoc task, using pooling and relevance judgments by people with medical backgrounds. The topics the second year were more focused, with six generic topic "templates" used to generate the 50 topics [92]. These templates allowed the topics to concentrate on known types of search problems, such as the role of a specific gene (where different genes could then be inserted in the template).

For the full-text documents for the next two years, the track moved to a passage retrieval task, e.g., find the relevant passages in the documents. Note that passages are difficult to define so this track had to create specific definitions, deal with the difficult relevance judgment issues that go with judging passage retrieval and finally use metrics that are appropriate for passage retrieval [131]. The passages were defined to have a maximum span of one paragraph, but could be much shorter, e.g., the minimum passage that would answer the topic/question. By the final year of the track (2007), a specific passage mean average precision (MAP) measure had been developed that compared character overlap between the submitted passages and those that had been selected as the "gold-standard" by the judges. A pooling mechanism was used with the passages, but then the submitted passages were judged not only for relevancy but their coverage of all the specific aspects of the answer. "To assess relevance, judges were instructed to break down the questions into the required elements (e.g., the biological entities and processes that make up the questions) and isolate the minimum contiguous substring that answered the question [96]." There was also an aspect-level MAP that looked at what percentage of the aspects were covered. The final evaluation used all three MAPs (document-level, passage-level and aspectual-level), making this one of the most complex evaluations at TREC.

2.4.4.2 Legal

The Legal track, started in 2006, resulted from a collaboration of information retrieval researchers with the legal community to "develop and apply objective criteria for comparing methods for searching large heterogeneous collections using topics that approximate how real lawyers would go about propounding discovery in civil litigation" [20]. This in turn was driven by new regulations governing how electronic data would be used in civil cases in federal courts. Doug Oard and David Lewis worked with Jason Baron of the U.S. National Archives and Records Administration to adapt traditional ad hoc testing methodology to this task.

The data collection was the IIT Complex Document Information Processing (CDIP) Test Collection [20] of about 7 million document records from the Legacy Tobacco Documents library. These documents, in the form of XML records, contain many kinds of electronic records such as

email, OCR'd versions of tables and handwritten documents, etc., and therefore represent a challenge both in size and in "noise" levels.

The topics were uniquely created to mirror actual legal searches. The starting points were several (3 in 2008) hypothetical complaints, such as "a shareholder class action suit alleging securities fraud advertising in connection with a fictional tobacco company's campaign". These complaints were used to generate the traditional ad hoc topics (there were 45 in 2008). These complaints and the resulting topics were created by the Sedona Conference Working Group on Electronic Document Retention and Production, a nonprofit group of lawyers, who also then created the baseline Boolean queries for these topics. These Boolean queries would have been the "normal" way of requesting information in this type of legal situation. For more on this, including a deeper background of the legal process, see [20].

There were three tasks in 2008 and 2009 [91]: an ad hoc task, a relevance feedback task, and an interactive task. For the ad hoc task, systems returned both a ranked list and their optimal cutoff (K) on a per topic basis, with the main metric being the F measure at K. Although recall is critical to the legal profession, it is also important to return "good" sets as opposed to just ranked lists. The relevance judgments were made using enormous pools of documents, with various sampling methods tried in different years . The final set of documents (an average of over 500 per topic), were judged by volunteer second and third-year law students. (The interactive task was very user oriented and is discussed in Chapter 3 along with the other interactive tasks in TREC).

2.4.4.3 Blogs

Iadh Ounis and Craig Macdonald from the University of Glasgow worked with NIST to design and create a track dealing with the blogsphere in TREC2006. They created a large blog collection [126] containing 88.8GB of Permalinks, along with associated homepages, crawled from 100,649 blog feeds during an 11-week period. The crawl was specifically designed to include topics of interest to the NIST assessors and to include assumed spam blogs (splogs) to allow for realistic tasks. Note that whereas blogs are not a specific domain such as genomics and legal, they have a different nature (and often language) than newswire or web text, and attract different types of user tasks.

The blog track in TRECs 2006, 2007 and 2008 worked with opinion finding tasks, where the definition of a document was the blog post plus all the associated comments as identified by a Permalink. The task was to find all documents containing an opinion on a given topic, and then to identify if that opinion was positive, negative, or "mixed". The 50 topics were generated by the NIST assessors, using a query log from BlogPulse as seeds for reverse-engineering of the full topic. The documents were judged as: not relevant to the topic (0), relevant but not containing an opinion (1), relevant and containing a negative opinion (2), same as 2 but mixed opinion (3), and same as 2 but positive opinion (4). Traditional ad hoc metrics were used both on the relevancy (marked 1 or higher), and on the relevant including opinions (2 or higher) [127].

A second obvious task for blogs was the blog distillation task run in TRECs 2007 and 2008, where the user model is a person seeking to find an interesting blog to follow or read in their RSS

reader. Here the documents were defined as the full blog, i.e., the aggregates of blog posts, and the topics were contributed and judged by the TREC participants. The guidelines for this were that the blog should be principally devoted to the given topic over the whole timespan of the collection.

The blog track in 2009 [120] had a much larger collection (Blogs08), with 28.5 million blog posts sampled from 1.3 million blog feeds from January 2008 to February 2009. A more complex distillation task required that systems use blog attributes (opinionated, personal, or in-depth) as well as the blog topic, with fifty topics developed and judged at NIST.

2.4.5 PUSHING THE LIMITS OF THE CRANFIELD MODEL

All of the test collections previously discussed were basically using the Cranfield model. Whereas the "documents" may be text, or speech, or websites, or blogs, and there were various user goals being modeled in the topics, the end result was usually a ranked list of unique items for judgment, and most of the collections could be considered re-usable. This final set of TREC tracks are grouped together because each of them required some major deviation from the Cranfield model, usually leading to either limitations on the resulting test collection or unsolved problems in the evaluation.

2.4.5.1 Question-Answering

The question-answering track, which ran for 9 years starting in 1999 (TREC-8), required systems to return *the answer* to a question instead of returning a ranked list of documents. This task both tested a new user model and created an opportunity to work with the natural language processing community. Whereas the document collections were the basic English ones used in the ad hoc task, the topics were questions, starting with simple fact-based, short answer questions (factoids) such as "How many calories in a Big Mac?", and progressing to "definition" questions such as "Who is Colin Powell?".

Note that the definition of a test collection had to be changed for this track. Whereas there are documents and topics/questions, the answer set is not the set of relevant documents, but rather "pieces" of one (or more) documents. In TRECs 8, 9 and 10, passages of lengths 250 and 50 words respectively were returned as the answers (along with the document id), and variable length strings were returned starting in 2002 [186].

The answer sets described above do not constitute a reusable test collection because the unit that is judged is only the answer string. Different runs seldom return exactly the same strings, and it is quite difficult to determine automatically if the difference between a new string and the judged string is significant with respect to the correctness of the answer. Any new system using this test collection is likely to return different answers which could be correct but would be "graded" wrong, making their experimental results not reliable. Whereas there is no known answer to this problem, NIST (and later Ken Litkowski) manually created answer patterns consisting of Perl string-matching algorithms from the answer strings [192] and researchers use these patterns along with the correct document ids to automatically score their new experiments (see software at `http://people.csail.mit.`

edu/gremio/code/Nuggeteer). Note that these are not truly correct scores (see [192] for more on this), but only approximations; clearly in violation of the Cranfield re-usable collection model!

Participants submitted questions the first year, however in the following years NIST created around 500 factoid questions using Excite logs as seeds, real questions from Encarta, and full questions coming from Microsoft and AskJeeves logs. The answer strings for the factoid questions were pooled and judged, with scoring based on mean reciprocal rank (MRR) similar to the known item searching for the OCR track. List questions were added, such as "Name 4 countries that can produce synthetic diamonds", which were not only graded for correctness but also for redundancy, with the score being accuracy (number of correct answers divided by the target number).

For TREC 2002 exact answers were required in order to "prove" that the systems knew the answer. Note that this would not be the actual user model, but could be considered as part of an interface where the exact answer would be highlighted in the text. Only one answer was allowed and the scoring was based on the number of correct answers using a new metric that also tested the confidence of the systems in their answers (see [186] for more details of this metric).

The last five years of the track had factoid, list and definition questions built in a series of questions such as a user might ask about a person, organization or a thing. For example, someone interested in the Hale Bopp comet might ask factoid questions such as "when was it discovered", list questions such as "what countries was it visible in on its last return", but also want to see "other" strings of interesting information. Whereas the factoid and list questions could be evaluated as before, the "other" strings required a new mechanism. The judging was done in two stages. In the first stage the assessors looked at a pool of all the system response strings and also added their own strings. These strings were then broken into "nuggets" for evaluation purposes, i.e., facts for which the assessor could make a binary decision as to whether a system response contained that nugget. Some of these nuggets were tagged as "vital", and systems were scored on recall (how many of the vital nuggets did they get), and precision, where precision was based on the length of the full set of strings (see [63, 182, 190] for more details).

2.4.5.2 Spam

The spam track was initiated in 2005 by Gordon Cormack of the University of Waterloo to encourage research in spam filtering. The major difficulty was the privacy issue since public distribution of personal email files is unethical and the few available public email files (such as the Enron data), or public forums do not accurately reflect real user experience. The track ran for 3 years, with a carefully constructed public corpus consisting of 75,419 messages (about two-thirds spam), and a private corpus from "MrX" consisting of 161,975 messages (95% spam). The public data from previous years could be used for training purposes, whereas the private corpus was accessed only via a toolkit that used 5 command-line operations (initialize, classify, train ham/spam, and finalize). The toolkit also was used in the evaluation to implement the various feedback tasks and to record and score results [58].

Four feedback tasks were designed to model different user behavior. The first gave immediate feedback, i.e., the correct answer was provided to the filter for quick "re-training". The second

and third methods (delayed and partial) gave answers either at random times or for a subset of the messages. Finally, the active on-line learning task allowed participants to request answers for a small quota of the messages. The messages were presented for classification to the filters by the toolkit in a time-ordered fashion, with answers returned depending on the appropriate feedback task. After all messages had been classified, the scoring was based on ham misclassification and spam misclassification, with both ROC curves and a single metric (logistic average misclassification average provided (see [58] for definition of this metric).

2.4.5.3 Routing and Filtering Tasks

TREC also had a second type of task right from the beginning: the routing task. Here it was assumed that the user model is someone with an on-going information need, someone who has done some searching (and therefore has some relevant documents for their need), but wants to continue collecting more information. So the topic is fixed, but the document set keeps changing/growing. This user model certainly reflects an intelligence agent, but also news-clipping services, users following a particular news story, etc.

For the first 5 TRECs, the Cranfield model was essentially unchanged. The topics used were generally the ones from the previous TRECs, with the relevance judgments available, and the documents were some similar data that was specially gathered for the routing task. The results on the new data were submitted as a ranked list, then evaluated using the ad hoc pooling techniques and metrics. This scenario mostly worked, although at times it was difficult to find new data that matched the old topics, and participating systems tended to use this task as a way of improving their ad hoc results rather than thinking about a routing application.

It became clear by TREC-5 that this task did not reflect many user models; a more realistic task deals with filtering a time-ordered set of documents into a set of relevant ones for the user "inbox". This task could be modeled in several ways, such as in the batch filtering mode with "training" documents from one time-ordered set, and the test set from a later time-ordered set, or as an adaptive filtering task where training is incremental based on feedback from the user. However either case results in a set-based retrieval rather than a ranked list.

These tasks, started in TREC-6, and running for five years, along with the routing task, posed several big challenges in terms of evaluation [136]. The first was the continuing need to find reasonable data, and various options were tried, some more satisfactory than others. The final year TREC-11 (2002) [161] had a filtering collection specifically built for the Reuters collection using new TREC-style topics with a "complete" set of relevant documents formed by doing successive searching. It is *strongly* recommended that new filtering experiments use these 50 manually built topics from 2002, rather than the ones built for the earlier TREC filtering tracks. (Note however that there was also a set of 50 "experimental" topics built automatically in 2002 using the category tags from Reuters; these were not successful and should not be used.)

Another problem involved the pooling for relevance judging since the systems were not using ranked lists. Several methods were tried [118], such as sampling the results from each system and

estimating the performance based on that sample, or sampling the pool of documents from all the systems. The fact that the routing task was continued in these years, allowing ranked list input to the pool somewhat helped the problem, but again, the 2002 filtering collection is strongly recommended because of its excellent set of judgments.

The final problem involved the metrics, which were generally utility measures. Note that because the systems are tuning on these metrics, one specific metric needs to be selected in advance, unlike the large range of metrics for the ad hoc task. Diverse utility measures were tried (again see [136]), including some F measures. The "correct" metric to use remains an unsolved problem as each metric reflects a different user model and the choice of the metric strongly effects both the results and the sampling methods that are used for pooling.

2.4.5.4 Terabyte, Million Query and the "new" Web Tracks

The Terabyte track started in TREC 2004, with the goal of working on a larger, very realistic web collection. The collection (GOV2) had 426 gigabytes crawled from U.S. government sites (the .gov sites), including extracted text from PDF, Word and postscript files. For the first year of the track, there were 50 ad hoc topics created in a similar manner to the original web track and the top 85 documents from each run were pooled for the relevance assessments. In 2005 and 2006 a named page finding task was also run, where the task was to find a particular page that a user might have found once and then tried to relocate (known item task with "near duplicates"). There was also an efficiency task.

Whereas the Terabyte track was a mainly scale-up of the old web track to a larger collection, it was suspected that the relevance judgments would not be "complete". A new measure *bpref* [35] specifically designed for incomplete collections was tried, and the pools were made larger (top 100 documents) in 2005. Additionally it was found [32, 47] that the pools were biased in that there was a very high occurrence of title words from the topics in the relevant documents. This means that runs using mostly title words could score higher than more "exotic" runs because documents without title words were simply not judged. This problem was likely caused by the huge scale of these collections, with many occurrences of relevant documents that never appear in the pools.

In 2006 the Terabyte track [37] made an effort to measure this effect by building three separate pools. The first pool (used for MAP evaluation) was the top 50 documents from each run. The second pool started at rank 400, selecting additional documents up to a total of 1000 documents per topic to gain some idea of how many additional relevant documents were not judged, with the metric being the *titlestat* measure [32]. Here it was found that there were indeed more relevant to be found at these deeper depths, with the titlestat measure still reading 0.6 at depths 400-500, down from 0.74 at depths 1-100. The third pool was a random sample using estimates of probability of relevance based on the relevance judgments from the first pool (top 50), and results of experiments done in 2006. This third pool was used for a new evaluation measure *inferred average precision, infAP* [202]. Three metrics for effectiveness were used in 2006: MAP, bpref and infAP.

Efforts to deal with the pooling problems led to the Million Query Track starting in 2007 [7]. The idea of this track was to investigate whether it was better to use many more queries/topics with

shallow judgments rather than the more traditional pooling for 50 topics. There were 10,000 queries selected from a search engine log with the only requirement being that there was at least one click on the GOV2 corpus. In 2007 there were 1700 of these that were judged for 40 documents, where the judgments were done at NIST and by the participants. The queries to be judged were selected in a "quasi-random" manner, with the assessor picking a query from a set of 10 (or skipping to the next set), and then using that query as a seed to build a full TREC topic statement. The documents that were judged for each topic did not come from traditional pooling but were sampled using one of two different methods, Minimal Test Collections (MTC) [41] or statAP [12, 13], using the appropriate evaluation measures for the results.

The track in 2008 modified several parameters in order to more deeply investigate these two methods. The 10,000 queries were equally divided into four sets: those with 6 or fewer words that had fewer than 3 clicks in GOV2, those with more words and more clicks, and the other two possibilities (few clicks/long query, more clicks/short query). The judgments were done only at NIST, with 792 queries judged for up to five different stopping points (8, 16, 32, 64, and 128 judgments). This allowed a finer-grained investigation of the strengths and weaknesses of the two different methods. It was found [6] that the MAPs were more stable using MTC for the smaller number of judgments (16/32), whereas statAP was more stable for 64/128 judgments. However MTC seemed more affected by the different query categories.

Both the Million Query and Terabyte (renamed Web) tracks continued in 2009 using a new collection ClueWeb09 (http://boston.lti.cs.cmu.edu/Data/clueweb09). This collection consists of roughly 1 billion web pages (25TB of uncompressed data) in multiple languages resulting from a crawl during January and February 2009. A smaller subset (Category B) of the collection consists of 50 million documents which are approximately the first 50 million English documents hit by the crawl plus the English version of Wikipedia. It should be noted that the Category B subcollection has many unique characteristics, including high-quality "seed" documents that were used to generate the rest of the crawl and results from this subcollection (or any subset of a large web collection) need to be carefully considered with respect to any conclusions or generalizations not tested on the full set. The web track in 2009 had an ad hoc task (50 topics using MTC pooling and evaluation), and a pilot diversity task where documents were also judged with respect to special subtopic fields.

2.5 OTHER EVALUATION CAMPAIGNS

Since 1992 there have been other evaluation campaigns, often starting in a similar vein as TREC and then branching out into new areas especially appropriate for their participants. In addition to the three discussed here, there is the FIRE evaluation (http://www.isical.ac.in/) with a concentration of work with various Indian languages and the Russian evaluation (http://romip.ru/en/).

2.5.1 NTCIR

An Asian version of TREC (NTCIR) started in 1999, with the conferences occurring every 18 months since then. Whereas NTCIR has run evaluations in a similar mode to TREC, there has always been a tighter connection to the NLP community, allowing for some unique tracks. Additionally NTCIR pioneered retrieval evaluations with patents, developing appropriate evaluation techniques for searching, classification, and translation efforts in this field. Online proceedings and more information about NTCIR can be found at `http://research.nii.ac.jp/ntcir`.

This first NTCIR (August 1999) worked with 339,483 Japanese abstracts from 65 Japanese academic societies; about half of these abstracts were also in English, and all were written by the authors. The 83 Japanese topics were gathered from researchers, with assessments done via pooling and at three levels (relevant, partially relevant and non-relevant). NTCIR-2 (March 2001) had a similar task, with the same collection but with the 49 new topics also translated to English (allowing CLIR from English to Japanese) and a fourth grade for relevance judgments (highly relevant).

By NTCIR-3 (October 2002) there were three languages for CLIR, with over two hundred thousand newspaper articles in Japanese and Chinese from 1998-1999, plus a smaller collection in Korean. Topics were built in four languages (including English), and were translated for the other languages, with tags in the topics to indicate both the source language and the target (translated) language. Note that because trec_eval was not set up for graded relevance judgments, there was a "rigid relevant" (highly relevant and relevant) score and a "relaxed relevant" (also including partial relevant) score. More newspapers were added in NTCIR-4 (June 2004), making almost equally sized collections for Japanese, Chinese, Korean and English (from English versions of Asian newspapers). The compatibility of the document sets across all four languages allowed complete multilingual testing to be done, and this continued in NTCIR-5 (December 2005) with new document sets from the 2000-2001 timeframe.

The NTCIR-6 (May 2007) CLIR task was run in two stages and without the English documents. In stage one there were 140 topics reused from NTCIR-3 and 4, but this time against the newer document sets from 2000-2001. Only shallow pools of the top 10 documents per run were judged, and 50 of the topics having enough relevant documents were used in the scoring. In addition to the rigid and relaxed scores, the multi-grade relevance scoring (nDCG and Q) was used. Stage 2 was a rerun of the NTCIR-6 systems, without parameter adjustment, on the NTCIR-3 thru 5 test collections. The goal of stage 2 was to provide information for cross-collection analysis.

The NLP connection in NTCIR encouraged both a summarization task (NTCIR-2-NTCIR-4) and a question answering task. The summarization task was done in Japanese only; looking at single and multiple document summaries of newspapers. Different evaluation methods were tried, including comparing the submitted summaries on a scale from 1-4 (best to worst) for completeness and readability and counting the number of character edits (insertion, deletions, substitutions) that were needed to turn the automatic summary into the human summary. The question answering task was also done in Japanese only and was similar to the TREC 2002 QA task in that exact answers were required for the 200 questions. However a subtask consisted of 40 "followup" questions, where

a question related to one of the 200 initial questions asked for more information using pronouns or ellipsis. This task continued in NTCIR 4 and 5 with series questions, but became a cross-language QA task starting in NTCIR-6.

The cross-language effort and the question-answering effort merged into one task in NTCIR-7 and 8, allowing for component-based evaluation in both the retrieval and the question components, plus providing a platform for exchanging interim results and combining modules for different systems.

NTCIR has had a major patent retrieval task since NTCIR-3 [71]. The first year the task was a technology survey using search topics inspired by newspaper articles in four languages against two years of Japanese patents with the goal of finding relevant patents. This task became a true patent invalidity task in NTCIR-4, against ten years of Japanese patents, including major experimentation into how to get large number of topics (invalid patents) more easily by using the rejection citations as the relevant documents. This method was enhanced in NTCIR-5, along with including a subtask using passages as opposed to full documents. NTCIR-6 (2007) saw more emphasis on CLIR in patents, with both a Japanese patent invalidity task and an English one using 10 years of U.S. Patent data. A translation task was added in NTCIR-7 and 8, including an aligned corpus of these patents. Patent classification has also been evaluated, along with assignment of patent classification codes to research papers, both monolingually in Japanese and English, and in a cross-language mode.

2.5.2 CLEF

The European CLIR task moved from TREC to a new conference, CLEF (Cross-Language Evaluation Forum) in 2000. Whereas one of the reasons for the move was that TREC needed to work on non-European languages, the major reason was that it become obvious that U.S. participants did not have the background (nor maybe the interest) to progress much further in the research. The European partners who had worked with TREC were interested in starting a new conference, led by Carol Peters of CNR Pisa, not only to continue the European CLIR effort but to expand it to more languages and more participants from Europe. Working notes from all of the CLEF workshops can be found at `http://www.clef-campaign.org/`, with formal proceedings produced in the Springer Lecture Notes in Computer Science each year.

In 2000 there were the initial four TREC languages (English, French, German and Italian), however, each year thereafter saw the addition of one or two new languages, starting with Spanish and Dutch in 2001, and eventually ending with 13 different languages for the documents, and even more languages for the topics. In order for a new language to be added, a native-speaking support group from that country needed to obtain adequate newspaper data from the 1994/1995 period, arrange for the data to be put in TREC-like formatting, and then provide personnel to create topics, perform relevance judgments, etc. This was a major effort and it is a credit to the CLIR community that so many groups were able to do this. As languages were added, research groups in the new language area were able to perform their research using their own language for the first time as monolingual ad hoc retrieval tasks were offered in these new languages. Additionally the cross-

language effort (bilingual and multilingual) was continued, with a main task of retrieving documents across 8 different languages, a smaller task of using only 4 languages (English, French, German and Spanish), and then different specific bi-lingual pairs, such as Finnish to German, Italian to Spanish, French to Dutch, etc. Not all language pairs were offered in a given year in order to concentrate the effort of the participants. The expansion of the CLEF CLIR ad hoc tasks into so many languages not only enabled many new research groups to join in, but for the first time allowed major investigations into the differences between languages with respect to retrieval [155].

There were several specific evaluation issues given the widespread distributed nature of the CLEF evaluation effort. First, topic creation needed to be managed very carefully to truly reflect the user model. The initial topics needed to mirror the types of questions asked within the various countries, and this was done (for example) in CLEF 2001 by asking each of the 5 language groups that year to generate a set of 15 topics, along with doing a pre-search to make sure the topics were covered in each of the 5 languages [201]. These 75 initial topics were then jointly pruned to 50 topics based on potential problems in one or more of the languages. The final set of topics was then translated directly to the other languages, with indirect use of the English master set only when necessary. Note that it is critical that the translation not be word-for-word, but take into account both the linguistic and the cultural background of the target languages, i.e., the final topics must represent how a user in a given country would actually ask that question.

Another issue for CLEF was pooling given the sparse nature of the submissions. For example, in 2003, the first year for Finnish and Swedish, there were only 7 groups for Finnish and 8 for Swedish. A "uniques" test was performed [30] to check for the completeness of the resulting test collection. This test involves creating multiple sets of relevance judgments, actually *n+1* sets containing the original (full) set plus *n* sets built by removing those relevant documents uniquely found by one participating group. These *n+1* sets are then each used to create the test results, with the goal of showing how a given group would have performed if their unique relevant documents had not been in the pool, i.e., not considered relevant. The results for Finnish and Swedish showed a maximum drop in performance of 0.31% for Finnish and 2.02% for Swedish, which can be considered not significant.

CLEF also attracted many new tracks over the years, such as cross-language question answering (24 groups doing working in 10 languages against target text of 9 languages in 2005), a geospatial track specifically looking at topics with geospatial information, and a structured data track (GIRT) using structured documents in 3 languages. One of the more unusual (and popular) tracks was ImageCLEF, which started work with captioned photographs in 2003, where the goal was to search topics built in 5 languages against the English captions. This track expanded to include over 20,000 "touristic" photographs with captions in English, German and Spanish, along with 50,000 medical images with annotations in several languages. Whereas the initial and continued main task was work with the captions, eventually the images themselves were also used in the retrieval task, either alone or in combination with the captions.

2.5.3 INEX

The INitiative for the Evaluation of XML retrieval (INEX) started in 2002 to provide evaluation of structured documents, in particular to investigate retrieval of document components that are XML elements of varying granularity. The initiative used 12,107 full-text scientific articles from 18 IEEE Computer Society publications, with each article containing 1,532 XML nodes on average. The collection grew to 16,819 articles in 2005 and moved on to using Wikipedia articles starting in 2006. Like its sister evaluations, INEX has also had auxiliary "tracks" over the years (see `http://inex.is.informatik.uni-duisburg.de/` for INEX through 2007; the current site is `http://www.inex.otag.nz/`).

The main task of ad hoc retrieval has run since the beginning but with major differences from TREC, NTCIR, and CLEF. Because the goal is retrieval of a particular element of the document that is demarcated by XML tags, the "relevant" documents can range from a paragraph to the whole document. The general idea has been to present the user with the "best" element in the article with respect to their information request, that is an element that exhaustively discusses the topic without including too many irrelevant topics. The notions of exhaustivity and specificity, although well known to the information retrieval community, are very difficult to measure, and this has caused extensive investigations of new metrics within INEX over the years [114]. This difficulty also extends to the relevance assessments and is part of the reason that all of the relevance assessments in INEX are done by the participants.

The ad hoc topics in INEX (also built by the participants) have reflected the structural nature of the task. The content-only (CO) topics resemble TREC topics, however the content-and-structure (CAS) topics include NEXI query language in the title section, which provides specific structural criteria. Results from both kinds of topics have been evaluated similarly, although the structural constraints in the CAS topics were interpreted variously.

The 2007 ad hoc track illustrates the complexity of the evaluation. There were four subtasks with different sets of results: thorough (ranked list of elements), focused (ranked list of focused non-overlapping elements), relevant in context (ranked list of the full articles, but including a set of non-overlapping elements), and best in context (ranked list of the articles, but including the best entry point). Each task models a different user approach, with the focused assuming a top-down look across the ranked elements, and the relevant in context and best in context assuming the user wants to see the full document, but with more complex displays for the elements within that document.

There were 4 additional tracks in 2005, building up to 8 tracks plus ad hoc by 2009, with the common thread being the use of XML. The longer running tracks have been interactive (user studies for XML documents), multimedia (retrieving images using the XML structures), and document mining (clustering and classification of XML structured documents). A book search track for digitized books, entity ranking and question answering within the XML environment started in 2007. Because the data is Wikipedia, there is a Link-the-Wiki track, and also an efficiency track.

2.6 FURTHER WORK IN METRICS

There has been considerable interest and work on new metrics since 1992, coming from perceived needs of TREC and other evaluations, but also from the explosion of information access available today. Since the emphasis in this chapter has been more on the design of test collections, this section is only a brief summary of the new metrics; for a much more detailed discussion, see [147].

By TREC-2 the new trec_eval was in place using non-interpolated average precision and the new R-precision (see section 2.3.2). But the known-item searching in the TRECs 5 and 6 OCR and speech tracks required another new metric, *mean-reciprocal-rank, MRR*, to allow proper averaging. MRR is the mean of the reciprocal of the rank where the known item was found, averaged across all the topics, and is mathematically equivalent to the mean average precision when there is only one target document [183].

The various web tracks have required new metrics, both because of the different tasks and because of the scaling issues for pooling. See the Terabyte track, Section 2.4.5.4, for discussion of the various metrics.

Another of the issues in TREC has been the averaging of results across the topics, even though there are wide variations in system performance when results are examined on a per topic basis. The statistical testing such as the t-test used earlier (see Chapter 5 in [147] for a excellent discussion of the use of statistics in information retrieval evaluation) is one possible way of examining these effects. However a new metric *GMAP* using the geometric mean of the average precision scores (as opposed to the arithmetic mean) was used in a variation of the ad hoc track in 2004 to emphasize success on the "harder" topics [187].

The final metric briefly covered in this section was not driven by TREC issues but by a general awareness that binary relevance judgments were not sufficient and that some type of grading of judgments was needed (similar to what is done for commercial search engines). The group at Tampere University [99] devised a framework for dealing with graded relevance judgments *Discounted Cumulative Gain, DCG* which uses both the relevance grades and the position on the document ranking to give a single score. This can be parameterized to give more or less discounting. A normalized version of this (nDCG) [100] is heavily used both in commercial search engine evaluations and in other evaluations having graded relevance assessing. The measure has also been recently adapted to handle multi-query sessions [101] and diverse query evaluation [46].

2.7 SOME ADVICE ON USING, BUILDING AND EVALUATING TEST COLLECTIONS

This final section departs from the more formal coverage of the batch evaluations by offering some advice on selecting appropriate test collections, building new test collections, and evaluating those collections. This advice is not meant to be a complete manual on this topic but rather comes from personal observations during many years of working with test collections.

2.7.1 USING EXISTING COLLECTIONS

The easiest evaluation method for batch experiments is to use an existing collection. This method not only cuts the major costs of building a collection, but also provides training material. An equally important issue is the universal acceptance of these test collections, including the ability to compare results to other work. However this decision should not be taken automatically; the user task and the assumptions about the users should be appropriately matched to the selected test collection and test collection characteristics need to be considered in any analysis of the results.

The most heavily used test collections for monolingual English retrieval are the TREC ad hoc ones. These collections are based on mainly on newspapers and newswires, along with government documents. The topics are general-purpose and domain-independent, and there is a reasonable assumption that the relevance judgments are complete. But there are 9 sets of topics (1-450), searched against different document sets, so which to pick? The best choice for most experiments are the 3 sets used in TRECs 6-8, consisting of 150 topics (numbers 301-450) searched against (mostly) the same data (disks 4 and 5). This set provides 150 topics, enough for good statistical analysis, plus this group is the most consistent in terms of topic format and relevance judgments.

Some of the earlier TREC ad hoc collections need to be used with caution. Topics 1-50, used for minimal training in TREC-1, are poor topics with only minimal relevance judgments. Topics 51-150, used in TRECs 1 and 2 have an expanded format; this may be useful for particular kinds of experiments, such as structured query experiments, however the topics themselves were created in a possibly unnatural manner, with the relevance judgments being done by another person. Topics 150-200 (TREC-3), were constructed with reference to the documents, and because they often use terms from the documents, are the "easiest" of the TREC topic sets. Topics 201-250 for TREC-4 have no narrative field, which may or may not be necessary depending on the experiment.

Other TREC track test collections are also available (`http://trec.nist.gov/data.html`), including test collections for OCR and speech, non-English ad hoc collections, and collections for web, blog, genomics, legal, etc. Whereas the Chinese and European language collections are available (the Chinese one for TREC-6 is not recommended because of issues with the topic building process), it is better to get the NTCIR and CLEF collections for these languages. The Arabic collections (TRECs 2001 and 2002) are available at NIST, although the second one (2002) is the recommended one since the improved systems meant better pooling. Other collections are also available, usually with the topics and other auxiliary data on the TREC web site and a pointer to the documents which are available elsewhere. Note however that these collections are all specialized, based on possibly narrower tasks and/or user models and these issues should be thoroughly understood (by reading track overviews about the collection characteristics) before using these collections.

What about using the much older small collections? These collections are much too small for testing and validating new technology; furthermore most of them use abstracts rather than full documents and results may well be misleading because of this. However one exception to this would be to use them (particularly the TIME collection) as teaching tools; but here again the benefits of

a much smaller collection for failure analysis need to be weighed against the different insights one gains when working on the larger collections.

2.7.2 SUBSETTING OR MODIFYING EXISTING COLLECTIONS

Another option is to subset an existing collection or modify it to better fit an experiment. Many of the collections are small and therefore cannot be subset in terms of topics, however one could concentrate on subsets of the documents. For example groups have studied the effects of hyphenation and effective retrieval of long structured Federal Register (FR) documents [199]. Other characteristics of the data such as spelling errors (especially in newswires), duplication of information (again mostly in newswires), British vs. American English (FT vs. WSJ), evolution over time of news stories, and other areas invite further study. Additionally some sources of documents provide fielded information ranging from titles and headlines to manual indexing terms to the heavily structured data in the patent collection. Any of these ideas could be used to generate subcollections of documents, however care needs to be taken that there are enough documents in the subset to generate valid results and any possible biases introduced in this subsetting need to be considered in result analysis.

Modifying a collection is more difficult and (possibly) loses some of the advantages of using an existing collection. One obvious modification would be to change the relevance judgments, either by changing the unit of judgment (doing passage retrieval for example) or by changing the definition of relevance. The TREC relevance judgments for the ad hoc task are the broadest type of judgments, i.e., the fact that a document contained ANY information about a topic/question was enough to make it relevant. This was important because the perceived definition of the TREC task/user was that of a high-recall task. But it also was important in terms of creating the most complete set of relevance judgments possible. The current judgments could be used as the starting point for other types of relevance judgments, such as removal of "duplicate" documents [24], or the use of graded relevance judgments [99, 162] or even the measurement of some type of learning effect (the TREC novelty track). This type of modification is very tricky; in essence a second relevance judgment is being made, with all of the consistency difficulties discussed earlier, which may affect the experiment. For some discussion of the problems, see [81, 203].

2.7.3 BUILDING AND EVALUATING NEW AD HOC COLLECTIONS

Building a new collection is a major step to take; it is costly in terms of time and money and full of pitfalls for those new to this task. However if it is to be done, then the critical thing is that the experiments and the new collection they use be modeled on some real user task and that the characteristics of likely users be considered as part of this task. Are the users searching the web for a nearby restaurant with good reviews, are they searching their company's intranet for patents, or are they browsing the web for information about some specific type of tree they want to plant? Each of these applications requires a different set of documents, different types of "topics", different definitions of relevance, and different types of metrics to use in evaluation. These decisions need to be made long before the collection is built.

Once this piece of the design is done, then the next step is to find some documents. Again the easiest place to start would be some existing set of documents that can either be used as is or be subsetted in some manner. Possibly the news collections from TREC/NTCIR/CLEF are useful, but more likely the new web collection (`http://boston.lti.cs.cmu.edu/Data/clueweb09`), or the blog collection [120] are candidates. Note that any sub-setting of these collections needs to be done carefully; both the blog and web collections were carefully sampled during their construction process and any subsets need to reflect the new user task design being envisioned. The small web collection [89] design gives some clues on how to do this, the construction of the WT10g [15] also discusses web collection design or see the discussion of the recent "Category B" of the ClueWeb09 collection. Note that subsets of these collections cannot be re-distributed, however one could make subsets using the "docids" for reuse by others. If completely new document sets need to be collected (a major job), then hopefully this can be done in some manner so that the work can be used by others. This means that any intellectual property rights need to be resolved and that the data has to be formatted for ease of use.

Given that the documents are collected, the next step is the topics. Again the user task/characteristics need to be modeled; ideally some real topics from a log can be gathered, or some topics can be built by "surrogate" users such as those used in the TREC/NTCIR/CLEF ad hoc tasks. Enough topics need to be built to overcome topic variation in performance [188, 189]; 25 can be taken to be the absolute minimum, with 50 a more reasonable number. The format of the topics needs to mirror the task; for example browsing the web looking for a specific item may need a series of related topics to mimic the interactive search process. Topics for a specific domain need to be appropriate for that domain, either by getting domain experts to build the topics (such as the legal track), or by using a survey such as that done before topic construction by the genomics track to gather representative needs of the genomics community. It is equally critical to closely examine the issues with searching a given genre, such as the patent tracks in NTCIR, TREC and CLEF. The NTCIR efforts in multiple years have tackled different pieces of the patent retrieval problem, but all based on realistic analysis of the needs of that community [71]. If multiple languages are involved, then the topics need to be constructed so that there is no bias towards one language [201].

The methodology for making the relevance judgments for the topics is again reliant on the user task and characteristics. Does the user want only document level judgments or are passage (or even sentence) level judgments needed (such as for genomics and question-answering)? What types of judgments are required (binary, graded or other), how many judgments per topic (a major cost factor), and of course setting up the mechanics of getting the documents to judge (pooling, manual search, or some other method such as the sampling done in the TREC Million Query Track). Many different approaches to pooling have been tried over the years (see section 6.3 in [147]), each with some advantages and weaknesses that need to be considered in light of the goals of the test collection. Finally, of course, the judging needs to be done.

Once the collection is built, some types of validation need to be made. If this collection is used only for a single experiment, the validation is needed only to understand any likely biases that affect

the analysis of the results. However if the test collection can be used by others, then some measures must be made of the consistency and completeness of the relevance judgments (see section 2.3.2 in this lecture or Chapter 6 in [147]).

2.7.4 DEALING WITH UNUSUAL DATA

The previous section built on the assumption that the test collection is a "standard" collection working with textual documents, but many interesting applications have more diverse data. For example, the TREC speech retrieval task (and the video retrieval task) not only worked with multi-media data but had to define the "boundaries" of the documents. Here, and also in the blog track, these definitions were coordinated with the specific task being modeled, with a document defined as the blog post plus all the associated comments as identified by a Permalink for opinion finding, but defined as the full blog (the aggregates of blog posts) for the blog distillation task. In both cases the definition of a document was based on the output a user might expect to see for these tasks. This was also the basis for the various versions of the question-answering track, and for the passage retrieval part of the genomics track.

It can be the situation that the data to be searched has a major impact on the tasks; this is clearly true of the various TREC web/enterprise tracks, but also true of the NTCIR patent tasks. There the patent data had multiple fields and was multi-lingual, leading to a series of different tasks including cross-language retrieval to related newspaper articles in four languages, patent invalidity tasks, patent classification tasks, and patent translation tasks. The ImageCLEF track in CLEF started with captioned black&white images, where the major emphasis was on using the captions, moved to color photographs with multilingual captions where the images became equally important, and tackled a series of medical retrieval tasks with x-rays. In each of these cases the data determined the task; a similar situation occurs in the TRECvid evaluations. INEX has worked with semi-structured data, including books, Wikipedia, etc., with the tasks, topics, and metrics highly influenced by the data to be searched.

2.7.5 BUILDING WEB DATA COLLECTIONS

The TREC terabyte and web track built test collections based on the assumed needs of the TREC research community. The documents (web) were basically what could be obtained, and the topics and relevance assessments changed over the years to allow different research "themes" to be investigated. However the commercial search engines also build test collections, where they have the luxury of the full web, a real set of questions, plus other interesting metadata about searching. There is little information about how this is done, although Chapter 3 of this lecture has a discussion of the work done in the mining of search logs.

A paper presented at SIGIR in 2009 [122] provides one example of how this was done at Yahoo. The document collection is the entire web (at some given time). The query collection has 22,822 queries, randomly sampled from some population. A similar collection built by Microsoft [198] had 10,680 unique query statements taken from interaction logs where the users had provided consent

to participate in this study (likely a similar case from the Yahoo test collection). In both these cases, the URLs of the retrieved documents were available. The Yahoo test collection had judgments for an average of 23 URLs for each query, done by human editors (not the users), on a rating of Perfect, Excellent, Good, Fair and Bad. The Microsoft collection also had relevance judgments (assigned on a 6-pt scale) by trained judges for those documents the users had examined during the study (note that this is going to be a special subset of the retrieved).

Both of these test collections emphasized large numbers of topics with shallow relevance judging, and likely the collections were built specifically for the given experiments. The metrics used depended on the goal of the experiments but they were all either high precision metrics or some form of graded relevance metrics. The huge amount of other metadata collected for the searches (see examples in Chapter 3 and in [175]) allows combining the relevance judgments with other information such as assigning grades based on what percentage of the user population clicked on this answer.

CHAPTER 3

Interactive Evaluation

3.1 INTRODUCTION

Information retrieval systems are built for users and it is critical to examine how the various information retrieval techniques (such as those evaluated by the batch retrieval methods) perform in operational settings. Conversely it is important to observe users in an information seeking task and then attempt to model how users behave in order to devise new retrieval techniques. This chapter looks at evaluation issues in user studies, with an emphasis on those studies aimed at understanding user behavior rather than testing the usability of systems. The chapter starts with a short review of evaluation in early user studies, followed by a discussion of the evaluation experiences in the interactive track in TREC. The third section is a look at some recent user studies, with the goal of examining how the high-level issues of experimental design affect the outcome of the final evaluations. The final section presents work using the massive log files from search engines, again with an emphasis on how the design of the experiment, including the selection of which log files to study, how to filter the files, etc. enable user behaviors to be teased out of that data.

It should be noted that the scope of this chapter is narrow; the recent tutorial (Methods for Evaluating Interactive Information Systems with Users) by Diane Kelly [111] provides detailed methodologies for organizing user studies, data collection and analysis techniques, etc. This lecture is meant to be a complement to that tutorial, examining the bridge between evaluation in user studies and in the batch laboratory mode. Other references for user studies are the recent book by Ingwersen and Järvelin [98] which presents frameworks for evaluation of various information seeking tasks, and the special issue of Information Processing and Management [28] on evaluation of interactive information retrieval systems.

3.2 EARLY WORK

Most of the early user studies involved either indexers or search intermediaries, the only users of retrieval systems at that time. Indeed the Cranfield I experiment discussed in Chapter 1 looked into different characteristics of indexers, including the knowledge of the subject matter being indexed and the level of general indexing experience. Both the Medlars study and the Case Western study also discussed in that chapter draw the clear conclusion that "human error" is the cause of various search failures, either at the indexing stage, or at the search construction stage. However this might be better called "human variation" since the issue is not one of error but rather a mismatch of search terms and document terms (index terms in this case), a scenario very familiar to today's searchers.

Michael Keen (earlier associated with both the Cranfield work and the SMART work) continued investigations into indexing. His first experiment looked at five different types of indexing languages in eight test comparisons to study specificity and language, specificity and exhaustivity, methods of co-ordination, and precision devices [105]. There were 800 documents, 72 requests from real users, and full relevance judgments using library staff. But in this experiment the searching was done in a less constrained manner by library science students, who were allowed a maximum of 40 minutes to search and were given a specific recall/precision target. A Latin Square design (see [111] for more information on Latin Square designs) was used so that each of the searchers processed each request once, but eventually used all of the different indexes, manually recording their searches for later analysis.

His followup experiment [109] emphasized the searchers' role, this time using different types of printed indexes and allowing complete freedom of search. The aim "was to test the effects of index entry variations on performance from the user's viewpoint", and 9 entry types were selected based on an extensive survey of printed indexes. These entry types involved features such as index term context, entry term order, and number of access points, and looked at constructions like lead term, KWOCs or KWACs. A subset of 50 requests, and 392 documents from the earlier testing was used, along with the earlier relevance judgments. The searchers were asked to "imagine that they were in a real-world search situation, to simulate delegated searches by subject experts, where a few relevant entries were required rapidly, and the search subsequently broadened, within a time limit of ten minutes for each search". The searchers kept extensive records of their search, including timing and relevance ratings, and annotated the index entry forms to track what they were using for the search. This was supplemented by audio recordings (think-alouds), with all this material then analyzed to provide not only some answers to what types of index entry forms were most useful, but also some insights into the mental processes of the search operations.

Note however that this set of experiments, and other similar experiments looking at the search intermediaries, were not addressing the "real" user of the system. It was assumed that the end users brought their information needs to a human search intermediary, who constructed a search (usually a series of searches) against some online data base such as MEDLINE or DIALOG. Taylor's prescient paper [174] explored how to better help end users by examining the user/search intermediary interaction, in particular noting that a user's initial question was "ambiguous, imprecise, and requires feedback from the system or from a colleague in order to provide an acceptable answer". Nick Belkin, working on a PhD in the late 1970s also wondered how end users might better communicate their information needs either to the intermediary or to a system directly. His ASK (Anomalous State of Knowledge) hypothesis was "that an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly" [23]. He further hypothesized [22] that these ASKs would fall into various classes requiring different types of retrieval strategies, and that a major goal of any information retrieval system would be to decide which types of strategies would be effective.

He and Bob Oddy did a design study [23] to test this model by tape recording interviews with 35 users of the Central Information Services of the University of London and then turning their stated information needs into structural representations of that ASK. Whereas the majority of this research involved creating these ASK structures, creating document structures and doing some type of matching, the research was the one of first to look in detail at these end user information needs. A summary of Table 6 of [23] gives five different ASK types based on these interviews, and the results are very similar to what is seen today.

A. Well-defined topic and problem

B. Specific topics. Problem well defined. Information wanted to back up research and/or hypotheses

C. Topics quite specific. Problem not so well defined. Research still at an early stage.

D. Topics fairly specific. Problems not well defined. No hypotheses underlying research

E. Topics and problems not well defined. Topics often unfamiliar.

By the early 1980s the end users were finally getting a chance to search on their own, with some of the larger libraries creating online versions (called OPACs) of their card catalogs. For example the MELVYL system at the University of California, Berkeley had a prototype online catalog of MARC records from all University of California campuses available to their library staff in 1980 and opened it to library patrons in 1981. In 1982 the National Library of Medicine (NLM) actually user tested [157] two different prototype access systems to their CATLINE online catalog, which contained over 500,000 records at the time of the study. One of the systems, CITE [64, 65] incorporated natural language input and ranking similar to the SMART system, with the other being a more conventional Boolean system (the ILS system). The user testing was done in three stages, with the CITE system available from late April through May, and the ILS system available from June through August. Over 600 surveys asked users about the amount of information retrieved, the system response time and user satisfaction with both the results and the system . Additionally 60 randomly selected library patrons used both systems (controlled as to order) before taking the survey. Siegel [157] noted that "the three studies of user acceptance yield strong and consistent patterns of user preferences, which were separately corroborated by the results of technical testing." CITE was the system was overwhelmingly preferred and was used as NLM's online catalog system for 2 or 3 years until replaced by Grateful Med (personal communication from Tamas Doszkocs).

As more OPACs came into being, more user studies were done, with user surveys being supplemented by transaction log analysis. Tolle [176] reported on work done with logs obtained from six libraries, including the NLM CATLINE (8 weeks of 441,282 transactions from over 11 thousand users), with the specific types of data collected including the terminal identification, the system used, the file searched, user commands, times of starting and finishing, and the system response. The data from these logs was mapped into 11 primary user action states, such as start of session, set creation/searching, display of records, access to help functions, and errors that occurred. These states were then used to construct a state transition matrix, which could be analyzed to answer

specific questions. For example it was found that errors tend to occur in clusters (high probability of transition to a second error), and that subject searches could result in zero hits in over 50% of the sessions. Borgman [27] contrasted user behavior in both bibliographic databases (such as DIALOG), mainly used by professional searchers, and the public OPACs, noting for example that OPAC users are mostly infrequent users who quit searching almost 20% of the time after the first error (9% of their input). Even once past the error barrier, these users had trouble using the Boolean logic, particularly in subject searching, which constituted over 50% of the searches.

The access problems of OPAC users attracted researchers in the 1980s, in particular a group at the Polytechnic of Central London library. A series of retrieval experiments [132], evaluated within an OPAC operational setting (the Okapi system), took place starting in 1984 [123], examining the effects of system changes on user search performance. Note that most subject searching based on keywords then had an implicit "AND" operator between input terms, causing massive failures (34% of 1000 subject searches with no records found [196]). Therefore the first Okapi experiment tried a best match system, with ranking based on relative term frequency in the index (a variation of IDF). The evaluation on a single terminal in the library included 70 structured interviews (user background, what they were searching for, and comments on Okapi), plus transaction logging from which 96 user sessions could be clearly isolated. In general users liked the new system, but it was found that only 38% of the searches were subject searches rather than for specific items (title/author) and that the average number of terms per subject search was just over two.

The Okapi 1987 experiments [196] investigated stemming and spelling correction since transaction logs from the 1984 Okapi system revealed that more than 25% of the subject searches would have been helped by simple stemming, and about 10% of the searches containing unnoticed spelling errors. Two versions of the system were installed on alternative days on two terminals at the library, with about 120 users interviewed over a one month period. The two versions were a control (CTL) system with weak stemming (including an "s" stemmer and "ed" / "ing" verb stems), and an experimental system (EXP) with a strong stemmer (Porter), spelling correction and a lookup table for phrases and equivalence classes of related terms. The evaluation involved both observation/interviews and transaction log analysis. The observations were mostly to note problems and to collect beginning and end session times, with the interviews asking about frequency of catalog use, what was the target of the search, the success of the search and any problems encountered. The transaction logs allowed the experimenters to repeat the exact (initial) search, on either of the two test systems or on a system with no stemming (the original system). It is worth noting that this two-stage approach was deemed necessary because earlier pilot studies had shown that most searches would return the same documents regardless of the system and therefore a direct comparison of the two systems (such as that done for the CITE system at NLM) would not have been sensitive enough to find differences. Indeed the results bore this out; even though the weak stemming retrieved more records in almost half of the initial searches repeated from the transaction logs, it "rarely turned a search from a complete failure into a success", and the strong stemming hurt performance as much as it helped.

A third set of experiments in 1989 [194] tackled a much more difficult user scenario, where the goal was to examine the effects of relevance feedback. The 1987 system (with weak stemming and no spelling correction) was the base system, with a second system (qe) offering query expansion (look for books similar to), and a third system (full) which also included a "shelf browsing" system based on the Dewey class number order. For reasons related to the logistics of dealing with the data underlying the OPAC system, it was impossible to set up true operational testing, and a laboratory experiment with over 50 subjects was done instead. The task was to build reading lists for specific essay questions and the subjects were asked to use the base system for 15 minutes and then use either the qe or the full system (brief demonstrations of these systems were also done before use). The data for analysis was the transaction log and transcripts of recorded interviews. Based on the interviews, the qe (query expansion) system was considered highly acceptable, with the full system (shelf browsing) much less so. More than two-fifths of records based on query expansion were chosen by the users, with only one-fifth of those from the full system chosen.

Later in 1989 the Okapi system moved to the Centre for Interactive Systems Research at City University, London and a second set of experiments using relevance feedback was done [77, 195], this time within an operational setting using both the City catalogue and a section of the INSPEC database (computer science and information technology). A terminal with Okapi query expansion was available for six months, with searches logged and 120 user interviews done in the last three months. Query expansion was offered only when searchers had chosen at least two items as relevant; this happened in 43% of the logged searches and expansion was used in 31% of those searches, or about 13% of the total searches. For over half of the time it was used, no additional items were selected by the users. Post search interviews revealed that 41 out of the 45 users that did not use query expansion had found all the books they wanted already. However replays of those searches did show that there was indeed more to be found 50% of the time.

What were the results of these experiments? Clearly the best match ranked results from the first experiment were better than the Boolean results, although there could be little actual user analysis. The weak stemming made a small improvement in some searches, with the spelling correction helping in only a few. Relevance feedback worked well for the particular task chosen in the laboratory experiment, but was not heavily used in an operational setting. Note that these experiments were the first time that search methodologies proven in a batch mode were then evaluated in an operational setting (with the exception of the single CITE evaluation by NLM). As such they also illustrated the difficulties in transitioning (and then evaluating) these technologies to real world settings. The differences between two techniques have to be noticed by users, the specific tasks being examined have to "require" these new techniques, and the huge amount of noise involved in user studies (variations in users, topics, etc.) can often swamp any significant results. All of these issues have continued to plague user testing, as will be seen throughout the rest of this chapter.

In addition to these early user experiments, considerable discussion (and work) was being done in user modeling, in particular looking at the meaning of relevance, a concept that is central to evaluation of information retrieval. The batch experiments, starting back with Cranfield, all carefully

defined relevance, and also made simplifying assumptions in order to operationalize the making of relevance judgements. These simple assumptions were criticized heavily, along with much theoretical discussion and some user experiments. One of the central figures looking at relevance was Tefko Saracevic, who started writing about relevance in the 1970s. A recent set of papers by him [153, 154] not only reviewed his work but surveyed other past work in relevance, both in the theoretical sense (the first paper), and in a review of the user studies (second paper).

Saracevic (and others) pointed out that the simplifying assumptions made about relevance in the batch evaluations did not reflect how users viewed relevance in the real world. The batch relevance judgments were based solely on topicality, they were usually binary, each document was considered independently, and it was assumed that judgments did not change over time (or across users). This was not meant just as a criticism of the batch evaluations (which clearly were doing useful research), but as a spur to further user experiments to better understand how users actually viewed relevance. Saracevic [154] then described various user studies that had been done to examine each of these assumptions.

Additionally he summarized a series of studies that looked at various relevance criteria, divided into the following categories:

- Content: topic quality, depth, scope, currency, treatment, clarity
- Object: characteristics of information objects, e.g., type, organization, representation, format, availability, accessibility, costs
- Validity: accuracy of information provided, authority, trustworthiness of sources, verifiability
- Use or situational match: appropriateness to situation or tasks, usability; value in use
- Cognitive match: understanding, novelty, mental effort
- Affective match: emotional responses to information, fun, frustration, uncertainty
- Belief match: personal credence given to information, confidence

These relevance criteria illustrate the huge variation that can be seen in how users approach information seeking, and in how they judge/value what is found. This variation continues to widen as information seeking has migrated from the simple OPACs to the vast amount of information available both on the web and through various social media such as Facebook. It is further multiplied by the explosion in the number and characteristics of users (far beyond the graduate student populations usually recruited for user studies!!).

3.3 INTERACTIVE EVALUATION IN TREC

Whereas TREC started in 1992 as a batch evaluation program only, there was always interactive involvement because the results were allowed to be created manually in addition to automatically

(the creation method needed to be declared). By TREC-3 this was also formalized into an interactive track, where the 50 routing topics were used with the goal of creating the optimal routing query. There were four groups that took part, with results showing that humans could not beat the automatic machine learning methods used in routing (or conversely that the machine learning methods were indeed very powerful), but more importantly how difficult it was to evaluate interactive methods in the same manner as batch methods. Some of this comes from the "normal" difficulty in evaluation of interactive systems, but use of the TREC methodology added major problems, including the issues of which results to select for submission and how to deal with the natural disagreement between the "users" and the TREC relevance assessors. For an excellent summary of these problems (and the TREC interactive track in general), see [67], and for more on the various individual results, see both the individual online TREC proceedings and a special issue of Information Processing and Management [94].

The TREC-4 interactive track was more specifically designed for the interactive research community, with 25 of the ad hoc topics being used and two tasks created (find and save the most relevant documents in 30 minutes or create the "best" final query). This helped avoid some of the earlier problems but still left the issue of how to deal with the relevance assessment differences. Nonetheless 10 groups took part, including three commercial search engines, exploring such areas as graphical interfaces, visualization, etc. These papers (included in the TREC-4 proceedings) serve as interesting examples of user studies because they share the common thread of the same task(s). Additionally TREC-4 emphasized the analysis of the data and the search process by specifying what kinds of data should be collected, such as search logs, timings, use of various search features, etc.

Despite this general success, the interactive research community wanted better ways of comparing their systems within a task that was realistic for interactive searching. The TREC-5 task was done on yet fewer topics (12 old ad hoc ones), but with the goal of finding relevant documents that cover different "aspects" of the topic. The evaluation of the TREC-5 interactive track was separate from the main ad hoc task, with a different pool created from all of the documents submitted for the aspectual task. This pool was used to create lists of unique aspects for each topic, including which documents contained these aspects, and interactive groups were scored against these lists. Systems were asked to submit detailed event logging, including query terms, documents judged relevant, etc. and a detailed narrative account for one search, as once again the emphasis was on the search process. TREC-5 also featured a new experimental design [113] to allow direct comparison of results across participating sites. The twelve topics were divided into four blocks, and a common control system (NIST ZPRISE) was provided. Participating groups also tested their own system and the common experimental design dictated the order of the topics/systems to be searched. The goal was to be able to compare the systems by measuring users' performance on the common control system and on the site's experimental system. Only two groups took part in TREC-5, but the same task (with only six topics) was done by nine systems in TREC-6. An analysis of variance model was used to study the effects of the topics, the users, the systems and the various interactions [113]. It was found that there were significant effects from all three factors, with the topic effects being the largest.

Note that whereas the TREC-6 effort represented the first true interactive cross-site comparison, several issues remained troublesome. There were doubts as to whether the common control system method fully removed site-related differences, with a few attempts made to validate this [170]. Equally important, there were serious concerns about the logistic requirements both to "waste" time on an "uninteresting" control system and to use so few topics (given the known heavy effects of topic differences). For these reasons, the common control system was dropped for TREC-7, with each of the eight participating sites substituting a control system of their own, but keeping the common experimental design. Although this did not allow strict comparison across systems, it did allow each site to make a clean comparison to their own control system (unfortunately the differences were generally nonsignificant). The aspectual task was continued, with eight ad hoc topics specifically tailored for aspectual retrieval. The measures used were aspectual recall, aspectual precision, and elapsed time for the search. TREC-8 continued the same task, for six topics and seven sites, with generally the same evaluation issues. It is important to note that whereas the evaluation issues remained the same, each of these TRECs provided a common task and common measures that allowed interesting user studies to take place and to be informally compared among the systems. TREC-9 focused on a question-answering task, with eight challenging questions that users had to answer in five minutes, including providing the supporting answer document. The scoring was based on getting an answer plus the elapsed time. Five minutes was probably too short for this task for some topics , but again there were interesting and useful results from this task.

For TREC-10 it was decided to emphasize observational studies as opposed to system comparisons, with the goal of defining a new task for TREC-11. The public web was the data to be searched, and a common focus was provided by picking four search tasks (finding medical information, shopping, travel planning, and research for a paper), with eight topics spread across these areas. This general task was continued in TREC-11, this time using the TREC gov collection rather than the public web. An optional common search engine (Panoptic) was provided, with the task sharpened to require either short answers or a good web site for the answers. Additionally the experimental design developed earlier for TREC-6 was used, again allowing for within-system comparison of results. This task was continued in TREC-12 as a subset of the main web track, where the task was topic distillation and eight of the web topics were modified for better interactive searching.

TREC-12 (2003) marked the end of the formal interactive track in TREC. Just as the batch-oriented TREC ad hoc tasks and test collections had been useful in improving the underlying search engines, the TREC interactive track came along at a time when user studies were migrating from the limited OPAC system work to the much larger web. In addition to all of the results from the user studies, much was learned about what kinds of (TREC) tasks were most useful for system comparison, what measures were informative, and what kinds of common data were interesting to collect. Additionally a control system/experimental design methodology was developed, modified, and thoroughly tested. However there was disappointment in the general failure to detect significant differences between methodologies. Dumais [67] concluded her comments on the track by saying "It is always difficult to interpret the failure to find significant effects–it could mean that there are no

effects or that there is still sufficiently high variability, making it difficult to detect all but the strongest effects. The continuing, strong desire for more experimental power to find significant effects, when they exist, requires a reduction in the variability or an increase in the number of searches. This might be accomplished by either increasing the number of tasks per searcher or focusing on subtasks."

Because of the logistics and expense of adding more tasks or searchers, the interactive community in TREC chose to move to a specific subtask. In particular the track looked at the effects of having more metadata about the user or topic, including the use of simple interactions, in order to improve early precision results (the High Accuracy Retrieval from Documents or HARD track). The participating groups first submitted their results based solely on the standard input topic (similar to the ad hoc task), but then had the opportunity to use additional information from the user (in this case the assessor who built the topic and also would make the final judgments). This information could take the form of metadata contained in the topic, or could be answers to a clarification form that the assessor completed. The second stage was then a submission of a new ranked list modified by the metadata and/or the interaction.

The 2003 HARD track should be considered a pilot, although interesting research was done in how to choose documents to submit for quick assessment (the simple interaction part). The 2004 HARD track [8] used 50 topics, all including metadata built into the initial topic. In addition to the standard TREC topic fields, there were new fields for *metadata-narrative*, explaining how the metadata is intended to be used, *retrieval-element*, specifying whether a full document or a passage was wanted, *familiarity*, *genre*, such as news-report, opinion-editorial, etc., *geography* (U.S. or not), *subject-domain*, *related-text.on.topic*, and *related-text.relevant*. The relevance judgments for the results were done at three grades: nonrelevant, hard relevant (relevant to topic and satisfies metadata), and soft relevant (relevant but does not satisfy metadata). Additionally if the granularity requested was at the passage level, then passage relevance judgments needed to be made (passages were defined by a document offset and their length). In addition to the metadata provided, the participants could submit a clarification form needing 3 minutes or less for completion (most groups used this form for quick judgments of keywords, documents or passages). In general the use of the metadata did not help results, often because the topics were such that the metadata did not provide useful additional information (a similar issue occurred in the elaborate TRECs 1 and 2 ad hoc topics). However some of the groups were able to obtain improved results using clarification forms. Note that several new metrics for passage retrieval were developed for this track [193] that were later used in the genomics track and in INEX.

The final running of the HARD track in 2005 had to be more constrained (no metadata and no passage judging) because of funding issues, however the clarification forms were continued and allowed to be more complex (but still only require three minutes to complete). Most systems were able to improve their baseline runs, but with a wide variation across topics as to which types of clarification worked best [9]. The HARD track morphed into a subtrack (ciQA) of the Question-Answering track in 2006 and 2007 [63]. The tasks in 2006 were similar to the HARD track, but here using the clarification forms for complex question-answering about relationships. The 2007

variation actually allowed the NIST assessors to interact with the participants' systems, with a limit of 5 minutes interaction time in which the participants could gather information. The basic goal of ciQA was the same as HARD in that groups hoped to get improvements in their baseline runs from these interactions. This did not happen consistently, partially due to some groups simply not using effective interactions, but also to problems with the task setup (assessors are not "naive" users after they have built a topic, had multiple interactions with clarification forms, etc.).

The final morph of the HARD track was as part of the current legal track. Starting in 2008, the interactive task [91] has used the Enron data and mirrored the actual legal discovery task where there is a "lead" attorney or topic authority who is the sole person who defines the task. There were 7 topics developed for this interactive task, and the participants were encouraged to interact with the topic authority (up to 10 hours worth of time) to clarify relevance issues, although other volunteer legal professionals made the relevance judgments (which could be appealed to the topic authority).

3.4 CASE STUDIES OF INTERACTIVE EVALUATION

TREC of course was not the only place where interactive evaluation was occurring; there were many other experiments. This section covers some of these, selected both to cover specific themes and to illustrate different issues/pitfalls in terms of higher level experimental design. They are organized by theme rather than chronologically.

The first theme is really a continuation of work done with the OPACs, i.e., do performance improvements seen in batch evaluations hold when tested interactively. There have been a series of papers dealing with this topic, in particular examining if the TREC batch system improvements resulted in improved performance for users. Hersh et al. [95] showed that better systems did lead to better interactive performance, however the differences were not statistically significant. This was quickly followed by a second paper (Turpin and Hersh [178]) investigating why this result occurred. Since the first paper used only six topics with 25 users and one task (the TREC-8 aspectual retrieval task), the second study also included the 8 questions from the TREC-9 question-answering task. Two different retrieval systems, a baseline system (tf*idf) and an "improved" system (Okapi), were set up using the same interface for each of the two tasks. There were 25 users for the TREC-8 instance task, who spent 20 minutes trying to find at least one document for as many aspects of the topic as possible. The experiment used appropriate randomization as to topics and systems and basically found that although the Okapi system had a 15% better instance recall, this only came from one of the topics out of the six. A second experiment was run using the TREC-9 question-answering task, with 25 users who had 5 minutes to find correct answers. This time the "improved" system actually had a slightly poorer performance (-6%).

A detailed analysis was then made to find out what was behind these results. Since the logs contained all the queries asked and documents found or saved, the researchers were able to trace how the systems performed with the actual user queries (instead of the TREC topics). They found that on average the users had indeed seen better results (precision at 10 was 55% higher, instances at 10 was 105% higher) with the Okapi system, and therefore could issue fewer queries. However, these

improvements did not show in the final results because the users could find all the necessary relevant simply by issuing more queries. There was also a sufficient number of relevant documents available so that 30% to 55% of the relevant documents were not even read by the users. "It appears that the extra benefit of the improved system was ignored in the question-answering experiment and at best played a small part in the instance recall experiment" [178]. An unexpected complication was that the baseline system tended to retrieve shorter documents, therefore cutting reading time. However at least for these two tasks, users seemed able to do the task well with either system.

One of the issues with these two papers is that only a small number of topics were used. Whereas it is critical to have enough users in a study to detect significant differences in results, having a small number of topics makes generalization of the results difficult. This illustrates a major dilemma for interactive retrieval experiments, where logistics/expense usually force a choice between large numbers of users or large numbers of topics, even though the performance variation seen across topics is just as large (or larger) than the variation across users. Azzah Al-Maskari et al. [5] did a study with only 56 users but 56 TREC topics in a task to find as many relevant documents as possible in 7 minutes using two different systems. Because systems have a highly variable performance across topics, the systems in this experiment were selected differently for each topic. Given the results from three "comparable" systems for a given topic, the "good" system was the one with the best average precision and the "bad" the worst . Table 1 in their paper shows significant user performance differences in the time taken for the task, the number of relevant documents found, and user satisfaction, easiness, etc. between the good system and the bad. The second part of their experiment grouped the results into four topic sets based on the size of gaps in batch performance, and here they were able to show that as the performance gap grows smaller, so do the differences in user performance, with the gap needing to be at least 30% in order to see consistently better performance by users. Note that these experiments considered documents relevant only if the user judgments matched the TREC relevance judgments. If only user judgments were used (ignoring the TREC judgments), there were still significant differences with all 56 topics but these disappear when fewer topics were used.

A recent paper from SIGIR 2010 provides yet another experimental methodology for comparing batch results to user results. Sanderson et al [148] performed a crowd-sourcing experiment with 296 Turkers, who were shown paired results from the diversity experiment in the TREC 2009 ClueWeb track. Topics where ALL of the 19 runs in the diversity track had found two or fewer relevant documents were removed, leaving a total of 30 topics in test. The pairs of runs were selected from the 19 runs submitted to TREC by picking those runs with the same number of relevant in top 10 documents but a minimum difference in average nDCG of 0.1. The users were shown the top ten results (title, snippet and URL) and asked which set of results they preferred based on one of the subtopics of the full diversity topic (using the full topic had shown inconsistent results in pilot study). Each pairing was seen by an average of 8 Turkers and the votes were cumulative for each system of the pairs. The level of agreement between the votes and systems' nDCG performance was significant, even at smaller gaps.

It is obvious from this discussion that whereas conclusions vary as to whether batch improvements translate into user improvements, these conclusions are likely to be dependent on many factors in the experimental design. One factor is the number of topics examined; another is the gap in performance between the systems being compared. Maybe most critical are the specifics of the user task, such as the time allotted for task, the amount of interaction allowed, the instructions to the users, where the judgments come from (the batch experiment or the users' themselves), etc.

This leads into the second related theme and that is the general area of user modeling, where user modeling here means studying how a particular group of users behave as opposed to modeling a single user. These studies range from highly controlled laboratory studies where users are exposed to a minimum number of different variations to wide open observational studies, possibly even in an operational setting. The study by Kelly et al. [112] is an example of a very tightly controlled study. Her goal was not only to investigate how differences in system performance affect users, but also to develop an experimental methodology that could be used to isolate specific issues in how users actually search. She used a total of 81 subjects working with four topics from the TREC 2005 HARD track to investigate whether the ranking of the relevant documents or the total number of relevant documents in the top 10 ranks made a difference in users preferences of systems. The users searched each of the four topics on each of four "search engines" (this was appropriately randomized) and were asked to rank those engines based on their preferences. They entered a single query, read the titles and the documents and made relevance decisions. The four search engines were not actually search engines were ranked lists of documents in which the ranks or the number of the relevant and non-relevant documents were manipulated. For the first two studies, there were exactly five relevant and five non-relevant documents, but they were ranked from best (first five relevant seen first) to worst (relevant at ranks 6-10), with two intermediate cases. The third study had different numbers of relevant documents, ranging from 6 to 3, but controlled for order. The users were asked to vote their preferences after each search and could change these preferences after they had seen all four search engines (few did so). Note here that the relevant documents were carefully selected by the experimenter to be documents that would likely have a high-degree of agreement between the TREC assessors and the users (90% was observed). Detailed statistical analysis of the results showed that improved ranking resulted in significantly higher user preference, however the total number of relevant in the ranked list was more important than the ranking.

A study by Smith and Kantor [158] kept tight control on the "systems" but much less on the users because the goal was to observe more natural user behavior. There were 36 users who each completed 12 searches using either a standard system or one of two degraded ones. The standard system was the normal results of a Google search, with the other two being the same list but with degraded rankings. The CLR (consistently-low-ranking) one had the ranked list starting with those documents found at rank 300 and beyond, whereas the ILR one was inconsistent, with different starting points in the ranking for each query. All users searched on the same topics and were told there was a bonus for the person who found the most good information sources and the fewest bad sources. As queries were entered, the top 20 documents were displayed in the typical Google

interface, but the links were disabled so that users had to make their decisions based on the titles, etc. The topics and users were carefully grouped into three blocks (appropriately balanced) to enable detailed statistical analysis of the results. The first and third blocks were searched using the standard (good) system, with the middle block using the two different degraded rankings. Users submitted their results to the researcher who graded each selected document as either a good, marginal or bad information source for the topic and the four evaluation measures were the average numbers for these counts plus the search time. The basic results show that users did just as well with the degraded systems as with the good ones by adapting their behavior, such as by issuing more queries (it takes less time to eliminate bad results) or by not resubmitting the same query to bad systems. Some of the same issues seen in the Turpin paper discussed earlier apply here, in that there could be many good sources in the good system, however by lessening the control of the users, more natural phenomena could be observed.

An example of an operational user study with minimal control is a 1992 study by Su [169] investigating what types of measures were correlated with user satisfaction. The 40 users (library patrons) were recruited as they came into the library to search for information. They then sat with one of six search intermediaries while the search was done (including paying the cost of the search), and were given or sent the set of documents retrieved in order to make relevance decisions. This represents a complete operational setting (at least in 1992). Almost 50% of the searches were for PhD research or other such high recall areas, with the rest being class assignments, grant proposals, etc. The complete search process was documented, including the time taken for the search, and the time to make relevance decisions, and there was a 60-minute interview asking the users about satisfaction, why they were satisfied, etc. Table 1 in the paper lists 20 different measures that were checked for correlation with satisfaction, with the highest correlation being the users' perception of the completeness of the search results (recall). In general precision was much less important, either because users were satisfied with only a few documents or because they were willing to look longer if they felt there were likely to be few relevant.

Another subject for user study involves investigation of how user groups tackle different search scenarios. Bhavnani [25] studied how domain specific knowledge affects searching, using 8 tasks adapted from the TREC 2001 interactive track (four tasks in healthcare and four in online shopping). His users were five healthcare search experts from medical libraries and four students who claimed three or more years of online shopping experience. This was an observational study, using think-alouds and interviews to collect the data. Analysis of these studies showed that the users employed different searching techniques when they were in their expert domain than when outside that domain, with generally better results in their domain. For example the shopping experts found cameras for $60 less than the healthcare experts, but the healthcare experts found the 9 categories of people needing flu shots by going to an average of 3.7 reliable sites whereas the shoppers went to 12 general purpose sites and none of them found all 9 categories. An important part of this study was to observe the searching process to determine the characteristics of domain specific searching. They found that both groups knew what types of websites to visit in their domain, including specific

URLS to search, and were also familiar with the internals of those sites so that they could efficiently and effectively search them. A log analysis study on this same topic is presented in the next section.

The final two studies presented on user modeling deal with comparing searching via social sources (such as friends, social networks), and searching with search engines. The two studies differ in the level of control but also in the goal of the study. Morris et al [124] did an informal experiment comparing using Facebook vs. using a search engine to find answers to an information need. The 12 participants brought their own search needs, most of which were relatively straight-forward requests for advice such as "Any tips for tiling a kitchen backsplash", or "Should I wait for ZuneHD or buy iPod touch (to gift someone)?". They then sent this request out to their Facebook friends (they needed to have at least 50 friends) and started their own searching. When they had finished searching, they checked with their Facebook network and captured a screenshot of the content and timestamps for any responses; this was repeated three days later. In general they got better information from the web, however social networking had added benefits, including sometimes different answers.

In contrast, Evans et al. [68] used two laboratory tasks, both of which were research questions about energy. Their questions were specifically picked not to be advice type questions, such as normally posted to Facebook, and were questions that are complex to search (one question was "What role does pyrolytic oil (or Pyrolysis) play in the debate over carbon emissions?"). Additionally they looked at all types of social search strategies, including calling or emailing friends, all the various social networking sites and other "socially-generated" sites such as question-answer sites, and blogs. They had 8 participants who worked in two blocks in a balanced experiment, with one block searching with social methods and the second using traditional search engines, databases, and Wikipedia, but not any of the "social" sites. There were semi-structured interviews, followed by detailed analysis of the various strategies that were used. One of their findings was that the results were very topic dependent, with the highly technical topics generally doing less well for social search strategies.

The last theme in this section is a continued look at relevance and the characteristics of relevant documents. The earlier work had begun looking at end users in libraries, but the information seeking tasks today are much broader. Barry and Schamber [21] compared two user studies that had the goal of finding relevance criteria. A more traditional one was done with 18 faculty and students who submitted a request for an online search and when presented with results (including full text documents, abstracts, citations, etc.) were instructed "to mark any portion of the materials that indicated something the respondent would or would not pursue." This was followed by open-ended interviews to collect various relevance criteria (989 responses to 242 documents). The other study was done within a specific domain with 30 users of weather information in areas such as construction, electric power utilities and aviation. In this case the information would be used to make planning decisions and they were asked to create time-lines of information seeking events leading to these decisions. The study interview then looked at three of these events, with the participants asked to evaluate the sources and presentation modes of these sources. In this case the "criteria were operationalized as ways in which sources or presentations made a difference to respondents in their

situations." The study then looked at criteria that were common across the two studies, and ones that were different. The common ones were divided into ten categories:

- Depth/Scope/Specificity: extent to which information is in-depth or focused; is specific to the user's needs
- Accuracy/Validity: extent to which information is accurate, correct or valid
- Clarity: extent to which information is presented in a clear and well-organized manner
- Currency: extent to which the information is current, recent, timely, up-to-date
- Tangibility: extent to which information relates to real, tangible issues; definite; proven information is provided; hard data or actual numbers are provided
- Quality of sources: extent to which general standards of quality or specific qualities can be assumed based on the source providing the information; source is reputable, trusted, expert
- Accessibility: extent to which some effort is required to obtain information; some cost is required to obtain information
- Availability of information sources: extent to which information or sources of information are available
- Verification: extent to which information is consistent with or supported by other information with the field
- Affectiveness: extent to which the user exhibits an affective or emotional response to information or sources of information; information or sources of information provide the user with pleasure, enjoyment or entertainment

Criteria that were different for each study "appear to be due to the differences in situational contexts and research task requirements: specifically, control for source type in the Barry study and control for topic in the Schamber study [21]".

Other studies have looked how relevance criteria have changed in the web searching arena. Tombros et al. [177] used three different tasks and asked 24 participants to indicate the features of the web pages that were important to them for relevance criteria. The researchers were interested not only in what web document features were used in assessments, but also how the different tasks affected that choice. There was no control on the searching (other than time), and a think-aloud process and questionnaires were used to capture information. All tasks were set within a search scenario and were of three types: a background search, a decision task, and a listing task. The pre-search questionnaires looked for familiarity with task topic, etc., while post-search questionnaires asked about the clearness and easiness of task and the participants' perception of the importance of certain aspects of the web pages they viewed (both positive and negative aspects). It was found that most useful aspect of the web sites was the text, including the content and the numbers (especially

for the decision task), followed by titles/headings and query terms. Another important criteria was the perceived quality of the web site, including its scope/depth, authority/source and recency. There was a difference across tasks in the importance of some features including a limited use of pictures for task 1 (the background search), the increased use of numbers for task 2 (the decision task) and the increased use of links for tasks 1 and 2. "As far as the scope/depth feature is concerned, its increased use in a decision task is based on that users required enough information in pages (e.g., enough details about prices, specifications, guarantees, availability of speakers, etc.) in order to make an informed comparison of the available choices [177]."

A crowd-sourcing experiment [11] also looked at relevance criteria on the web, including e-commerce tasks. There were 83 needs taken from the most frequent queries in Yahoo Buzz and Amazon product searches for two weeks. Thirty-five of them were e-commerce searches, whereas forty-eight were classic needs looking for items such as IRS tax forms, government jobs, or health insurance. For each need, there were 17 available criteria, with eight of them adapted from the work by Barry and Schamber [21] and nine more added specifically for the e-commerce task. The experiment used crowd-sourcing (2450 results), asking Turkers to select one or more criteria for these needs. Figure 1 in that paper shows that accuracy and validity were most important for both sets of needs, with availability coming in next (higher for e-commerce as would be expected). Differences can be seen in the two tasks, with more importance for depth/scope in the non-ecommerce needs, and price/value more important in the e-commerce ones.

The user studies in this section differ in the amount of control being exerted during the experiment, and in the number of users/topics/tasks addressed in the study, with most of the groups preferring to study more users and fewer tasks. This decision allows tighter focus in the analysis, such as in the study by Tombros et al. with only three tasks, where they were able to observe how the different tasks affected relevancy decisions. Note that in this study, and also in the Bhavnani study on domain-specific searching, the details in the higher-level experimental design (such as what tasks to pick, what to observe, etc.) were critical in getting useful final results. The two crowd-sourcing studies were able to have many users, and therefore were able to have more confidence in their end results, but could be successful only because they carefully focused the tasks (and set up the crowd-sourcing correctly). It is also interesting to contrast the tightly controlled study by Kelly et al. with the Smith and Kantor one, with one being able to use tight control to learn "micro" analysis of how users evaluate ranking and the other able to observe how users adapt to badly ranked results. There is no "right" way to design experiments, including how much control to use, how to balance the number of users versus the number of tasks, etc., but the key to the success of these various studies is that they selected clear, highly-focused initial goals and then determined which issues were most important to control based on these goals.

3.5 INTERACTIVE EVALUATION USING LOG DATA

One way to study many users is with the log data collected by the various search engines (and other places). Unfortunately this data is not usually publicly available, however research be-

ing done within the search engine companies provides some interesting insights into web usage. This section looks at some of that research and at some of the issues in dealing with web logs. A valuable resource for this work is the recent tutorial by Dumais, Jeffries, Russell, Tang and Teevan (http://research.microsoft.com/en-us/um/people/sdumais/Logs-talk-HCIC-2010.pdf). This tutorial starts with listing the advantages of log analysis, including that the users' actual behavior is recorded as opposed to being reported or recalled or being the subjective impressions from laboratory experiments. Additionally the large (huge) sample permits subdivision of the data to any level of resolution, allowing for very focused experimental design. The disadvantages of the log files is that they are not controlled, not annotated with "macro-events", and there is no method of understanding why some micro-event has happened. A big challenge also is that these logs are massive, requiring serious efforts in focusing the experiment, in cleaning the data, in partitioning the data, and in understanding and interpreting the results. The papers presented in this section serve as illustrations of how this can be done successfully.

One of the first issues in log analysis is to decide what basic events/measures/counts to track. A workshop at the 2003 SIGIR conference [66] looked into this issue and one of the results was a paper by Fox et al. [70] detailing methods of collecting and then using the implicit user behavior available in web logs. The paper starts with a description of the embedded add-in for the client that allowed detailed tracking of users. The goal of the study was to collect both explicit feedback and implicit feedback in order to allow correlation of the various measures, and 146 internal Microsoft employees volunteered for the 6 week study. Two types of explicit feedback were gathered, including asking about the "relevance" of each individual search result visited ("liked it", "interesting but need more information", "didn't like it", "didn't evaluate it") and session level evaluation ("is this a new search", and "what is your level of satisfaction with the old search"); both of these were handled by system prompts based on the user actions. Tables II and III in the paper list 30 different implicit measures/counts made at both the result level (time spent on page, scrolling count, time to first click, etc.) and session level (number of queries, number of results returned, end action, etc.). The rest of the paper deals with statistical methods to combine the implicit feedback to predict the explicit feedback, proving that appropriate combination methods outperform simple uses of the implicit measures.

The number of implicit measures and features that are collected, and the sophistication with which these features can be analyzed, has exploded since then, with recent papers displaying large (and different) samples of the features now used. One example is the Teevan et al. [175] paper investigating the diversity of information needs behind similar queries and looking for methods that could be created to aid users in getting better results from these "ambiguous" queries. This study used a large sample of 44,002 distinct queries, each input by at least 10 different people. The types of features examined include some that are based only on the query itself (such as query length, contains URL fragment or time of issuance), some that need the result sets (such as the clicks, the number of ODP (Open Directory Project) categories or the portion of the urls that end in ".com") and some that need multiple instances of the same query (such as the number of times the query

has been issued, the average number of ads displayed or the average number of results). The paper discusses ways these features can be combined to create models that improve the results of these types of queries. For example, query length and the use of URL fragments in the query help distinguish between navigational queries and informational queries.

The use of click data was the subject of a workshop at the Second ACM International Conference on Web Search and Data Mining (WSDM 2009), where a large query log from Microsoft was made available to researchers (14.9 million entries from one month in 2006, including 6.62 million unique queries). One of the studies [54] using this data looked at how click data could be combined with data from other sources to identify "diverse" queries (similar to the ambiguous queries in the Teevan et al. paper). One of the measures used was click entropy, which measures the spread of search results that are clicked on by multiple issuers of the same query (higher click entropies mean that users clicked on many different results). This click entropy was measured across the collection, showing that queries with low entropy accounted for 80.2% of the queries and that 95% of these returned names of organizations. The researchers used a subset of the data (queries with 50 or more repetitions and with five or more clicks) to investigate the correlation between click entropy and various other clues to ambiguity. They found little correlation between the query terms having many word senses (such as measured by WordNet or Wikipedia) and click entropy, however there was a positive correlation between the size of a Wikipedia article and click entropy, indicating that high click entropy might indicate broad topics with a need for aspectual retrieval.

The tutorial on log analysis also discussed the need to properly partition the logs based on the particular goals of a given study, such as partitioning by language, by time, and by the type of device being used to access the web. The study by White et al. [197] looking at how domain expertise affects the way people search the web is an excellent example of careful data partitioning. They started with 900 million browser trails from 90 million search sessions collected over a three-month period. The search sessions were then partitioned (by automatically classifying the visited pages using the ODP) into sessions where the users had (mainly) searched in one of four domains (medicine, finance, legal and computer sciences). They then selected the target group of users from this subset as those who had viewed 100 or more pages in these domains and whose page views contained 1% or more of domain-related pages. However this set of users are not necessarily domain experts, and the last partition was to identify as experts those users in this set that had visited PubMed, online financial services, Westlaw, or the ACM Digital Library, noting that three of these cost money and are not liable to be used casually. Once the users had been divided into experts and non-experts, the various other features of searching could be correlated with the two groups, showing for example that experts searched longer, issued more queries and visited more pages in unique domains in a given session.

In addition to careful partitioning of the logs, often there also needs to be "cleaning" of those logs, such as removal of spam, robot visits, and allowances for anomalies such as data drops, capped values, or "censored" data. This is particularly important in studying patterns, such as the study by Adar et al. [2] on web revisitation patterns. The goal of this study was to characterize how people revisit web sites using a 5-week web interaction log, followed by a user survey to identify intent.

The log was filtered to include only users with data for over 7 days, and a series of "outliers" were cleaned from the data such as users in the top 1% of activity (measured in different ways) who were assumed to be robots or other "badly behaved" users. This still left 612,000 valid users, and their visits to 54,788 URLs became the basis of the revisitation study. These URLs were binned by different criteria such as number of unique visitors (4 bins), median per-user revisits (5 bins) and inter-arrival times for a total of 120 possible bins to examine for usage patterns. Interesting clusters could be seen in these patterns, such as fast revisits could be porn or spam sites, with slightly different patterns indicating shopping or reference sites, whereas "medium" revisits could mean portals such as bank pages or news pages. These various patterns were then verified by a user study with 20 volunteers who were tracked for some of these URLs and also surveyed on a small selection of their revisits.

This final section illustrates the power of log studies to pinpoint and verify user behavior. In each of the cases there was a clear focus before the experiment on specific goals and this provided the researcher with a "gameplan" to analyze the data. Because the experiments basically started with a specific problem to solve, it is likely that the results can then be incorporated into new systems for the commercial search engines.

CHAPTER 4

Conclusion

4.1 INTRODUCTION

This lecture concludes with some thoughts on how to design an experiment, pulling together some of the ideas from earlier chapters. Additionally there is discussion of some very recent issues in evaluation, both in methodology and in metrics, and a personal look ahead at some future challenges.

4.2 SOME THOUGHTS ON HOW TO DESIGN AN EXPERIMENT

In 1992 Jean Tague-Sutcliffe wrote "The Pragmatics of Information Retrieval Experimentation, Revisited" [172], where revisited meant that she had written an earlier article for the Karen Spärck Jones 1981 book [167]. In the 1992 article she discussed the series of decisions that needed to be made before any experimentation could start, and most of these ideas are still valid almost 20 years later.

Her first decision was "To test or not to test", where she said "An experiment should have a purpose; it is a means to an end, not an end in itself. It is therefore essential that the investigator delineate clearly the purpose of the test, the addition to knowledge that will result from its execution, and ensure that this addition has not already been made." In these days of fast computers, readily available data such as the TREC data, and a push to publish, it is sometimes easy to forget this. But the most critical key to the success of an experiment is to have such a focused goal that the variables are "easy" to define and operationalize, it is "obvious" whether the experiment was a success or "failure" (either is OK), and the results can be presented in such a concise fashion that readers immediately understand what was learned. Note that this applies whether the experiment involves the study of search logs, an operational user study or using a test collection. The successful log studies discussed in Chapter 3 were strongly focused and therefore were able to draw conclusions, such as the one looking at the effects of domain expertise [197] from massive data files. The Medlars operational user study [115] had a clear goal even though it was necessarily less focused, and that goal drove the intricate design of the recall and precision databases to allow non-biased measurements of performance. TREC (and similar) evaluations are even less focused, with the data and the task defined, but the actual experiments up to the participants. Whereas this allows greater freedom, it also can lead to results where it is not clear what has been learned.

Another decision is how to operationalize the variables. The documents to be searched need to be identified, and the definition of a document for the particular experiment needs to be determined.

This may seem to be a simple task, however as information retrieval has branched beyond OPACs, the unit to be retrieved could be passages, blog threads, sections of video, etc. Again the goal of the experiment, and therefore the user application that is being modeled, *must* determine this. Evaluations such as INEX, TRECvid and the genomics track at TREC have done extensive work on this difficult problem.

Tague-Sutcliffe also discussed the issue of "information representation" because most of her documents were manually indexed. Whereas this does not seem to be a part of today's experimental scene, it is really a hidden variable. The search log experiments rely on many types of information, with new types continuing to be tried (see [175] for an excellent example of this). In test collection experiments the representation of the information may be encoded in the system, such as the use of language modeling, making it more difficult to understand the underlying effects of data.

The third variable Tague-Sutcliffe discussed was the users. Here she was looking particularly at user studies, and listed four types of categories:

- type of user –student, scientist, businessperson, child;
- context of user – occupational, educational, recreational;
- kinds of information needed – aid in understanding, define a problem, place a problem in context, design a strategy, complete a solution;
- immediacy of information need – immediate, current, future.

Note however that users are not just for user studies, they also are the basis of user models and therefore need to be considered as part of *all* experiments in the operationalization of other variables.

The fourth variable is the source (and format) of the user questions/queries/search statements. Here the user model is critical, determining not only the types of questions that are being used for testing, but also how those questions are collected/assembled/created. One of the important lessons learned in the CLEF cross-language evaluations was how to build questions that faithfully mirrored the way that native speakers would query in their language and also reflected the type of questions that might be asked within each language area [201]. Likewise the format of the questions needs to mimic how users might search within a given application; good examples here are the TRECvid questions and the work done with patents in NTCIR.

Another issue is how to measure the performance and how to analyze the data. This involves appropriate selection of metrics such that the user and the application will be cleanly modeled. It also includes (at least in this lecture) the choice of how to measure the correct answers or relevance judgments. These two issues are closely related, although not tightly coupled. For example the TREC ad hoc user model was a high recall user where relevance was taken to mean any document that could be useful in a report. Note that this would reasonably include "duplicate" information since an intelligence analyst (or a newspaper reporter) might well regard duplicate information as verification of some fact. It also meant that a document was either useful or not (relevant or not), implying binary judgments. Additionally the TREC ad hoc results were always considered ranked and therefore needed a recall/precision metric. However other high recall users could be patent

examiners, where a single patent would invalidate the application. Here the judgments are again binary, but the metrics could be those used for known-item retrieval, such as the mean reciprocal rank (MMR).

High recall can also mean high recall of different aspects of a given search, such as finding books by a specific author or finding recipes. This user model occurred in the various TREC interactive tracks, where aspectual recall was used, and in the genomics track. It also was behind the diversity track for the web in which the goal was to identify the various subtopics of the query. The definition of relevant there was slightly changed, incorporating the novelty of the information. Other possible evaluations could incorporate the order the information is presented, the perceived validity of the source, or the currency of the information, in fact any of the relevance criteria discussed in Chapter 3. Some of these definitions of relevance will require different or multiple metrics, such as graded relevance metrics like nDCG (see [46] for an example).

Web searching is usually considered low recall, where the goal is to get the user something useful quickly. This leads to metrics like success@1, or nDCG measured at the top 10 or 20 documents retrieved. The search engines all use graded relevance judgments, often with five or more grades, implying that they consider these fine grades important to users. Another issue involving web search is the huge number of potential relevant documents and this scaling issue has led to new sampled pooling methods (requiring new metrics) in the recent TREC web tracks (see Chapter 2).

Finally there is the critical decision of how to analyze the data (and what to present to the readers). Large tables of recall/precision averages showing small differences from baselines are not very interesting, and one could question what is learned from that. Use of statistical testing can be done (see for example [97, 150, 159]) but has not been heavily adopted by the community. A personal preference would be to actually look at the data, examining issues like how many of the testing questions showed improvement vs. lack of improvement, how large that improvement was, and most importantly WHY this occurred. The Turpin and Hersh paper [178] exploring why user evaluations did not give the same results as batch evaluations is an excellent example of this, and the failure analyses done in some of the TREC research papers contribute more to understanding of the underlying issues of retrieval than these large tables of averages. However plans for this type of analysis need to be incorporated in the experimental design early on to ensure that appropriate data is collected.

4.3 SOME RECENT ISSUES IN EVALUATION OF INFORMATION RETRIEVAL

A brief survey of recent SIGIR and CIKM proceedings shows three main categories of current interest in evaluation. The first of these categories is deeper examination of the implications of current evaluation metrics and methodologies.

Robertson [135] analyzed the Geometric Mean Average Precision (GMAP) used in the TREC Robust track and noted that the difference between GMAP and MAP "has to do solely with emphasis on different parts of the effectiveness distribution over topics". In particular, the GMAP

measure emphasizes improvements in topic scores where the initial score was close to zero, rather than giving equal emphasis to improvements across all the topic scores such as done by MAP. He expanded on this by adding that the use of these different methods of averaging (or indeed the use of different metrics) allow researchers to observe more clearly how their systems are operating (such as his comment that the use of relevance feedback becomes more questionable when measured using GMAP).

There have been several papers re-examining the use of statistical methods. Sanderson and Soboroff's poster at SIGIR 2007 [149] showed some problems with the use of Kendall's tau rank correlation method, where the issue was the dependency of the threshold on the range of scores in the set of runs being compared. Voorhees' poster at SIGIR 2009 [188] revisited the issue of how many topics are needed in order to show significant differences across systems. In this paper she noted that even with 50 topics, there will be a small number of results that will be falsely declared significant (well within the 5% probability of a Type I error) and that researchers should validate results on multiple test collections. A poster by Smucker et al. in SIGIR 2009 [159] compared different significance testing methods to investigate the effects of varying sample sizes on the results.

Another continuing topic for research is the quality/consistency of the relevance judgments and how this affects evaluation. This has been investigated starting back with the SMART system. Table 1 in a paper by Bailey et al. at SIGIR 2008 [16] not only summarizes this work but adds new observations comparing the CSIRO assessors (the gold standard) in the TREC 2007 Enterprise track to a silver standard group (science communicators outside of CSIRO) and TREC participants (the bronze standard), with the latter being shown as not as reliable. A recent paper by Carterette and Soboroff [44] used simulation to examine the effects of different types of assessors (optimist vs. pessimist) using the TREC Million Query collection; one of the goals of this paper was to speculate on possible problems using completely untrained assessors, such as in crowdsourcing.

A second category of recent evaluation papers involves better/easier ways of building test collections. Evaluations (such as in TREC) have shown problems with the very large collections, where it can be assumed that the judgments are not "complete" and therefore that the collection cannot be declared reliable in terms of reusability by systems very different from those used in building the pool. The Million Query track was started in TREC 2007 to investigate better ways to select documents for judging and Carterette et al. in SIGIR 2008 [43] presented an analysis of this work, concluding that it was more effective and efficient to judge fewer documents using more queries. A follow-on paper by Carterette et al. in SIGIR 2010 [42] looked at a methodology for designing reusable test collections at this large scale and then being able to validate (or not) their reusability.

Another way of dealing with incomplete judgments is to use metrics less sensitive to incompleteness. The bpref metric is one of these and He et al. in SIGIR 2008 [90] compared the use of different metrics (MAP, bpref, infAP and nDCG) in effectively training systems as the levels of completeness were varied (MAP was the least effective).

Test collections are very expensive to build, with the relevance judgments being the most costly. One option has always been to get participants in the evaluations to do the judgments, however this option has a mixed record in terms of dependability. Kazai et al. for the INEX evaluation tried a Book Explorer game [104] to gather assessments, investigating the use of incentives to entice players and to control for quality. The use of Amazon Mechanical Turk or other crowdsourcing platforms continues this work, with Alonso and Mizzaro's poster at SIGIR 2009 showing the use of Turkers to determine relevance criteria in e-commerce [11], and another paper [10] by the same group in a workshop at that conference asking if TREC assessors could be replaced by Turkers!! SIGIR 2010 had both a workshop entitled Crowdsourcing for Search Evaluation, and a paper by Sanderson et al. [148] using crowdsourcing to compare the effectiveness of different retrieval systems where Turkers voted their preferences of paired results.

The final category of recent evaluation papers involves better ways of "mirroring" the user in batch evaluations. Järvelin and Kekäläinen developed the nDCG metric [100] based on graded relevance judgments, and it has seen heavy use in commercial search engine evaluations. In 2008 Järvelin et al. extended the metric to handle sessions of multiple queries, clearly better modeling real user interactions with a system. This same theme has been continued in the TREC Sessions track started in 2010. Note that there has always been work on multi-query sessions in user studies, and also in the work with the search engine logs, however trying to simulate this in batch evaluations remains a challenge.

Another obvious poor modeling of users in batch evaluations are the simplistic relevance criteria generally used. The idea of novelty and diversity has inspired the diversity task in the TREC web track, and also a paper by Clarke et al. in SIGIR 2008 [46] proposing a framework for evaluation of novelty and diversity, including a new metric based on nDCG. Additionally there have been two recent workshops at SIGIR: "Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments" at SIGIR 2008 and "Redundancy, Diversity, and Interdependent Document Relevance" at SIGIR 2009.

A paper by Turpin et al. [179] at SIGIR 2009 suggested that evaluations could better model users if they included summaries of the documents in the judgment process and suggested a metric suitable to accommodating this. Whereas work on sessions has concentrated on the multiple query aspect in batch evaluation, the process of deciding "where to click" has only be investigated in user studies. A workshop at SIGIR 2010 ("Simulation of Interaction: Automated Evaluation of Interactive IR") further expanded on how users could be modeled for batch evaluation efforts.

The commercial search engine community has enormous amounts of logging data that can be exploited to better model users. Chapter 3 discusses some of the recent papers in this area. Additionally Radlinski and Craswell presented a paper at SIGIR 2010 [129] that examined the process of interleaving results from different search engines as a new way of evaluation without test collections. They showed that interleaving was better at detecting very small differences in retrieval effectiveness (note that small differences do add up) than a more traditional test collection methodology, at least at high precision tasks.

4.4 A PERSONAL LOOK AT SOME FUTURE CHALLENGES

Information retrieval has always placed a high importance on evaluation, and this has allowed great progress over the years in terms of improvements. However the downside of this is that areas that are difficult to evaluate tend not to attract much research. Therefore the development of new and different evaluation methodologies and metrics is a critical key to attacking future challenges. What follows is a discussion of four areas in information retrieval that hopefully will see more research in the future, given that there are better ways of evaluating them.

The first area is better understanding of how our retrieval processes/search engines actually work. The "Ideal" test collection proposed by Spärck Jones and van Rijsbergen [166] envisioned a series of different document types, different indexing schemes, etc., with the goal of seeing how the retrieval systems needed to be modified to deal with these differences. Although this collection was never built, current evidence from TREC and other evaluations tends to indicate that results from systems are fairly invariant to the domain and the type of textual documents (newspapers vs. medical documents vs. social science documents). The various indexing schemes (free vs. controlled terms vs. bibliographic information) have been mainly ignored in experimentation today. The requests now come in natural language from different users but have (unfortunately) never reached the 700 to 1000 they proposed. What might be useful to do today?

Experiments aimed at better understanding of our systems could be envisioned at two levels. The first level would be to investigate how the systems are working in the *simple* task of retrieval from newspapers, i.e., the TREC ad hoc task. Once this is better understood, a second level would look at how results would change as the systems move into truly diverse environments, such as different tasks, different types of documents (blogs vs. web vs. mixed media such as in the legal track in TREC), or different user populations.

For the first level we need better metrics and methodology for diagnostics. Robertson's analysis with GMAP [135] and his suggestion of using different metrics to tease apart system behavior is an example of working with metrics. The work with metrics likely needs to be tightly coupled with new diagnostic methodologies, however. A fruitful area for this investigation would be the causes of huge variations in system performance across topics, and there have been multiple papers trying to find ways of predicting performance on a given topic (for example [40, 62] and another lecture in this series "Estimating the Query Difficulty for Information Retrieval by David Carmel and Elad Yom-Tov").

Beyond predicting, however, a major source of improvement would be the ability to modify/tune the system on a per topic basis. This unfortunately turns out to be quite difficult; a 6-week workshop [83] in the summer of 2003 tried to determine the characteristics of topics (from the TREC ad hoc track) that would be helped by the use of relevance feedback, and could not reach any clear conclusions. The workshop employed eight of the best retrieval systems, along with their research teams, to do a series of controlled experiments with the TREC data *and* to do many hours of extensive failure analysis [31] using a specially-built tool. The results were often surprising and

revealed just how complex the current systems are, with considerable interaction between the various (often minor) components of the systems (see for example [45]).

Another way to tackle the same problem would be to analyze the topic sets looking for patterns of interaction between systems and topics. An attempt was made to do this with statistics [18], again mostly revealing how complex the interactions are. Guiver et. al [75] showed that small sets of topics were able to predict system effectiveness, with those sets composed of different topics depending on how many topics were used. This implies that there *may* be some way of generalizing topic characteristics and *one day* being able to test more specific components.

The second area where more research is needed is better integration of users into batch evaluations. Chapter 3 illustrated the logistical difficulties of doing user studies at a large enough scale to deal with both topic variation and user variation issues. Small controlled studies are hard to generalize, however the mixed results from applying successful batch methodologies into user applications shows the importance of better understanding of how users interact with the systems. The commercial search engines are able to track, and often generalize, these interactions, but may have different goals in terms of how to apply their user models.

One possible approach would be a type of user simulation within a batch environment. This could start with large scale observations of users doing a simple, well-defined task. The goal would be to identify some specific issues that could then be modeled in a batch environment. Results of that modeling would then be user-tested as proof of "improvements". For example the TREC-3 interactive track task was to find as many relevant documents as possible in up to 30 minutes, with 25 topics being searched (by multiple users). There were 9 systems that tackled this task, usually trying different interfaces to "help" the users with the task. Suppose that from this user study it could have been observed that specific types of document surrogates (or document space representations) were better than others. This could then be further developed within a batch environment, with results brought back for user testing. This example may be too simplistic (or naive), however the goal of integrating the users into the batch process is too important to just ignore. Moving forward will call for both new evaluation methodologies and metrics. There is some promising recent work on user simulation, with a workshop at SIGIR 2010 [14], and the new Sessions track at TREC. This needs to be pushed by both the user study community and the batch evaluation community.

Another approach would be to take more advantage of the facts that we already know based on previous user studies. An example of this would be to try different categories of relevance. Saracevic [154] listed many criteria such as novelty, validity, usability, fun, etc. that could be the basis of research (and evaluation). Some work has been done in novelty, and the criteria of validity has been somewhat tackled in the web world by the use of links. Research on removing spam is also relevant here, at least as a part of usability. A second source for relevance criteria is the Alonso and Mizzaro [11] study on one type of web searching. A major barrier to research in these different versions of relevance is the evaluation in a batch mode, that is, how do we effectively model these criteria in test collections.

A third area for future research is better integration of information retrieval technology into other human language technology areas, or for that matter, into other related areas. The text retrieval part of information access is only one part of the information access process; other technologies such as machine translation, summarization, and speech recognition are also needed. Whereas TREC and other evaluations have worked across areas, such as with cross-language information retrieval, or the various question-answering projects, this is usually done by connecting components. There have been many instances of good question-answering components dealing with poor retrieval components, or poor machine translation components feeding into information retrieval components, but little research on true interactions of these components.

The NTCIR evaluation program has made several efforts in this area. One was the investigation of the effects of machine translation accuracy on patent retrieval [72], and another has been a deliberate linking up of the question-answering results to the retrieval results in the more recent NTCIR meetings. In both these cases, and in other such research, the goal has been to measure the effects rather than trying to better integrate the components. But suppose that it was known that specific machine translations (words) were poor, would it be useful to integrate that knowledge into the retrieval process? Maybe not as information retrieval seems rather tolerant of "errors" in documents, but if that translation involved a key word in the query, is there a way to avoid poor results, such as by trying multiple translations. As another example, results from poor retrieval can guarantee poor question-answering results, but can retrieval systems be more tightly coupled into question-answering systems to allow a true interaction, such as getting more documents if it is obvious that the answer has not been found (several groups have tried this with success). Certainly part of the problem is that each of these technologies have very distinct communities, with different training, different evaluation criteria, and often different publishing venues. One of the goals of the various evaluations that have spanned communities (such as the TREC speech retrieval, the CLEF and NTCIR cross-language retrieval and the multiple QA evaluations) has been to bring these communities together, but this has generally not led to better integration.

The same types of "non-interaction" occurs with other related areas, such as database technology and image retrieval technology. Effective use of metadata and work with semi-structured information would likely be helped by database technology. The various image and video retrieval challenges employ both image and text and speech, and here there has been more integration, with text seen as only one component of the retrieval process. It is not clear if this has happened because of the characteristics of the communities involved or because the evaluations have encouraged specific component evaluation by requiring runs using a single component for comparison. Maybe this type of evaluation can be tried with more success in other cross-community evaluations or maybe evaluations modeled after Spärck Jones's cascading evaluations [168] will emphasize the effects of the various components and encourage tighter integration.

The final area for suggested research in evaluation addresses a much broader issue. The OPACs and "plain text" retrievals that started our field have exploded in many directions. A look at the vast array of tasks supported by today's commercial search engines, such as e-commerce and people/URL

location accessing, in addition to the standard "text retrieval", gives an idea of the breadth of challenges to retrieval (and evaluation). Add to this the multimedia aspect, such as working with maps, images, and video, and the social aspects such as Facebook, Twitter, and blogs and the prospect of evaluation becomes mind-boggling. And this is just the web; what about enterprise search, patent retrieval, medical data location, or legal e-discovery? How do we go about producing meaningful evaluations (or research) in this diverse environment?

One way is by targeting very specific tasks within a specific environment. The TREC tracks were formed to investigate specific issues in (plain text) retrieval, such as the robust track, but mostly to deal with these different environments, giving us the legal track, the genomics track, the enterprise track, etc. NTCIR, CLEF, INEX and FIRE have also taken this approach. This at least has allowed investigations into the issues in these different areas and has made some initial exploration of effective retrieval (and evaluation) strategies. Possibly this is all that can be done, and it is certainly useful as there are large potential audiences for this work.

But can we (or should we) try to learn more general principles about the retrieval process (including the user aspect of this process)? Are there commonalities across these tasks and applications that can be harnessed to improve retrieval in general or at least allow faster/better adaptation of systems to these diverse environments? A workshop (MINDS [38]) attempted to find some answers, and at a minimum laid out some of the challenges requiring research (and evaluation) that occur in most of these environments.

- heterogeneous data: spam, audio, video, slides, notes, images, metadata/structure
- heterogeneous context: diverse tasks such as finding, learning, monitoring, communicating, planning
- beyond the ranked list: information analysis and organization
- better understanding of what the user is actually doing

These are simply the challenges, not the answers, but it is critical that we as a community expand our research horizons, and necessarily our evaluation methodologies, to tackle today's world. This will not be easy; the SIGIR 2009 "Workshop on The Future of IR Evaluation" had good discussions but as many different opinions as to what could be done as there were workshop attendees. But limiting the scope of research because our evaluation methodology is too restricted is not an option.

Bibliography

[1] The Truth, the Whole Truth ... *American Documentation*, 6:56, 1955. Cited on page(s) 2

[2] Eytan Adar, Jaime Teevan, and Susan T. Dumais. Large Scale Analysis of Web Revisitation Patterns. *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206, 2008. DOI: 10.1145/1357054.1357241 Cited on page(s) 74

[3] J. Aitchison and C.W. Cleverdon. A Report on a Test of the Index of Metallurgical Literature of Western Reserve University. Aslib Cranfield Research Project, Cranfield, England, 1963. Cited on page(s) 4

[4] T.M. Aitchison, A.M. Hall, K.H. Lavelle, and J.M. Tracy. Comparative Evaluation of Index Languages. Institute of Electrical Engineers, London, England, 1970. Cited on page(s) 22

[5] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. The Good and the Bad System: Does the Test Collection Predict Users' Effectiveness. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–66, 2008. DOI: 10.1145/1390334.1390347 Cited on page(s) 67

[6] J. Allan, J. A. Aslam, V. Pavlu, E. Kanoulas, and B. Carterette. Million Query Track 2008 Overview. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008. Cited on page(s) 46

[7] J. Allan, B. Carterette, B. Dachev, J. A. Aslam, V. Pavlu, and E. Kanoulas. Million Query Track 2007 Overview. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, pages 85–104, 2007. Cited on page(s) 45

[8] James Allan. The HARD Track overview of the TREC 2004 High Accuracy Retrieval from Documents. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, pages 25–35, 2004. Cited on page(s) 65

[9] James Allan. The HARD Track overview of the TREC 2005 High Accuracy Retrieval from Documents. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, pages 51–68, 2005. Cited on page(s) 65

[10] Omar Alonso and Stefano Mizzaro. Can We Get Rid of TREC Assessors. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009. Cited on page(s) 81

[11] Omar Alonso and Stefano Mizzaro. Relevance Criteria for E-Commerce: A Crowdsourcing-based Experimental Analysis. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 760–761, 2009. DOI: 10.1145/1571941.1572115 Cited on page(s) 72, 81, 83

[12] Javed A. Aslam and Virgil Pavlu. A Practical Sampling Strategy for Efficient Retrieval Evaluation. Technical Report, College of Computer and Information Science, Northeastern University, 2007. Cited on page(s) 46

[13] Javed A. Aslam, Virgiliu Pavlu, and Emine Yilmaz. A Statistical Method for System Evaluation using Incomplete Judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 541–548, 2006. DOI: 10.1145/1148170.1148263 Cited on page(s) 46

[14] Leif Azzopardi, Kalervo Järvelin, Jaap Kamp, and Mark D. Smucker. Report on the SIGIR 2010 Workshop on the Simulation of Interaction. *SIGIR Forum*, 44(2):35, 2010. DOI: 10.1145/1924475.1924484 Cited on page(s) 83

[15] Peter Bailey, Nick Craswell, and David Hawking. Engineering a Multi-Purpose Test Collection for Web Retrieval Experiments. *Information Processing and Management*, 39(6):853–871, 2003. DOI: 10.1016/S0306-4573(02)00084-5 Cited on page(s) 38, 54

[16] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance Assessment: Are Judges Exchangeable and Does It Matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674, 2008. DOI: 10.1145/1390334.1390447 Cited on page(s) 39, 80

[17] K. Balog, I. Soboroff, P. Thomas, Peter Bailey, Nick Craswell, , and A.P. de Vries. Overview of TREC 2008 Enterprise Track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008. Cited on page(s) 39

[18] David Banks, Paul Over, and Nien-Fan Zhang. Blind Men and Elephants: Six Approaches to TREC Data. *Information Retrieval*, 1:7–34, 1999. DOI: 10.1023/A:1009984519381 Cited on page(s) 83

[19] F.H. Barker, D.C. Veal, and B.K. Gray. Retrieval Experiments based on Chemical Abstracts Condensates. Research report No. 2, UKCIS, Nottingham, England, 1974. Cited on page(s) 22

[20] J.R. Baron, D.D. Lewis, and D.W. Oard. TREC 2006 Legal Track Overview. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, pages 79–98, 2006. Cited on page(s) 40, 41

[21] Carol L. Barry and Linda Schamber. Users' Criteria for Relevance Evaluation: A Cross-Situational Comparison. *Information Processing and Management*, 34(2/3):219–236, 1998. DOI: 10.1016/S0306-4573(97)00078-2 Cited on page(s) 70, 71, 72

[22] Nicholas Belkin. Anomalous States of Knowledge as a Basis for Information Retrieval. *The Canadian Journal of Information Science*, 5:133–143, 1980. Cited on page(s) 58

[23] N.J. Belkin, R.N. Oddy, and H.M. Brooks. Ask for Information Retrieval: Part II. Results of a Design Study. *Journal of Documentation*, 38:145–164, 1982. DOI: 10.1108/eb026726 Cited on page(s) 58, 59

[24] Yaniv Bernstein and Justin Zobel. Redundant Documents and Search Effectiveness. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, pages 736–743, 2005. DOI: 10.1145/1099554.1099733 Cited on page(s) 53

[25] Suresh K. Bhavnani. Important Cognitive Components of Domain-Specific Search Knowledge. In *Proceedings of the Ninhth Text REtrieval Conference (TREC-9)*, pages 19–26, 2001. Cited on page(s) 69

[26] Abraham Bookstein. Relevance. *Journal of the American Society for Information Science*, pages 269–273, September 1979. Cited on page(s) 28

[27] Christine L. Borgman. Why are Online Catalogs Hard to Use? Lessons Learned from Information-Retrieval Studies. *Journal of the American Society for Information Science*, 37:387–400, 1986.
DOI: 10.1002/(SICI)1097-4571(198611)37:6%3C387::AID-ASI3%3E3.0.CO;2-8 Cited on page(s) 60

[28] Pia Borlund and Ian Ruthven. Evaluation of Interactive Information Retrieval Systems. *Information Processing and Management*, 44:1–142, 2008. DOI: 10.1016/j.ipm.2007.03.006 Cited on page(s) 57

[29] M. Braschler, P. Schäuble, and C. Peters. Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 25–34, 2000. Cited on page(s) 37

[30] Martin Braschler. Clef 2002 – Overview of Results. In *Evaluation of Cross-Language Information Systems, the Third Workshop of the Cross-Language Forum*, pages 9–27. Springer LNCS 2785, 2003. DOI: 10.1007/978-3-540-45237-9_2 Cited on page(s) 49

[31] Chris Buckley. Why Current IR Engines Fail. *Information Retrieval*, 12(6):652–665, 2009. DOI: 10.1007/s10791-009-9103-2 Cited on page(s) 82

[32] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen M. Voorhees. Bias and the Limits of Pooling. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 619–620, 2006. DOI: 10.1145/1148170.1148284 Cited on page(s) 45

[33] Chris Buckley and Ellen Voorhees. Retrieval System Evaluation. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3. The MIT Press, 2005. Cited on page(s) 34

[34] Chris Buckley and Ellen M. Voorhees. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2000. DOI: 10.1145/345508.345543 Cited on page(s) 32

[35] Chris Buckley and Ellen M. Voorhees. Retrieval Evaluation with Incomplete Information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, 2004. DOI: 10.1145/1008992.1009000 Cited on page(s) 45

[36] Robert Burgin. Variations in Relevance Judgments and the Evaluation of Retrieval Performance. *Information Processing and Management*, 28(5):619–627, 1992. DOI: 10.1016/0306-4573(92)90031-T Cited on page(s) 28

[37] S. Buttcher, C.L.A. Clarke, and I. Soboroff. The TREC 2006 Terabyte Track. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, pages 128–141, 2006. Cited on page(s) 45

[38] Jamie Callan, James Allan, Charles L. A. Clarke, Susan Dumais, David A. Evans, Mark Sanderson, and ChengXiang Zhai. Meeting of the MINDS: An Information Retrieval Research Agenda. *SIGIR Forum*, 41(2):25–34, 2007. DOI: 10.1145/1328964.1328967 Cited on page(s) 85

[39] B. Capps and M. Yin. The Effectiveness of Feedback Strategies on Collections of Differing Generality. In *Scientific Report ISR-18 to NSF*, chapter IX. Cornell University, Ithaca, N.Y, 1970. Cited on page(s) 15

[40] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. What Makes a Query Difficult? In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 390–397, 2006. DOI: 10.1145/1148170.1148238 Cited on page(s) 82

[41] Ben Carterette, James Allan, and Ramesh K. Sitaraman. Minimal Test Collections for Retrieval Evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268–275, 2006. DOI: 10.1145/1148170.1148219 Cited on page(s) 46

[42] Ben Carterette, Evangelos Kanoulas, Virgiliu Pavlu, and Hui Fang. Reusable Test Collections through Experimental Design. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 547–554, 2010. DOI: 10.1145/1835449.1835541 Cited on page(s) 80

[43] Ben Carterette, Virgiliu Pavlu, Evangelos Kanoulas, Javed A. Aslam, and James Allan. Evaluation over Thousands of Queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 651–658, 2008. DOI: 10.1145/1390334.1390445 Cited on page(s) 80

[44] Ben Carterette and Ian Soboroff. The Effect of Assessor Error on IR System Evaluation. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–546, 2010. DOI: 10.1145/1835449.1835540 Cited on page(s) 80

[45] Charles L. A. Clarke, Gordon V. Cormack, Thomas R. Lynam, Chris Buckley, and Donna Harman. Swapping Documents and Terms. *Information Retrieval*, 12(6):680–694, 2009. DOI: 10.1007/s10791-009-9105-0 Cited on page(s) 83

[46] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–666, 2008. DOI: 10.1145/1390334.1390446 Cited on page(s) 51, 79, 81

[47] C.L.A. Clarke, F. Scholer, and I. Soboroff. The TREC 2005 Terabyte Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, pages 109–119, 2005. Cited on page(s) 45

[48] C.W. Cleverdon. Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems. Aslib Cranfield Research Project, Cranfield, England, 1960. Cited on page(s) 2

[49] C.W. Cleverdon. Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Aslib Cranfield Research Project, Cranfield, England, 1962. Cited on page(s) 2

[50] C.W. Cleverdon. The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages. Cranfield Library Report No. 3, Cranfield, England, 1970. Cited on page(s) 33

[51] C.W. Cleverdon and E.M. Keen. Factors Determining the Performance of Indexing Systems, Vol. 2: Test Results. Aslib Cranfield Research Project, Cranfield, England, 1966. Cited on page(s) 3, 8

[52] C.W. Cleverdon, J. Mills, and E.M. Keen. Factors Determining the Performance of Indexing Systems, Vol. 1: Design. Aslib Cranfield Research Project, Cranfield, England, 1966. Cited on page(s) 3, 7

[53] Cyril Cleverdon. The Signficance of the Cranfield Tests on Index Languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1991. DOI: 10.1145/122860.122861 Cited on page(s) 2

[54] Paul Clough, Mark Sanderson, Murad Adouammoh, Sergio Navarro, and Monica Paramita. Multiple Approaches to Analysing Query Diversity. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 734–735, 2009. DOI: 10.1145/1571941.1572102 Cited on page(s) 74

[55] W.S. Cooper. Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems. *American Documentation*, pages 30–41, January 1968. DOI: 10.1002/asi.5090190108 Cited on page(s) 25, 36

[56] W.S. Cooper. A Definition of Relevance for Information Retrieval. *Information Storage and Retrieval*, 7:19–37, 1971. DOI: 10.1016/0020-0271(71)90024-6 Cited on page(s) 28

[57] Gordon V. Cormack, Charles L. A. Clarke, Christopher R. Palmer, and Samuel S. L. To. Passage-based Refinement(MultiText Experiments for TREC-6). In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 303–320, 1998. Cited on page(s) 34

[58] G.V. Cormack. TREC 2007 Spam Track Overview. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, pages 123–131, 2007. Cited on page(s) 43, 44

[59] Nick Craswell, A.P. de Vries, and Ian Soboroff. Overview of TREC 2005 Enterprise Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, pages 17–24, 2005. Cited on page(s) 39

[60] Nick Craswell and David Hawking. Overview of TREC 2002 Web Track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, 2002. Cited on page(s) 38

[61] Nick Craswell and David Hawking. Overview of TREC 2004 Web Track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, pages 89–97, 2004. Cited on page(s) 39

[62] S. Cronen-Townsend, Y. Zhou, and W.B. Croft. Predicting Query Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2002. DOI: 10.1145/564376.564429 Cited on page(s) 82

[63] H.T. Dang, D.Kelly, and J.Lin. Overview of the TREC 2007 Question Answering Track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, pages 105–122, 2007. Cited on page(s) 43, 65

[64] Tamas E. Doszkocs. From Research to Application: the CITE Natural Language System. In *Research and Development in Information Retrieval*, pages 251–262, 1982. DOI: 10.1007/BFb0036350 Cited on page(s) 59

[65] Tamas E. Doszkocs and Barbara A. Rapp. Searching MEDLINE in English: a Prototype User Interface with Natural Language Query, Ranked Output, and Relevance Feedback. In *Proceedings of the 42nd Annual Meeting of the American Society for Information Science*, pages 131–139, 1979. Cited on page(s) 59

[66] Susan Dumais, Krisha Bharat, Thorsten Joachims, and Andreas Weigend. SIGIR 2003 Workshop Report: Implicit Measures of User Interests and Preferences. *SIGIR Forum*, pages 50–54, 2003. DOI: 10.1145/959258.959266 Cited on page(s) 73

[67] Susan T. Dumais and Nicholas J. Belkin. The TREC Interactive Tracks: Putting the User into Search. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 6. The MIT Press, 2005. Cited on page(s) 63, 64

[68] Brynn M. Evans, Sanjay Kairam, and Peter Pirolli. Do Your Friends Make You Smarter? an Analysis of Social Strategies in Online Information Seeking. *Information Processing and Management*, 46(6):679–692, 2010. DOI: 10.1016/j.ipm.2009.12.001 Cited on page(s) 70

[69] E. Fox. Characteristics of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts. Technical Report TR 83-561, Cornell University: Computing Science Department, 1983. Cited on page(s) 19

[70] Steve Fox, Kuldeep Karnawat, Mark Myland, Susan Dumais, and Thomas White. Evaluating Implicit Measures to Improve Web Search. *ACM Transactions on Information Systems*, 23(2):147–168, 2005. DOI: 10.1145/1059981.1059982 Cited on page(s) 73

[71] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Introduction to the special issue on patent processing. *Information Processing and Management*, 43:1149–1153, September 2007. DOI: 10.1016/j.ipm.2006.11.004 Cited on page(s) 48, 54

[72] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Evaluating Effects of Machine Translation Accuracy on Cross-Lingual Patent Retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 674–675, 2009. DOI: 10.1145/1571941.1572072 Cited on page(s) 84

[73] J.S. Garofolo, C.G.P Auzanne, and E.M. Voorhees. The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 107–130, 2000. Cited on page(s) 36

[74] H. Gilbert and K. Sparck Jones. Statistical Bases of Relevance Assessment for the "Ideal" Information Retrieval Test Collection. British Library Research and Development Report 5481, Computer Laboratory, University of Cambridge, 1979. Cited on page(s) 25

[75] John Guiver, Stefano Mizzaro, and Stephen Robertson. A Few Good Topics: Experiments in Topic Set Reduction for Retrieval Evaluation. *ACM Transactions of Information Systems*, 27(4), 2009. DOI: 10.1145/1629096.1629099 Cited on page(s) 83

[76] D. Gull. Seven Years of Work on the Organisation of Materials in a Special Library. *American Documentation*, 7:320–329, 1956. DOI: 10.1002/asi.5090070408 Cited on page(s) 3

[77] Micheline Hancock-Beaulieu and Stephen Walker. An Evaluation of Automatic Query Expansion in an Online Library Catalogue. *Journal of Documentation*, 48(4):406–421, 1992. DOI: 10.1108/eb026906 Cited on page(s) 61

[78] Donna Harman. Overview of the Second Text REtrieval Conference (TREC-2). In *Proceedings of the Second Text REtrieval Conference (TREC-2)*, pages 1–20, 1994. DOI: 10.1016/0306-4573(94)00047-7 Cited on page(s) 32, 33

[79] Donna Harman. Overview of the Third Text REtrieval Conference (TREC-3). In *Overview of the Third Text REtrieval Conference (TREC-3) [Proceedings of TREC-3.]*, pages 1–20, 1995. Cited on page(s) 32

[80] Donna Harman. Overview of the Fourth Text REtrieval Conference (TREC-4). In *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1–23, 1996. Cited on page(s) 33

[81] Donna Harman. Overview of the TREC 2002 Novelty Track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, pages 46–56, 2002. Cited on page(s) 53

[82] Donna Harman and Chris Buckley. Overview of the Reliable Information Access Workshop. *Information Retrieval*, 12(6):615–641, 2009. DOI: 10.1007/s10791-009-9101-4 Cited on page(s) 32

[83] Donna Harman and Chris Buckley. Overview of the Reliable Information Access Workshop. *Information Retrieval*, 12(6):615–641, 2009. DOI: 10.1007/s10791-009-9101-4 Cited on page(s) 82

[84] S.P Harter. The Cranfield II Relevance Assessments: a Critical Evaluation. *Library Quarterly*, 41:229–243, 1971. DOI: 10.1086/619960 Cited on page(s) 8

[85] Stephen P. Harter. Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, 1996. DOI: 10.1002/(SICI)1097-4571(199601)47:1%3C37::AID-ASI4%3E3.3.CO;2-I Cited on page(s) 28

[86] David Hawking and Nick Craswell. Overview of TREC-7 Very Large Corpus Track. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 91–106, 1999. Cited on page(s) 38

[87] David Hawking and Nick Craswell. The Very Large Collection and Web Track. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. The MIT Press, 2005. Cited on page(s) 38

[88] David Hawking and Paul Thistlewaite. Overview of TREC-6 Very Large Corpus Track. In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 93–107, 1998. Cited on page(s) 38

[89] David Hawking, Ellen Voorhees, and Nick Craswell. Overview of TREC-8 Web Track. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 131–151, 2000. Cited on page(s) 38, 54

[90] Ben He, Craig Macdonald, and Iadh Ounis. Retrieval Sensitivity under Training using Different Measures. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 67–74, 2008. DOI: 10.1145/1390334.1390348 Cited on page(s) 80

[91] B. Hedin, S. Tomlinson, J.R. Baron, and D.W. Oard. Overview of the TREC 2009 Legal Track. In *Proceedings of the Einhteenth Text REtrieval Conference (TREC 2009)*, 2009. Cited on page(s) 41, 66

[92] W. Hersh, R.T. Bhupatiraju, L. Ross, A.M. Cohen, D.F. Kraemer, , P. Johnson, and M. Hearst. Overview of TREC 2005 Genomics Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, pages 25–50, 2005. Cited on page(s) 40

[93] W. Hersh, A. Cohen, J. Yang, R.T. Bhupatiraju, P. Roberts, and M. Hearst. Overview of TREC 2004 Genomics Track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, pages 13–24, 2004. Cited on page(s) 40

[94] William Hersh. Interactivity at the Text Retrieval Conference (TREC). *Information Processing and Management*, 37(3):365–541, 2001. DOI: 10.1016/S0306-4573(00)00052-2 Cited on page(s) 63

[95] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kraemer, Lynetta Sacherek, and Daniel Olsen. Do Batch and User Evaluations Give the Same Results. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 17–24, 2000. DOI: 10.1145/345508.345539 Cited on page(s) 66

[96] William Hersh and Ellen M. Voorhees. TREC Genomics Special Issue Overview. *Information Retrieval*, 12:1–15, 2009. DOI: 10.1007/s10791-008-9076-6 Cited on page(s) 40

[97] David A. Hull. Stemming Algorithms: A Case Study for Detailed Evaluation. *Journal of the American Society for Information Science*, 47(1):70–84, 1996. DOI: 10.1002/(SICI)1097-4571(199601)47:1%3C70::AID-ASI7%3E3.3.CO;2-Q_ Cited on page(s) 79

[98] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, Dordrecht, The Netherlands, 2005. Cited on page(s) 57

[99] Kalervo Järvelin and Jaana Kekäläinen. Ir Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, 2000. DOI: 10.1145/345508.345545 Cited on page(s) 51, 53

[100] Kalervo Järvelin and Jaana Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October 2002. DOI: 10.1145/582415.582418 Cited on page(s) 51, 81

[101] Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 4–15, Berlin, Heidelberg, 2008. Springer-Verlag. DOI: 10.1007/978-3-540-78646-7_4 Cited on page(s) 51

[102] Joseph John Rocchio Jr. Document Retrieval Systems – Optimization and Evaluation. Scientific Report ISR-10 to NSF, Cambridge, Massachusetts, 1966. Cited on page(s) 13

[103] R.V. Katter. The Influence of Scale on Relevance Judgments. *Information Storage and Retrieval*, 4:1–11, 1968. DOI: 10.1016/0020-0271(68)90002-8 Cited on page(s) 28

[104] Gabriella Kazai, Natasa Milic-Frayling, and Jamie Costello. Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 452–459, 2009. DOI: 10.1145/1571941.1572019 Cited on page(s) 81

[105] E. Michael Keen. The Aberystwyth Index Languages Test. *Journal of Documentation*, 29(41):1–35, 1973. DOI: 10.1108/eb026547 Cited on page(s) 58

[106] E.M. Keen. Document Length. In *Scientific Report ISR-13 to NSF*, chapter V. Cornell University, Ithaca, N.Y, 1967. Cited on page(s) 16

[107] E.M. Keen. Evaluation Parameters. In *Scientific Report ISR-13 to NSF*, chapter II. Cornell University, Ithaca, N.Y, 1967. Cited on page(s) 13

[108] E.M. Keen. Test Environment. In *Scientific Report ISR-13 to NSF*, chapter I. Cornell University, Ithaca, N.Y, 1967. Cited on page(s) 16

[109] E.M. Keen. On the Performance of Nine Printed Index Entry Types. British Library Report 5475, Aberystwyth, Wales, 1972. Cited on page(s) 58

[110] E.M. Keen and J.A. Digger. Report of an Information Science Index Languages Test. British Library Report 5120, Aberystwyth, Wales, 1972. Cited on page(s) 22

[111] Diane Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval*, 3:1–224, 2010. DOI: 10.1561/1500000012 Cited on page(s) 57, 58

[112] Diane Kelly, Xin Fu, and Chirag Shah. Effects of Postition and Number of Relevant Documents Retrieved on Users' Evaluations of System Performance. *ACM Transactions of Information Systems*, 28:1–29, 2010. DOI: 10.1145/1740592.1740597 Cited on page(s) 68

[113] Eric Lagergren and Paul Over. Comparing Interactive Information Retrieval Systems Across Sites: The TREC-6 Interactive Track Matrix Experiment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 164–172, 1998. DOI: 10.1145/290941.290986 Cited on page(s) 63

[114] Mounia Lalmas and Anastasios Tombros. Evaluating XML Retrieval Effectiveness at INEX. *SIGIR Forum*, 41(1):40–57, 2007. DOI: 10.1145/1273221.1273225 Cited on page(s) 50

[115] F.W. Lancaster. Evaluation of the MEDLARS Demand Search Service. National Library of Medicine, Washington, D.C., 1968. Cited on page(s) 9, 10, 77

[116] M. Lesk, D. Harman, E. Fox, and C. Buckley. The SMART Lab Report. *SIGIR Forum*, pages 2–22, 1997. DOI: 10.1145/263868.263870 Cited on page(s) 11, 16, 18, 19, 20

[117] M.E. Lesk and G. Salton. Relevance Assessments and Retrieval System Evaluation. In *Scientific Report ISR-14 to NSF*, chapter III. Cornell University, Ithaca, N.Y, 1968. DOI: 10.1016/0020-0271(68)90029-6 Cited on page(s) 16, 18, 33

[118] David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. Training Algorithms for Linear Text Classifiers. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–306, 1996. DOI: 10.1145/243199.243277 Cited on page(s) 44

[119] H.P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal*, October:309–317, 1957. DOI: 10.1147/rd.14.0309 Cited on page(s) 11

[120] C. Macdonald, I. Ounis, and I. Soboroff. Overview of TREC 2009 Blog Track. In *Proceedings of the Einhteenth Text REtrieval Conference (TREC 2009)*, 2009. Cited on page(s) 42, 54

[121] R. Merchant, editor. *The Proceedings of the TIPSTER Text Program—Phase I*, 1994. Morgan Kaufmann Publishing Co. San Mateo, California. Cited on page(s) 27

[122] Donald Metzler, Jasmine Novak, Hang Cui, and Srihari Reddy. Building Enriched Document Representations using Aggregated Anchor Text. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 219–226, 2009. DOI: 10.1145/1571941.1571981 Cited on page(s) 55

[123] Nathalie Mitev, Gillian Venner, and Stephen Walker. Designing an Online Public Acess Catalogue: Okapi, a catalogue on a local area network. Library and Information Research Report 39, London: British Library, 1985. Cited on page(s) 60

[124] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. A Comparison of Information Seeking Using Search Engines and Social Networks. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 291–294, 2010. Cited on page(s) 70

[125] E.M. Needham and K. Sparck Jones. KEYWORDS AND CLUMPS: Recent work on information retrieval at the Cambridge Language Research Unit. *Journal of Documentation*, 20(1):5–15, 1964. DOI: 10.1108/eb026337 Cited on page(s) 22

[126] I. Ounis, C. Macdonald, M. de Rijk, G. Mishne, and I. Soboroff. Overview of TREC 2006 Blog Track. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, pages 17–31, 2006. Cited on page(s) 41

[127] I. Ounis, C. Macdonald, and I. Soboroff. Overview of TREC 2008 Blog Track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008. Cited on page(s) 41

[128] J.W. Perry. Operational Criteria for Designing Information Retrieval Systems. *American Documentation*, 6:93–101, 1955. DOI: 10.1002/asi.5090060209 Cited on page(s) 7

[129] Filip Radlinski and Nick Craswell. Comparing the Sensitivity of Information Retrieval Metrics. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674, 2010. DOI: 10.1145/1835449.1835560 Cited on page(s) 81

[130] A.M. Rees. Evaluation of Information Systems and Services. In *Annual Review of Information Science and Technology*, chapter 3. Interscience, 1967. Cited on page(s) 8

[131] P.M. Roberts, A.M. Cohen, and W.R. Hersh. Tasks, Topics and Relevance Judging for the TREC Genomics Track: Five Years of Experience Evaluating Biomedical Text Information Retrieval Systems. *Information Retrieval*, 12:81–97, 2009. DOI: 10.1007/s10791-008-9072-x Cited on page(s) 40

[132] S. E. Robertson and M. M. Hancock-Beaulieu. On the Evaluation of IR Systems. *Information Processing and Management*, 28(4):457–466, 1992. DOI: 10.1016/0306-4573(92)90004-J Cited on page(s) 60

[133] S.E. Robertson. The Parametric Description of Retrieval Tests. *Journal of Documentation*, 25:1–27, 1969. DOI: 10.1108/eb026466 Cited on page(s) 26

[134] S.E. Robertson. On the history of evaluation in IR. *Journal of Information Science*, 34:439–456, 2008. DOI: 10.1177/0165551507086989 Cited on page(s) 1

[135] Stephen Robertson. On GMAP: and Other Tansformations. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, pages 78–83, 2006. DOI: 10.1145/1183614.1183630 Cited on page(s) 79, 82

[136] Stephen Robertson and Jamie Callan. Routing and Filtering. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 5. The MIT Press, 2005. Cited on page(s) 44, 45

[137] G. Salton. Automatic Processing of Foreign Language Documents. In *Scientific Report ISR-16 to NSF*, chapter IV. Cornell University, Ithaca, N.Y, 1967. DOI: 10.3115/990403.990407 Cited on page(s) 18

[138] G. Salton. The "Generality" Effect and the Retrieval Evaluation for Large Collections. In *Scientific Report ISR-18 to NSF*, chapter II. Cornell University, Ithaca, N.Y, 1970. Cited on page(s) 15

[139] G. Salton. The "Generality" Effect and the Retrieval Evaluation for Large Collections. *Journal of the American Society for Information Science*, pages 11–22, January-February 1972. DOI: 10.1002/asi.4630230105 Cited on page(s) 15

[140] G. Salton. A New Comparison Between Conventional Indexing (Medlars) and Automatic Text Processing (SMART). In *Scientific Report ISR-21 to NSF*, chapter I. Cornell University, Ithaca, N.Y, 1972. Cited on page(s) 18

[141] G. Salton. A New Comparison Between Conventional Indexing (Medlars) and Automatic Text Processing (SMART). *Journal of the American Society for Information Science*, pages 75–84, March-April 1972. DOI: 10.1002/asi.4630230202 Cited on page(s) 18

[142] G. Salton and M.E. Lesk. Information Analysis and Dictionary Construction. In *Scientific Report ISR-11 to NSF*, chapter IV. Cornell University, Ithaca, N.Y, 1966. Cited on page(s) 16

[143] G. Salton and D.K. Williamson. A Comparison Between Manual and Automatic Indexing Methods. In *Scientific Report ISR-14 to NSF*, chapter VI. Cornell University, Ithaca, N.Y, 1968. DOI: 10.1002/asi.4630200109 Cited on page(s) 18

[144] G. Salton and C.S. Yang. On the Specification of Term Values in Automatic Indexing. *Journal of Documentation*, 29(4):351–372, 1973. DOI: 10.1108/eb026562 Cited on page(s) 19

[145] Gerard Salton. The Evaluation of Automatic Retrieval Procedures– Selected Test Results Using the SMART System. *American Documentation*, 16(3):209–222, 1965. DOI: 10.1002/asi.5090160308 Cited on page(s) 16

[146] Gerard Salton, editor. *The SMART Retrieval System*. Prentice-Hall, Englewood Cliffs, New Jersey, 1971. Cited on page(s) 11, 13, 15, 16, 18, 33

[147] Mark Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*, 4:247–375, 2010. DOI: 10.1561/1500000009 Cited on page(s) 1, 7, 26, 32, 34, 51, 54, 55

[148] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Vanja Josifovski. Do User Preferences and Evaluation Measures Line Up. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–562, 2010. DOI: 10.1145/1835449.1835542 Cited on page(s) 67, 81

[149] Mark Sanderson and Ian Soboroff. Problems with Kendall's tau. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 839–840, 2007. DOI: 10.1145/1277741.1277935 Cited on page(s) 80

[150] Mark Sanderson and Justin Zobel. Information Retrieval System Evaluation: Effort, Sensitivity and Reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169, 2005. DOI: 10.1145/1076034.1076064 Cited on page(s) 79

[151] Tefko Saracevic. Linking Research and Teaching. *American Documentation*, pages 398–403, October 1968. DOI: 10.1002/asi.5090190407 Cited on page(s) 20

[152] Tefko Saracevic. Selected Results from an Inquiry into Testing of Information Retrieval Systems. *Journal of the American Society for Information Science*, pages 126–139, March-April 1971. DOI: 10.1002/asi.4630220212 Cited on page(s) 20, 21

[153] Tefko Saracevic. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance. *Journal of the American Society for Information Science*, 58(13):1915–1933, 2007. DOI: 10.1002/asi.20682 Cited on page(s) 62

[154] Tefko Saracevic. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance. *Journal of the American Society for Information Science*, 58(13):2126–2144, 2007. DOI: 10.1002/asi.20681 Cited on page(s) 62, 83

[155] Jacques Savoy. Stemming Strategies for European Languages. In *Proceedings of the 10th International Conference on Innovative Internet Community Services(IICS)*, pages 545–557, 2010. Cited on page(s) 49

[156] P. Sheridan, J.P. Ballerini, and P. Schäuble. Building a Large Multilingual Test Collection from Comparable News Documents. In *Cross-language Information Retrieval*, pages 137–180. Kluwer Academic Publishers, 1998. Cited on page(s) 37

[157] Elliot.R. Siegel, Karen Kameen, Sally.K. Sinn, and Frieda O. Weise. Research Strategy and Methods used to Conduct a Comparative Evaluation of Two Prototype Online Catalog Systems. In *Proceedings of the National Online Meeting*, pages 503–511, 1984. Cited on page(s) 59

[158] Catherine L. Smith and Paul B. Kantor. User Adaptation: Good Results from Poor Systems. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 147–154, 2008. DOI: 10.1145/1390334.1390362 Cited on page(s) 68

[159] Mark D. Smucker, James Allan, and Ben Carterette. Agreement Among Statistical Significance Tests for Information Retrieval Evaluation at Varying Sample Sizes. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 630–631, 2009. DOI: 10.1145/1571941.1572050 Cited on page(s) 79, 80

[160] Cees Snoek and Marcel Worring. Multimodal Video Indexing: A Review of the State-of-the-art. *Multimedia Tools Appl.*, 25(1):5–35, 2005. DOI: 10.1023/B:MTAP.0000046380.27575.a5 Cited on page(s) 36

[161] Ian Soboroff and Stephen Robertson. Building a Filtering Test Collection for TREC 2002. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 243–250, 2003. DOI: 10.1145/860435.860481 Cited on page(s) 44

[162] E. Sormunen. Liberal Relevance Criteria of TREC—Counting on Negligible Documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–330, 2002. DOI: 10.1145/564376.564433 Cited on page(s) 53

[163] K. Sparck Jones. Collection Properties Influencing Automatic Term Classifications Performance. *Information Storage and Retrieval*, 9:499–513, 1973. DOI: 10.1016/0020-0271(73)90036-3 Cited on page(s) 22

[164] K. Sparck Jones and R.G. Bates. Report on a Design Study for the "Ideal" Information Retrieval Test Collection. British Library Research and Development Report 5488, Computer Laboratory, University of Cambridge, 1977. Cited on page(s) 23, 30

[165] K. Sparck Jones and D.M. Jackson. The Use of Automatically-Obtained Keyword Classifications for Information Retrieval. *Information Storage and Retrieval*, 5:175–201, 1970. DOI: 10.1016/0020-0271(70)90046-X Cited on page(s) 22

[166] K. Sparck Jones and C. van Rijsbergen. Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975. Cited on page(s) 22, 82

[167] Karen Sparck Jones, editor. *Information Retrieval Experiment*. Butterworths, 1981. Cited on page(s) 1, 25, 26, 77

[168] Karen Sparck Jones. Towards Better NLP System Evaluation. In *Proceedings of the workshop on Human Language Technology*, pages 102–107, 1994. DOI: 10.3115/1075812.1075833 Cited on page(s) 84

[169] Louise T. Su. Evaluation Measures for Interactive Information Retrieval. *Information Processing and Management*, 28(4):503–516, 1992. DOI: 10.1016/0306-4573(92)90007-M Cited on page(s) 69

[170] Russell C. Swan and James Allan. Aspect Windows, 3-D Visualizations, and Indirect Comparisons of Information Retrieval Systems. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 173–181, 1998. DOI: 10.1145/290941.290987 Cited on page(s) 64

[171] D.R. Swanson. Some Unexplained Aspects of the Cranfield Tests of Indexing Language Performance. *Library Quarterly*, 41:223–228, 1971. DOI: 10.1086/619959 Cited on page(s) 8

[172] Jean Tague-Sutcliffe. The Pragmatics of Information Retrieval Experimentation, Revisited. *Information Processing and Management*, 28:467–490, 1992. DOI: 10.1016/0306-4573(92)90005-K Cited on page(s) 77

[173] Jean Tague-Sutcliffe and James Blustein. A Statistical Analysis of the TREC-3 Data. In *Overview of the Third Text REtrieval Conference (TREC-3) [Proceedings of TREC-3.]*, pages 385–398, 1995. Cited on page(s) 32

[174] Robert S. Taylor. Question-Negotiation and Information Seeking in Libraries. *College and Research Libraries*, 28:178–194, 1968. Cited on page(s) 58

[175] Jaime Teevan, Susan T. Dumais, and Dan Liebling. To Personalize or not to Personalize: Modeling Queries with Variation in User Intent. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 163–170, 2008. DOI: 10.1145/1390334.1390364 Cited on page(s) 56, 73, 78

[176] John E. Tolle. Monitoring and Evaluation of Information Systems Via Transaction Log Analysis. In *Research and Development in Information Retrieval*, pages 247–258, 1984. Cited on page(s) 59

[177] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. How Users Assess Web Pages for Information Seeking. *Journal of the American Society for Information Science*, 56(4):327–344, 2005. DOI: 10.1002/asi.20106 Cited on page(s) 71, 72

[178] Andrew Turpin and William Hersh. Why Batch and User Evaluations do not Give the Same Results. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 225–231, 2001. DOI: 10.1145/383952.383992 Cited on page(s) 66, 67, 79

[179] Andrew Turpin, Falk Scholer, Kalvero Järvelin, Mingfang Wu, and J. Shane Culpepper. Including Summaries in System Evaluation. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 508–515, 2009. DOI: 10.1145/1571941.1572029 Cited on page(s) 81

[180] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, 1975. Cited on page(s) 26

[181] P.K.T. Vaswani and J.B. Cameron. The National Physical Laboratory Experiments in Statistical Word Associations and their Use in Document Indexing and Retrieval. Publication 42, Division of Computer Science, National Physical Laboratory, Teddington, 1970. Cited on page(s) 22

[182] Ellen Voorhees. Question Answering in TREC. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 10. The MIT Press, 2005. DOI: 10.1145/502585.502679 Cited on page(s) 43

[183] Ellen Voorhees and John S. Garofolo. Retrieving Noisy Text. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 8. The MIT Press, 2005. Cited on page(s) 36, 51

[184] Ellen Voorhees and Donna Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005. Cited on page(s) 27

[185] Ellen M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing and Management*, 36(5):697–716, 2000. DOI: 10.1016/S0306-4573(00)00010-8 Cited on page(s) 33

[186] Ellen M. Voorhees. Overview of the TREC 2002 Question Answering Track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, pages 57–68, 2002. Cited on page(s) 42, 43

[187] Ellen M. Voorhees. Overview of TREC 2004 Robust Track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, pages 70–79, 2004. Cited on page(s) 51

[188] Ellen M. Voorhees. Topic Set Size Redux. In *SIGIR09*, pages 806–807, 2009. DOI: 10.1145/1571941.1572138 Cited on page(s) 54, 80

[189] Ellen M. Voorhees and Chris Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, 2002. DOI: 10.1145/564376.564432 Cited on page(s) 32, 54

[190] Ellen M. Voorhees and H.T. Dang. Overview of the TREC 2005 Question Answering Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*, pages 69–80, 2005. Cited on page(s) 43

[191] Ellen M. Voorhees and Donna Harman. Overview of the Fifth Text REtrieval Conference (TREC-5). In *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 1–28, 1997. Cited on page(s) 32, 36

[192] Ellen M. Voorhees and Dawn T. Tice. Building a Question Answering Test Collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207, 2000. DOI: 10.1145/345508.345577 Cited on page(s) 42, 43

[193] J.C. Wade and J. Allan. Passage Retrieval and Evaluation. Technical Report IR-396, CIIR, Department of Computer Science, University of Massachusetts, Amherst, 2005. Cited on page(s) 65

[194] Stephen Walker and Rachel de Vere. Improving Subject Retrieval in Online Catalogues: 2. relevance feedback and query expansion. British Library Research Paper 72, London: British Library, 1989. Cited on page(s) 61

[195] Stephen Walker and Micheline Hancock-Beaulieu. Okapi at City, an evaluation facility for interactive IR. British Library Research Report 6056, London: British Library, 1991. Cited on page(s) 61

[196] Stephen Walker and Richard M. Jones. Improving Subject Retrieval in Online Catalogues: 1 . stemming, automatic spelling correction and cross-reference tables. British Library Research Paper 24, London: British Library, 1987. Cited on page(s) 60

[197] Ryen W.. White, Susan T. Dumais, and Jaime Teevan. Characterizing the Influence of Domain Expertise on Web Search Behavior. *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*, 2009. DOI: 10.1145/1498759.1498819 Cited on page(s) 74, 77

[198] Ryen W. White and Dan Morris. Investigating the Querying and Browsing Behaviour of Advanced Search Engine Users. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 255–262, 2007. DOI: 10.1145/1277741.1277787 Cited on page(s) 55

[199] Ross Wilkinson. Effective Retrieval of Structured Documents. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 311–317, 1994. Cited on page(s) 53

[200] D. Williamson, R. Williamson, and M.E. Lesk. The Cornell Implementation of the SMART System. In *Scientific Report ISR-16 to NSF*, chapter I. Cornell University, Ithaca, N.Y, 1967. Cited on page(s) 15

[201] Christa Womser-Hacker. Multilingual Topic Generation within the CLEF 2001 Experiments. In *Evaluation of Cross-Language Information Systems, the Second Workshop of the Cross-Language Forum*, pages 389–393. Springer LNCS 2406, 2001. DOI: 10.1007/3-540-45691-0_36 Cited on page(s) 49, 54, 78

[202] Emine Yilmaz and Javed A. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, pages 102–111, 2006. DOI: 10.1145/1183614.1183633 Cited on page(s) 45

[203] Y. Zhang, J. Callan, and T. Minka. Novelty and Redundancy Detection in Adaptive Filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–88, 2002. DOI: 10.1145/564376.564393 Cited on page(s) 53

[204] Justin Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, 1998. DOI: 10.1145/290941.291014 Cited on page(s) 32

Author's Biography

DONNA HARMAN

Donna Harman graduated from Cornell University as an Electrical Engineer, and started her career working with Professor Gerard Salton in the design and building of several test collections, including the first MEDLARS one. Later work was concerned with searching large volumes of data on relatively small computers, starting with building the IRX system at the National Library of Medicine in 1987, and then the Citator/PRISE system at the National Institute of Standards and Technology (NIST) in 1988. In 1990 she was asked by DARPA to put together a realistic test collection on the order of 2 gigabytes of text, and this test collection was used in the first Text REtrieval Conference (TREC). TREC is now in its 20th year, and along with its sister evaluations such as CLEF, NTCIR, INEX, and FIRE, serves as a major testing ground for information retrieval algorithms. She received the 1999 Strix Award from the U.K Institute of Information Scientists for this effort. Starting in 2000 she worked with Paul Over at NIST to form a new effort (DUC) to evaluate text summarization, which has now been folded into the Text Analysis Conference (TAC), providing evaluation for several areas in NLP.

Foundations and Trends® in
Information Retrieval
Vol. 4, No. 4 (2010) 247–375

DOI: 10.1561/1500000009

now
the essence of knowledge

Test Collection Based Evaluation of Information Retrieval Systems

By Mark Sanderson

Contents

Foundations and Trends® in
Information Retrieval
Vol. 4, No. 4 (2010) 247–375

DOI: 10.1561/1500000009

now
the essence of knowledge

Test Collection Based Evaluation of Information Retrieval Systems

Mark Sanderson

The Information School, University of Sheffield, Sheffield, UK
m.sanderson@shef.ac.uk

Abstract

Use of test collections and evaluation measures to assess the effectiveness of information retrieval systems has its origins in work dating back to the early 1950s. Across the nearly 60 years since that work started, use of test collections is a de facto standard of evaluation. This monograph surveys the research conducted and explains the methods and measures devised for evaluation of retrieval systems, including a detailed look at the use of statistical significance testing in retrieval experimentation. This monograph reviews more recent examinations of the validity of the test collection approach and evaluation measures as well as outlining trends in current research exploiting query logs and live labs. At its core, the modern-day test collection is little different from the structures that the pioneering researchers in the 1950s and 1960s conceived of. This tutorial and review shows that despite its age, this long-standing evaluation method is still a highly valued tool for retrieval research.

1

Introduction

An examination of the opening pages of a number of Information Retrieval (IR) books reveals that each author defines the topic of IR in different ways. Some say that IR is simply a field concerned with organizing information [210]; and others emphasize the range of different materials that need to be searched [286]. While others stress the contrast between the strong structure and typing of a database (DB) system with the lack of structure in the objects typically searched in IR [262, 244]. Across all of these definitions, there is a constant, IR systems have to deal with incomplete or *underspecified* information in the form of the queries issued by users. The IR systems receiving such queries need to fill in the gaps of the users' underspecified query.

For example, a user typing "nuclear waste dumping" into the search engine of an academic repository is probably looking for multiple documents describing this topic in detail, he/she probably prefers to see documents from reputable sources, but all he/she enters into the search engine are three words. Users querying on a web search engine for "BBC" are probably looking for the official home page of the corporation, yet they fully expect the search engine to infer that specific information request from the three letters entered. The fact that the

content being searched is typically unstructured and its components (i.e., words) can have multiple senses, and different words can be used to express the same concept, merely adds to the challenge of locating relevant items. In contrast to a DB system, whose search outputs are deterministic, the accuracy of an IR system's output cannot be predicted with any confidence prior to a search being conducted; consequently, empirical evaluation has always been a critical component of Information Retrieval.[1]

The typical interaction between a user and an IR system has the user submitting a query to the system, which returns a ranked list of objects that hopefully have some degree of relevance to the user's request with the most relevant at the top of the list. The success of such an interaction is affected by many factors, the range of which has long been considered. For example, Cleverdon and Keen [61, p. 4] described five.

(1) *"The ability of the system to present all relevant documents*
(2) *The ability of the system to withhold non-relevant documents*
(3) *The interval between the demand being made and the answer being given (i.e., time)*
(4) *The physical form of the output (i.e., presentation)*
(5) *The effort, intellectual or physical, demanded of the user (i.e., effort)."*

To this list one could add many others, e.g.:

- the ability of the user at specifying their need;
- the interplay of the components of which the search algorithm is composed;
- the type of user information need;
- the number of relevant documents in the collection being searched;
- the types of documents in the collection;

[1] This is not to say that researchers haven't tried to devise non-empirical approaches, such as building theoretical models of IR systems. However, Robertson [197] points out that a theory of IR that would allow one to predict performance without evaluation remains elusive.

- the context in which the user's query was issued; and
- the eventual use for the information being sought.

Evaluation of IR systems is a broad topic covering many areas including information-seeking behavior usability of the system's interface; its broader contextual use; the compute efficiency, cost, and resource needs of search engines. A strong focus of IR research has been on measuring the *effectiveness* of an IR system: determining the *relevance* of items, retrieved by a search engine, relative to a user's information need.

The vast majority of published IR research assessed effectiveness using a resource known as a *test collection* used in conjunction with *evaluation measures*. Such is the importance of test collections that at the time of writing, there are many conferences and meetings devoted purely to their use: including three international conferences, TREC, CLEF, and NTCIR, which together have run more than 30 times since the early 1990s. This research focus is not just a feature of the past two decades but part of a longer tradition which was motivated by the creation and sharing of testing environments in the previous three decades, which itself was inspired by innovative work conducted in the 1950s. The classic components of a test collection are as follows:

- a collection of documents; each document is given a unique identifier, a *docid*;
- a set of topics (also referred to as queries); each given a query id (*qid*); and
- a set of *relevance judgments* (often referred to as *qrels* — query relevance set) composed of a list of qid/docid pairs, detailing the relevance of documents to topics.

In the possession of an appropriate test collection, an IR developer or researcher simply loads the documents into their system and in a batch process, submits the topics to the system one-by-one. The list of the docids retrieved for each of the topics is concatenated into a set, known as a *run*. Then the content of the run is examined to determine which of the documents retrieved were present in the qrels and

which were not. Finally, an evaluation measure is used to quantify the effectiveness of that run.

Together, the collection and chosen evaluation measure provide a *simulation* of users of a searching system in an operational setting. Using test collections, researchers can assess a retrieval system in isolation helping locate points of failure, but more commonly, collections are used to compare the effectiveness of multiple retrieval systems. Either rival systems are compared with each other, or different configurations of the same system are contrasted. Such determinations, by implication, predict how well the retrieval systems will perform relative to each other if they were deployed in the operational setting simulated by the test collection.

A key innovation in the IR academic community was the early recognition of the importance of building and crucially sharing test collections.[2] Through sharing, others benefited from the initial (substantial) effort put into the creation of a test collection by re-using it in other experiments. Groups evaluating their own IR systems on a shared collection could make meaningful comparisons with published results tested on the same collection. Shared test collections provided a focus for many international collaborative research exercises. Experiments using them constituted the main methodology for validating new retrieval approaches. In short, test collections are a catalyst for research in the IR community.

Although there has been a steady stream of research in evaluation methods, there has been little survey of literature covering test collection based evaluation. Salton's evaluation section [210, Section 5] is one such document; a chapter in Van Rijsbergen's book [262] another; Spärck Jones's edited articles on IR experiments [242] a third. Since those works, no broad surveys of evaluation appear to have been written; though Hearst has recently written about usability evaluation in IR [116, Section 3]. The sections on evaluation in recent IR books provided the essential details on how to conduct evaluation, rather than reviewed

[2] Indeed, it would appear that the academic IR community is one of the first in the Human Language Technologies (HLT) discipline of computer science to create and share common testing environments. Many other areas of HLT, such as summarization, or word sense disambiguation did not start building such shared testing resources until the 1990s.

past work. There are notable publications addressing particular aspects of evaluation: Voorhees and Harman's book detailed the history of the TREC evaluation exercise and outlined evaluation methods used [280]; a special issue of Information Processing and Management reflected the state of IR evaluation in 1992 [98]; another special issue in the *Journal of the American Society for Information Science* provided a later perspective [253]. More recently, Robertson published his personal view on the history of IR evaluation [199]. However, there remains a gap in the literature, which this monograph attempts to fill.

Using test collections to assess the effectiveness of IR systems is itself a broad area covering a wide range of document types and forms of retrieval. IR systems were built to search over text, music, speech, images, video, chemical structures, etc. For this monograph, we focus on evaluation of retrieval from documents that are searched by their text content and similarly queried by text; although, many of the methods described are applicable to other forms of IR.

Since the initial steps of search evaluation in the 1950s, test collections and evaluation measures were developed and adapted to reflect the changing priorities and needs of IR researchers. Often changes in test collection design caused changes in evaluation measures and vice versa. Therefore, the work in these two distinct areas of study are described together and laid out in a chronological order. The research is grouped into three periods, which are defined relative to the highly important evaluation exercise, TREC.

- **Early 1950s–early1990s**, Section 2: the initial development of test collections and measures. In this time, test collection content was mostly composed of catalogue information about academic papers or later the full-text of newspaper articles. The evaluation measures commonly used by researchers were strongly focused on *high recall* search: finding as many relevant items as possible.
- **Early 1990s–early 2000s**, Section 3: the "TREC ad hoc" period. Scale and standardization of evaluation were strong themes of this decade. The IR research community collaborated to build a relatively small number of large test

collections mainly composed of news articles. Evaluation was still focused on high recall search.

- **Early 2000s–present**, Section 4: the post ad hoc period (for want of a better name). Reflecting the growing diversity in application of search technologies and the ever-growing scale of collections being searched, evaluation research in this time showed a diversification of content and search task along with an increasing range of evaluation measures that reflected user's more common preference for finding a small number of relevant items. Run data gathered by TREC and other similar exercises fostered of a new form of evaluation research in this period: studying test collection methodologies. This research is covered in Section 6.

The one exception to the ordering can be found in the section on the use of significance testing. Apart from a recent book [74], little has been written on the use of significance in IR evaluation and relatively little research has been conducted; consequently, I chose to describe research in this area, in Section 5, more as a tutorial than a survey.

Such an ordering means that descriptions of or references to evaluation measures are spread throughout the document. Therefore, we provide an index at the conclusion of this work to aid in their location.

Note, unless explicitly stated otherwise, the original versions of all work cited in this document were obtained and read by the author.

2

The Initial Development of Test Collections

The genesis of IR evaluation is generally seen as starting with the work of Cleverdon and his Cranfield collections, built in the early 1960s. However, he and others were working on retrieval evaluation for most of the 1950s. In his article looking back over his career, Cleverdon [60] stated that along with a collaborator, Thorne, he created a small test collection in 1953. The intention was to test the effectiveness of librarians at locating documents indexed by different library cataloguing systems when faced with information requests from library users. This work was first described by Thorne two years after it was completed [257].

Thorne described the motivation for conducting this testing in terms that have a strong a resonance with the motivations of IR researchers today. "*the author has found the need for a 'yardstick' to assist in assessing a particular system's merits ... the arguments of librarians would be more fertile if there were quantitative assessments of efficiency of various cataloguing systems in various libraries*". In describing their methodology for testing, Thorne stated "*Suppose the questions put to the catalogue* [from users] *are entered in a log, and 100 test questions are prepared which are believed to represent typically such a log. If the*

test questions are based on material known to be included in the collection, they can then be used to assess the catalogue's probability of success".

The paper listed 50 statements of information need that were used to assess a series of library cataloguing systems. Thorne and Cleverdon tests were essentially a form of *known item searching*. To illustrate, the following is an information need taken from Appendix C of the paper: "*The pressure distributions over the nose of a body of revolution of fineness ratio 6 for angles of attack 0° to 8° at high subsonic Mach number* ($RN > 4 \times 10^{\circ}$)." This request was generated by the authors from a document known to be catalogued in a library. Assessments were based not only on success in finding the known item, but also consideration of the costs of implementing the cataloguing system. Note in Salton's writings, this early test collection was often referred to as Cranfield I (though Cleverdon called a different collection by that name).

In the same year of Cleverdon and Thorne's early efforts, Gull (who published the work in 1956, [97]) also reported building a form of test collection. Composed of 15,000 catalogue entries, the collection was built to compare two library cataloguing systems, each built by a separate group. In total, 98 queries (called requests by Gull) were created and searchers from each group worked to locate as many relevant documents for these requests as possible. Each group formed its own relevance judgments independently, which proved to be problematic, as they discovered that their judgments were quite different from each other based on different interpretations of the queries. Gull stated that one group took a more liberal view of relevance than the other. (Cleverdon stated in [60], that after seeing the problems created by independently formed qrels, he decided to centralize relevance judgments for his collections.)

In the 1950s, computers started to be used for searching of library catalogues. An early mention of "*machines*" being involved in IR was by Kent et al. [155], who proposed an evaluation methodology that they called "*a framework of reference for analyzing the performance of an IR system*". The framework described was similar to a modern test collection. Maron et al. [173] as part of their work in experimenting with

probabilistic indexing described a form of evaluation using a collection of 110 documents and 40 queries. Fels [84] detailed a methodology for testing retrieval effectiveness proposed by Mooers [181]. Bryant [35] briefly described the work by Borko [29] who, according to Bryant, constructed a test collection composed of 612 abstracts. In an appendix of his evaluation survey paper Robertson [195] details a number of other early tests; see also the books from Lancaster [159] and from Spärck Jones [242] for more on the early developments in IR evaluation.

Given that the very first uses of computers for searching only date back to the late 1940s,[1] evaluation of searching systems was clearly an early and important priority for IR researchers. These works, however, are little remembered by today's researchers due to the detailed construction of a test collection that Cleverdon started in the late 1950s.

2.1 Cleverdon's Cranfield Collection

In his reflective piece, Cleverdon cited an editorial from American Documentation[2] (now renamed JASIST) stating that "*evaluation of all experimental results is essential in rating the efficiency of* [IR] *systems*". Cleverdon argued that it wasn't good enough for groups to evaluate their own systems, an independently run evaluation was needed. Consequently, he was funded to test four competing indexing approaches on a collection composed of 18,000 papers [57]. The papers were manually indexed using each of the four classification methodologies. Once the indexes were built, the papers were searched with 1,200 "search questions". The questions were designed to retrieve one of the collection

[1] Holmstrom described a "*machine called the Univac*" capable of searching for text references associated with a subject code. The code and text were stored on a magnetic steel tape. Holmstrom stated that the machine could process "*at the rate of 120 words per minute*" [123]. Note, the UNIVAC isn't generally thought to have come into existence until 1951, the date when the first machine was sold, Holmstrom presumably saw or was told about a pre-production version. See also Mooers — creator of the term information retrieval — for further historical references to mechanical searching devices of the early twentieth century [181].

[2] 1955, "The Truth, The Whole Truth..." American Documentation. Vol. 6 p. 58; it was not possible to locate this editorial.

papers; if that paper was retrieved, the search was considered a success. The collection became known as Cranfield I. Cleverdon reported on the results of his comparison of the four methods and from the experience of this collection, decided to develop Cranfield II.

Cleverdon felt that the relatively large size of Cranfield I was not important in ensuring that measurements were reliable. Therefore, the new collection was composed of 1,400 "documents" (titles, author names, and abstracts) derived from the references listed in around 200 recent research papers. The authors of those papers were contacted and asked to write a question that summarized the problem their paper addressed, these became the collection topics. The authors were also asked to rate each reference in their paper on a scale of 1–5 on how relevant the reference was to the stated question and if possible to provide additional references. Cleverdon's students checked all documents against all questions and contacted the authors of each question asking them if they considered any additional documents found to be relevant. All this work resulted in a collection comprising 1,400 documents, 221 topics, and a set of complete variable level relevance judgments.

Cleverdon was not alone in creating test collections, Salton instigated the creation of a series of test collections: collectively known as the SMART collections (named after the experimental retrieval system that Salton and his students built). In 1968, along with Lesk [164], he described research using two collections, the ADI, a collection of short academic papers, and the IRE-3 collection composed of the abstracts of computer science publications. Later, Salton and Yu [215] described two more: Time and MEDLARS, the first is composed of 425 full-text articles from Time magazine; the second composed of 450 abstracts of medical literature. Note, this MEDLARS collection is different from the test collection with the same name built by Lancaster [158] who in an extensive evaluation of the MEDLARS system created a test collection from 410 actual search requests submitted to the system. Another popular test collection was the NPL created by Vaswani and Cameron [263]. To illustrate the scale of these collections, a number of the more commonly used are detailed in the following table. For details on others, see Spärck Jones and Van Rijsbergen's survey [246].

Name	Docs.	Qrys.	Year[3]	Size, Mb	Source document
Cranfield 2	1,400	225	1962	1.6	Title, authors, source, abstract of scientific papers from the aeronautic research field, largely ranging from 1945 to 1962.
ADI	82	35	1968	0.04	A set of short papers from the 1963 Annual Meeting of the American Documentation Institute.
IRE-3	780	34	1968	—	A set of abstracts of computer science documents, published in 1959–1961.
NPL	11,571	93	1970	3.1	Title, abstract of journal papers
MEDLARS	450	29	1973	—	The first page of a set of MEDLARS documents copied at the National Library of Medicine.
Time	425	83	1973	1.5	Full-text articles from the 1963 edition of Time magazine.

2.2 Evaluating Boolean Retrieval Systems on a Test Collection

With the creation of test collections came the need for effectiveness measures. Many early IR systems produced Boolean output: an unordered set of documents matching a user's query; evaluation measures were defined to assess this form of output. Kent et al. [155], listed what they considered to be the important quantities to be calculated in Boolean search:

- n — the number of documents in a collection;
- m — the number of documents retrieved;
- w — the number that are both relevant and retrieved; and
- x — the total number of documents in the collection that are relevant.

[3] Year is either the year when the document describing the collection was published or the year of the first reference to use of the collection.

Inspired by work from Vickery [265, p. 174], Cleverdon and Keen [61, p. 34] produced a *contingency table* of all possible quantities that could be calculated. The table is reproduced below including Kent et al.'s original labels.

	Relevant	Not-relevant	
Retrieved	$a(w)$	b	$a + b(m)$
Not retrieved	c	d	$c + d$
	$a + c(x)$	$b + d$	$a + b + c + d(n)$

Both Kent et al. and Cleverdon and Keen listed measures that could be created out of combinations of the table's cells. The three that are probably the best known are

$$\text{Precision} = \frac{a}{a+b} \qquad \text{Recall} = \frac{a}{a+c} \qquad \text{Fallout} = \frac{b}{b+d}$$

Where *precision* measures the fraction of retrieved documents that are relevant, *recall* measures the fraction of relevant documents retrieved and *fallout* measures the fraction of non-relevant documents retrieved. Although commonly described in IR text books, fallout is by far the least used in published IR research.

Of all the measures that were proposed, two — precision and recall — dominated evaluation from the start. Reporting on his 1953 test collection work, Gull [97] appeared to be the first to describe recall, measuring competing systems by dividing "*actual retrieval*" by "*optimum retrieval*". Precision and recall were first described together by Kent et al. [155]. In their paper, precision was referred to as a "*pertinence factor*"; recall was called "*recall factor*". Kent et al. stated that neither factor could be used on its own; both measures had to be taken into account to determine effectiveness of a retrieval system. Cleverdon, who called the precision and recall measures, respectively, "*relevance ratio*" and "*recall ratio*" described an inverse relationship between the two [58, pp. 71–72], showing that if one issued Boolean searches that precisely targeted relevant documents and avoided the retrieval of non-relevant, precision would likely be high, but recall would be low. If a query could be broadened in some way to improve recall, the almost inevitable consequence was that more non-relevant documents would

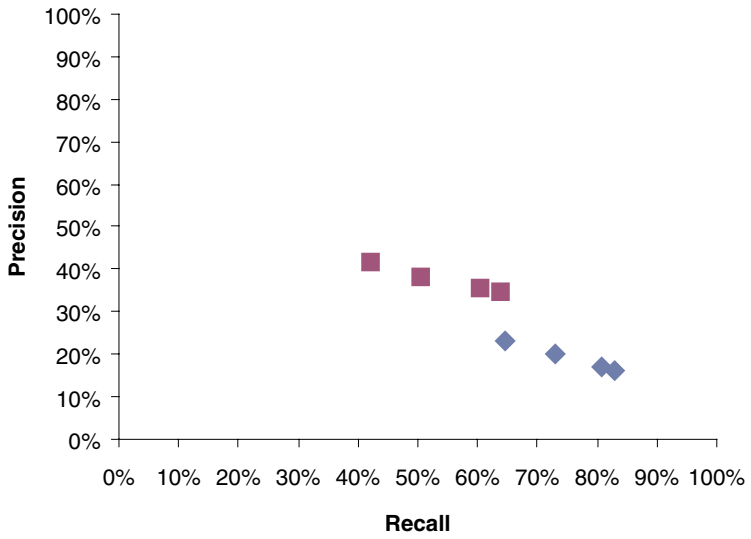


Fig. 2.1 A reproduction of Cleverdon's original recall precision graph, comparing two forms of retrieval.[4]

be returned, causing precision to drop. Using the Cranfield II test collection, he graphed recall/precision data points corresponding to the different Boolean queries; the graph is reproduced in Figure 2.1.

2.2.1 Summarizing Precision and Recall to a Single Value

A great deal of evaluation research addressed the question of how to conflate the two measures into a single value. Van Rijsbergen [261] surveyed a range of such measures. Later in his book, he proposed a measure, which is one minus the weighted harmonic mean of recall and precision, which he called e. Although this measure was sometimes used [73], the weighted harmonic mean was more extensively used in IR literature; it is commonly referred to as f, and is defined as follows.

$$f = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}.$$

[4] Note, the axes and their labels are changed here from the way that Cleverdon drew the graph, so as to reflect the modern convention in presenting such data.

The tuning constant α is used to indicate a preference for how much influence precision or recall has on the value of f. It is common for α to be set to 0.5, which then allows f to be defined as:

$$f = \frac{1}{\frac{1}{2}\left(\frac{1}{R} + \frac{1}{P}\right)} \quad \text{or the equivalent form} \quad f = \frac{2 \times P \times R}{P + R}.$$

2.3 Evaluating over a Document Ranking

The measures described so far work over an unordered set of retrieved documents as would be returned by a Boolean IR system. The development of ranked retrieval — see for example, Maron et al., [173] — required changes in evaluation as the balance of relevant and non-relevant documents varied over the ranking. For any query that was a reasonable reflection of a user's information need, the retrieved documents that matched the query well tended to be mostly relevant and ranked highly. Relevant documents in the collection that matched the query less well appeared further down the ranking mixed in with progressively greater numbers of non-relevant documents.

Swets [249] formally described this situation. He suggested that for any given document ranking, the proportion of relevant to non-relevant documents could be described by two distributions: one for the relevant documents and one for the non-relevant. Swets did not have search output to work with, and so could only speculate on the shape of the distributions: he initially suggested that they would both be normal, though later described other possibilities [250]. Bookstein [28] described potential problems with Swets's model with normal distributions in place. Much later, researchers such as Manmatha, et al. [171] analyzed large sets of ranks and confirmed Swets's formalisms. They found that relevant documents adhered to a normal distribution and the non-relevant followed a negative exponential. Graphs illustrating the distributions of two retrieval systems are shown in Figure 2.2. High scoring documents (on the right of the graphs) are all or nearly all relevant, but for documents that match the query progressively less well (moving to the left), the balance of relevant to non-relevant shifts to a greater proportion of non-relevant being retrieved. The exact nature of the balance

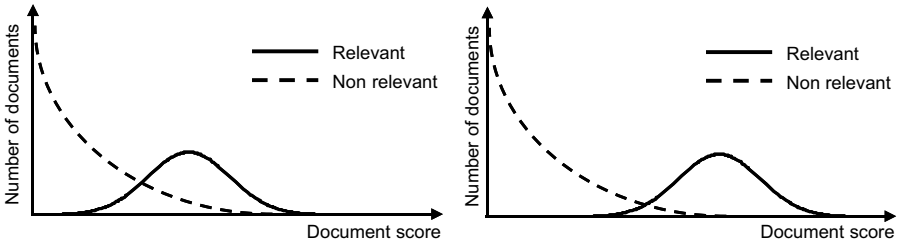


Fig. 2.2 Representations of the balance between relevant and non-relevant documents across a document ranking. The graph on the left represents a less effective retrieval system than the graph on the right.

and the way that it changes across a rank for a particular retrieval system will depend on that system's effectiveness: a good system will produce a ranking that has a strong separation between the distributions (e.g., graph on right of Figure 2.2) and a poor system the opposite (e.g., graph on the left), Swets suggested that this approach could form the basis of an evaluation measure for IR systems, however, the idea was not taken up by the community, who instead choose to focus on adapting precision and recall to ranked retrieval.

2.3.1 Plotting Precision on a Graph

An early popular approach to evaluation of ranked retrieval was to graph precision values, measured over the document ranking averaged across multiple topics. However, as can be seen from the example in Figure 2.3, plotting recall and precision computed at each relevant document for the ranks of two topics results in a scatter plot of discreet points that before being averaged need to be transformed to a pair of continuous functions using interpolation.

Many methods of interpolation were considered by researchers: Cleverdon and Keen [61] defined one; Keen discussed others [151, p. 90]. In the end, one of Keen suggestion's was commonly adopted, which Keen named *semi-Cranfield*, sometimes also called *neo-Cleverdon* Williamson et al. [285]. Here, the interpolated value of precision measured at any recall level (r_i) was defined to be the highest precision measured at any level (r') greater than or equal to r_i. Manning et al. [172]

Rank	Rel	Pr	Rcl
1	1	100%	20%
2	0		
3	1	67%	40%
4	0		
5	1	60%	60%
6	0		
7	0		
8	0		
9	0		
10	0		
∞	10	%	80%
∞	1	0%	100%

Rank	Rel	Pr	Rcl
1	0		
2	0		
3	0		
4	0		
5	0		
6	1	17%	33%
7	0		
8	0		
9	1	22%	67%
10	1	30%	100%

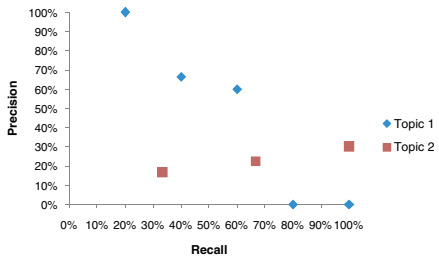


Fig. 2.3 Recall and Precision calculated and plotted for ranks resulting from two topics. In the first topic, there are five relevant documents, two of which were not retrieved; in the second topic, there are three relevant, all of which were retrieved in the top ten of the ranking.

formulated Keen's description thus[5]:

$$P_{interp}(r_i) = \max_{r' \geqslant r_i} P(r').$$

The result of the interpolation method on the points of the two topics in Figure 2.3 can be seen in the graph in Figure 2.4. The values of precision for each function were averaged at a series of pre-chosen recall levels, commonly eleven levels from 0% to 100%; although researchers also used ten levels (dropping the 0%), three levels (25%, 50%, 100%) and twenty-one recall levels (0%, 5%, 10%, 15%, ..., 95%, 100%). The resulting graph of precision averaged across both topics is shown in Figure 2.5. Note, it is common to draw such a graph with a simple interpolated line drawn between the averaged points.

By measuring the precision of every relevant document including those that were not retrieved (implied by measuring precision at recall 100%), there was an assumption in the design of this measure that users were interested in achieving such a high level of recall.

Precision at each of the standard recall levels can itself be averaged and is referred to as *interpolated average precision* or sometimes n-point average precision, where n is the number of recall levels. For

[5] Note, this interpolation measure is sometimes mistakenly thought to be the maximum precision measured between the recall levels r_i and r_{i+1}. See Harmandas et al. [108] and Baeza-Yates and Ribeiro-Neto [18, Section 3] as examples of researchers who made this error.

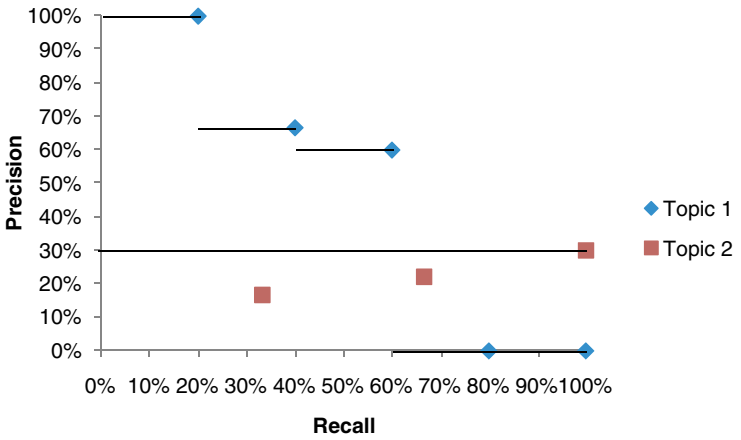


Fig. 2.4 Recall precision graph with interpolation.

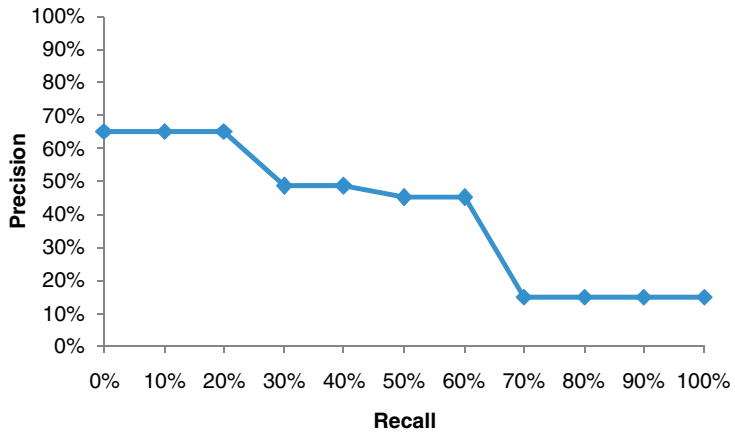


Fig. 2.5 Recall precision graph with precision averaged over two topics.

example, from the graph in Figure 2.5 one could compute 11-point average precision.

Recall precision graphs were a common form of reporting effectiveness: in his book, Salton mentioned little else; in Van Rijsbergen's evaluation section, [262], much space was devoted to the ways of computing such graphs.

At first glance, it might appear that the interpolation was an unusual choice as it ensured that the continuous function could only be

a flat or a monotonically decreasing line. Williamson et al. [285] stated that it was chosen as the standard used by the SMART retrieval system [211]. Ten years later, and it would appear quite independently, Van Rijsbergen also declared Keen's interpolation as the most appropriate to use [262, Section 7]. Both Van Rijsbergen and Williamson stated that they preferred this method over others as it was a conservative interpolation that did not inflate the values of precision for a topic. For topic 1 in Figure 2.4, this would appear to be the case; however, for topic 2, the interpolation would appear to be anything but conservative.

Keen appeared to explain this feature of the interpolation by stating that it computed "*the theoretical maximum performance that a user could achieve*". Van Rijsbergen, in a later personal communication stated that his reasoning for choosing the interpolation was one of normalization against potential errors. The general trend of precision recall graphs was that of a monotonically decreasing line, the increasing precision of topic 2 went against that trend and so should be viewed as an error to be normalized. Van Rijsbergen also stated that at the time, retrieval systems ranked documents based on similarity scores with a coarse granularity. Often, sets of documents were assigned exactly the same score. The order that such documents were ranked by was commonly the order the documents were stored in the IR system; Cooper [65] described this form ranking as a *weak ordering*. Van Rijsbergen's concern was that these blocks of weakly ordered documents caused many of the increases in precision seen for topics. Consequently, he chose the interpolation function as it would normalize such increases.

2.3.2 Another Early Evaluation Measure

Cooper was interested in evaluating ranked retrieval using a single evaluation measure, but was not convinced that finding every relevant document was always the priority of searchers. In his 1968 publication, he stated "*most measures do not take into account a crucial variable: the amount of material relevant to [the user's] query which the user actually needs*"; he went onto say "*the importance of including user needs as a variable in a performance measure seems to have been largely*

overlooked". He proposed a measure called Expected Search Length (ESL) [65], which determined the amount of a ranking that had to be observed by a searcher in order to locate a pre-specified quantity of relevant documents. Cooper was aware that the ranked retrieval algorithms of the time commonly sorted retrieved documents into weak orderings with large numbers of documents being given the same score. Consequently he ensured the measure provided an effectiveness score that accounted for these blocks of equally retrieved documents. Although relatively un-used by researchers at the time, ESL was later influential, most notably in the cumulative gain family of measures from Järvelin et al. described in Section 4.2.1.

2.4 Challenging the Assumptions in the Early Collections

There were a number of common themes to the test collections created in the first few decades of IR research, particularly regarding topics and the definition of relevance. Topics tended to be sentence length statements that mimicked the types of information request issued to librarians. Although not explicitly stated in the literature at the time, there was an assumption that users would query a computer-based IR system in the same way they used the service of a librarian: with a written detailed natural language information request. Early on some researchers pointed out that the assumption was often wrong: Fairthorne stated that users could issue a query that was "*an exceedingly ambiguous phrase*" [83]. However, this disparity between test collection queries and actual user queries wasn't addressed for some considerable time (see Section 4.1 for more detail).

From the initial testing work of Cleverdon and Thorne and of Gull, relevance was assumed to be a form of *topical relevance*, where a user would judge a document as relevant to an information need if its content either partially or wholly addressed the need. Further, search engine users were assumed to be people who would want to find as much relevant information on a particular topic as possible.

Although this straightforward notion of relevance persisted in the test collections that were built, many researchers were aware early on of the potential limitations of these assumptions. Verhoeff et al. [264]

stated that it was highly likely that different users could issue the same query but consider different documents as relevant. They proposed that retrieved documents should be ranked based on the probability of a population of users judging those documents as relevant: the more users who considered a particular document relevant, the higher its rank. Goffman, then considered the interdependence of relevance, pointing out that a document may not be viewed as relevant if retrieved documents containing the same information were previously seen by the user [95]. Such pioneering views on the importance of considering diversity in relevance and redundancy of information was only taken up in earnest much later; see Section 4.1.2.

Cooper [66] proposed that there should be a distinction between topical relevance (in his paper he referred to this as *logical relevance*) and what he called *utility*. He stated that ultimately an IR system needs to be useful and while it is possible to conceive of systems that retrieve a wide range of documents that have some level of topical relevance to an information need, the more important question to ask was which of those documents were actually useful to the user? Cooper pointed out that the credibility of a source or the recency of a document might be important factors in determining the utility of a relevant document to a user. Later [67] he suggested that writing style or even a human assessment of document quality could be a factor in utility. See Saracevic for a broader survey [224] of relevance research.

With hindsight, it can be seen that such suggestions were important features to consider when designing both test collections and search engines. However, it was many decades before such ideas were put into practice and in the intervening period the challenges made to the core assumptions of test collections were largely ignored.

2.5 Assessor Consistency

One of the primary concerns about test collection formation was that the relevance assessments made to form qrels were subjective judgments made by individuals rather than objectively true decisions. An early critic of test collections, Faithorne [83] argued for relevance judgments made by groups rather than individuals. Katter stated "*A recurring*

finding from studies involving relevance judgments is that the inter- and intra-judge reliability of relevance judgments is not very high" [145]. With such low levels of agreement, the concern was that effectiveness scores calculated from test collections using a single set of judgments were not accurate or representative. See also Burgin [41] who detailed past studies on the wide range of influences shown to affect assessment, including the order and way in which documents are presented, the definitions of relevance used, instructions given to assessors, the experience of assessors, etc.

Lesk and Salton [214] studied the question of assessor consistency by gathering pairs of relevance judgments for a test collection composed of 1,268 abstracts and 48 queries: judgments from the creator of the query were paired with those of a "subject expert" who assessed documents independently. The researchers evaluated three configurations of a search engine using different combinations of the paired judgments, determining which configuration was the best. The conclusion of their work was that regardless of the judgments used, the ranking of the different versions of the engine always came out the same. Lesk and Salton analyzed the reasons for this consistency and found that although on average, assessor consistency was low; the disparity between assessors was largely to be found for lower ranked documents. They stated that the reason for this result was that top-ranked documents tended to be most similar to the query, therefore judgments about such documents were easier to make. Most of the effectiveness measures used to assess search engines were more influenced by the rank position of top-ranked documents, therefore the ranking of the three configurations tested was consistent across the different qrel sets.

A similar experiment was conducted by Cleverdon [59] who worked with the Cranfield II collection and its multiple relevance judgments. Like Lesk and Salton, Cleverdon found variations in assessments but also found that they did not impact on the ranking of different configurations of a retrieval system. Burgin [41] also looked at a collection with multiple relevance judgments and again confirmed the Lesk and Salton result, though he cited one work (that we have been unable to locate) that was said to show variations across assessors could impact on ranking of systems. Harman [100] mentioned briefly

an assessor consistency experiment that she reported showed an 80% overlap in relevance judgments.

As a final footnote to this section, it is worth noting that the work here focused on *absolute* judgments of relevance. Rees and Schultz [193], examined the consistency of users at making judgments of documents *relative* to each other. In this test, they reported "*It is evident that the Judgmental Groups agree almost perfectly with respect to the relative ordering of the Documents.*" (p. 185). This early important observation was noted by a number of other researchers, but little work on capturing or exploiting relative judgments was reported until recently, see Section 6.3.4.

2.6 The Practical Challenges of Creating and Using Test Collections

In the early years of IR research, there were a series of practical challenges that faced test collection builders.

2.6.1 The Challenge of Obtaining Documents

The only digitized materials widely available for collection construction were catalogue information about document collections. It would appear that obtaining large quantities of full-text was virtually impossible. The only early test collection with complete documents was the Time collection built by Salton's group. It would appear that Salton got students at his institution to transcribe news articles from copies of the actual magazine. The issues used were preserved by researchers at NIST in the United States and are pictured in Figure 2.6. Later, Salton and Yu implied another collection, (MEDLARS) was also created through manual transcription [215].

2.6.2 Distribution of Collections

Although IR researchers were keen to share test collections, the practicalities of sharing could be challenging. Although in the 1970s and 1980s many institutions were connected to each other via some form of network, across the world, a range of different protocols were used to

Fig. 2.6 The copies of TIME magazine, the articles of which were manually transcribed to form the TIME test collection.

transfer data. Broad adoption of the Internet's TCP/IP beyond North America did not occur until the late 1980s. Removable storage devices such as magnetic tapes were a data transfer option, but a range of formats existed and often only a few devices to read each format were found in an organization such as a University. Because of these obstacles, sharing of collections between research groups was patchy and ad hoc.

No broadly applicable solution to distribution was found until the early 1990s, when the first large-scale distribution of test collections was achieved with the creation of the Virginia Disc One: a CDROM containing many of the commonly used test collections [86]. Several hundred copies of the discs were distributed world-wide.[6] With the increased ubiquity of networks and data transfer protocols, by the early 1990s, networked-based distribution of collections started; an early example of which is the University of Glasgow IR group's test collections web page, created in 1994 by Sanderson.[7]

2.6.3 The Challenge of Scale — Limited by qrels

Commercial IR systems were by the early 1960s searching the subject keywords of several tens of thousands of documents [78]. By the mid-1970s, it is recounted that searching hundreds of thousands of documents was routine [23], yet the test collections of the time were orders of magnitude smaller. A key reason for this appears to be researchers'

[6] A personal communication to the author.
[7] http://ir.dcs.gla.ac.uk/resources/test_collections/ (accessed April 2010).

insistence on knowing the total number of relevant documents for a query.

Cleverdon, when building Cranfield II, employed people to manually scan the full collection for all relevant documents for all topics. Others investigated less resource intensive approaches. Kent et al. [155] and later Bornstein [32] proposed collection sampling to locate some of the missing relevant and estimate the numbers remaining unfound. Maron et al. [173, p. 79] described a method of using multiple queries generated by searchers to create a set of retrieved documents that were then assessed for relevance, this approach was also suggested by Salton [210, p. 294]. Lancaster [158] used subject experts to both search and draw on their knowledge of existing documents to build up what he called the "recall base". Fels [84] stated that Mooers [180] proposed creating a form of known item search test collection. The methodology, which Fels tested, involved randomly sampling documents from a collection and creating topics that would be highly relevant, relevant or irrelevant to the sampled document. These topics would then be issued to a searching system and tests of success or failure would be determined by the presence or absence of the known items.

Despite a plethora of suggestions listed above, however, there appears to be little evidence of researchers actually trying these suggestions to build bigger test collections. Neither does there appear to be a willingness to give up on the notion of finding all relevant documents as advised by Cooper [67]. The conclusion amongst IR researchers at the time was that a way had to be found to produce larger test collections while at the same time locate as many relevant documents as possible.

Spärck-Jones and Van Rijsbergen proposed [245] a methodology for building larger test collections (what they referred to as an *ideal test collection*). Their proposal was motivated by concern that existing test collections were not only small but often carelessly built and/or inappropriately used by researchers (p. 3 of their report). They proposed the creation of one or more large test collections built using well-founded principles and distributed to the community by a common organization. Addressing the problem of assessors not being able to judge every document in large collections they proposed a solution using a technique they referred to as *pooling*.

Spärck-Jones and Van Rijsbergen suggested that for a particular topic, assessors judge the documents retrieved by "*independent searches using any available information and device*" [245, p. 13]. Pooling would create a small subset of documents containing a sufficiently representative sample of relevant documents. The relationship of pool size and its impact on the accuracy of comparisons between retrieval systems was analyzed later in some detail by Spärck-Jones and Bates [244, p. A31]. In order to manage the number of relevance judgments needing to be made, the later report also described random sampling from a pool (see pp. 20–21).

The impact of the ideal test collection report was initially limited. Some further small collections were built using exhaustive relevance judgments, such as the CISI [85] and the LISA collections [76]. Some collections were built using pooling but did not appear to be aware of Spärck-Jones and Van Rijsbergen's work. Katzer built the INSPEC test collection, composed of 12,684 abstracts [146, 147]. Katzer stated that 84 topics were created for the collection and relevance judgments made on a pool of documents composed of the union of seven searches conducted by search intermediaries. Later Salton augmented the pool with the ranked document lists of two retrieval systems configured to use different ranking algorithms [213, p. 1030]. Although earlier evaluation was conducted using approaches similar to pooling, such as Lancaster's MEDLAR tests, INSPEC appears to be the first shared test collection built using pooling. Fox described another collection, the CACM, composed of 3,204 documents where Katzer et al.'s seven search variations were used to build a pool for the collection [85].[8] Fuhr and Knorzs built a 300 query, 15,000 document collection with pooling [91]. Blair and Maron [25] later Blair [24], constructed a test collection for estimating the true recall of a Boolean search engine, using a series of broad searches to locate as many relevant documents as possible. None of this work cited Spärck-Jones and Van Rijsbergen.

Details of these somewhat larger collections are provided in the following table. By contrast, commercial search engines were by that

[8] Note, the literature is a little confused on how the CACM collection's relevance judgments were formed. A later article briefly mentioned the CACM stating that all documents were examined for relevance [213, p. 1030].

time routinely searching multi-million document collections and calls were made by the industry for the research community to start testing on larger data sets [160]. A few research groups obtained such collections: researchers in Glasgow used 130,000 newspaper articles for testing a new interface to a ranked retrieval system [218]; IR researchers at NIST conducted experiments on a gigabyte collection composed of 40,000 large documents [107]; and Hersh et al. [118] released a test collection composed of around 350,000 catalogue entries for scholarly articles.

Name	Docs.	Qrys.	Year[9]	Size, Mb	Source document
INSPEC	12,684	77	1981	—	Title, authors, source, abstract, and indexing information from Sep to Dec 1979 issues of Computer and Control Abstracts.
CACM	3,204	64	1983	2.2	Title, abstract, author, keywords, and bibliographic information from articles of Communications of the ACM, 1958–1979.
CISI	1,460	112	1983	2.2	Author, title/abstract, and co-citation data for the 1,460 most highly cited articles and manuscripts in information science, 1969–1977.
LISA	6,004	35	1983	3.4	Taken from the Library and Information Science Abstracts database.

Spärck-Jones and Van Rijsbergen's *ideal test collection report* is often cited for its introduction of the idea of pooling, however, the researchers had more ambitious goals. On page 2 of the report can be found a series of recommendations for the IR research community:

(1) "*that an ideal test collection be set up to facilitate and promote research;*

[9] Year is either the year when the document describing the collection was published or the year of the first reference to use of the collection.

(2) *that the collection be of sufficient size to constitute an adequate test bed for experiments relevant to modern IR systems...*
(3) *that the collection(s) be set up by a special purpose project carried out by an experienced worker, called the Builder;*
(4) *that the collection(s) be maintained in a well-designed and documented machine form and distributed to users, by a Curator;*
(5) *that the curating (sic) project be encouraged to, promote research via the ideal collection(s), and also via the common use of other collection(s) acquired from independent projects.*"

This vision of larger test collections built by a curating project that fostered their use in research was finally realized in the formation of TREC.

3

TREC and Its Ad Hoc Track

In 1990, the US government agency DARPA funded the National Institute of Standards and Technology (NIST) to build a large test collection to be used in the evaluation of a text research project, TIPSTER. In 1991, NIST proposed that this collection be made available to the wider research community through a program called TREC — the Text REtrieval Conference. The annual evaluation event started in November, 1992. Operating on an annual cycle, the multiple goals of TREC were to:

- create test collections for a set of retrieval tasks;
- promote as widely as possible research in those tasks; and
- organize a conference for participating researchers to meet and disseminate their research work using TREC collections.

In the early years, TREC organizers annually created gigabyte-sized test collections, each with 50 topics and a set of qrels built using pooling, see Voorhees and Harman [280] for a detailed history of the exercise. As can be seen in the overlap between the TREC goals and the Spärck-Jones and Van Rijsbergen recommendations, the ideal test collection document appeared to have influenced the construction

of TREC, however, its initiators, headed up by Harman, still had to instantiate them. Key to making TREC a success was their solution to gathering the independent searches that Spärck-Jones and Van Rijsbergen described.

Harman and her colleagues appear to be the first to realize that if the documents and topics of a collection were distributed for little or no cost, a large number of groups would be willing to load that data into their search systems and submit runs back to TREC to form a pool, all for no cost to TREC. TREC would use assessors to judge the pool. The effectiveness of each run would then be measured and reported back to the groups. Finally, TREC would hold a conference where an overall ranking of runs would be published and participating groups would meet to present work and interact. It was hoped that a slight competitive element would emerge between groups to produce the best possible runs for the pool.

The benefits of the "TREC approach" were that research groups got access to new test collections; and at the conference, compared their methods against others. The benefits of the approach to TREC were that their chosen area of IR research became a focus of interest among the research community. The benefit of the approach to all was that new test collections were formed annually for all of the IR community to use. Other fields of Human Language Technology used such collaborative/competitive approaches before TREC: e.g., the Message Understanding Conference [96]. However, the continued running of TREC, now approaching its third decade, is a testament to the particular success of the approach Harman and her colleagues applied to IR.

TRECs had a profound influence on all aspects of evaluation, from the formatting of test collection documents, topics, and qrels, through the types of information needs and relevance judgments made, to the precise definition of evaluation measures used. In particular, the first eight years of TREC, when the *ad hoc track* was run, established the norms on which a great deal of other TREC and broader IR evaluation work was based. Consequently that period in TREC is described here in some detail. The section starts with an explanation of how each of the three components of a test collection was created followed by a detailing of some of the tasks that TREC chose

to focus on in its early years. Finally, the evaluation measures used are explained.

3.1 Building an Ad Hoc Test Collection

The TREC ad hoc collections were built with the searching task of an information analyst in mind: a person who was given topics to search for on behalf of someone else Harman [105]. The topic given to them was well described and the analyst was expected to locate as much relevant material as possible. The topics for the ad hoc track were created by members of the TREC document assessment team at a rate of 50 per year. The numbers and exact procedures for forming the topics varied over the eight years of TREC ad hoc, Voorhees and Harman [280, p. 28]. However, certain aspects of the method remained constant. The creators of the topics would create a set of candidate topics, these were then trialed by searching on the ad hoc collections to estimate how many relevant documents each topic would return. Topics with too many or too few relevant documents were rejected [99, 278].

TREC topics were structured to provide a detailed statement of the information need that lay behind the query, which was intended to ensure that the topic was fully understood. The topics were formatted into an XML-like scheme (Figure 3.1), the structure of which varied over the years, but its main components were:

- a topic *id* or number;
- a short *title*, which could be viewed as the type of query that might be submitted to a search engine;
- a *description* of the information need written in no more than one sentence; and
- a *narrative* that provided a more complete description of what documents the searcher would consider as relevant.

Obtaining large quantities of text to build a collection involved persuading the copyright owners of a large corpus of material to allow their content to be used. Through connections with news publishers, TREC organizers obtained US and UK newspaper and magazine articles, as

```
<top>

<num> Number:  200

<title> Topic: Impact of foreign textile imports on U.S. textile industry

<desc> Description:  Document must report on how the importation of foreign
textiles or textile products has influenced or impacted on the U.S. textile
industry.

<narr>  Narrative:  The impact can be positive or negative or qualitative.
It may include the expansion or shrinkage of markets or manufacturing volume
or an influence on the methods or strategies of the U.S. textile industry.
"Textile industry" includes the production or purchase of raw materials;
basic processing techniques such as dyeing, spinning, knitting, or weaving;
the manufacture and marketing of finished goods; and also research in the
textile field.

</top>
```

Fig. 3.1 Example TREC ad hoc topic.

well as US government documents. TREC standardized the gathered documents in a similar XML scheme as used in the topics.

The documents and topics were sent out to participating groups who were given a limited time to generate and return a series of runs. Each run contained a maximum of 1,000 top-ranked documents retrieved for each of the TREC topics. The top n documents (most often $n = 100$, or more recently 50) from each run were merged into the pool to be judged. In order to make pool judgment tractable, TREC organizers sometimes had to limit the number of runs that contributed to the pool. In such situations, participating groups nominated a subset of submitted runs to be assessed.

TREC defined two types of run:

- *automatic* runs, defined as runs where no manual intervention took place between the submission of topics to a group's retrieval system and the outputting of the run.
- *manual* runs, where any amount of human intervention in the generation of search output was allowed. For some manual runs, the list of documents submitted was a concatenation of the best results from multiple queries. For details of how individual manual runs were created, see Voorhees and Harman's overview of one of the years of TREC (e.g., [99,

100, 101, 277, 278]). Although such runs appeared to have limited scientific value, TREC organizers encouraged their submission as they were found to be rich sources of relevant documents for the pool. Kuriyama et al. [156], showed the importance of manual searching in effective pool formation.

In order to be seen to be fair to all participants, TREC assessors viewed all top n documents in the pool; documents were sorted by docid so that the rank ordering of documents did not impact on the assessment. TREC organizers tried to ensure that the creator of the topic was also the assessor of its qrels. Unlike a number of earlier test collections, which had degrees of relevance, in TREC, assessors made a binary relevance judgment. They were instructed that "*a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document)*",[1] which resulted in a liberal view of what documents were viewed as relevant.

TREC, particularly, its ad hoc collections continue to have a profound impact on the academic IR community. The collections are used extensively: a search on a well-known scholarly search engine (conducted in May 2010) revealed that the phrase "TREC collection" occurred in nearly 1,210 papers; in a small survey conducted for this paper, examining 40 of the 60 papers in ACM SIGIR 2004, 28 of the papers used TREC collections (70%), 17 of which used at least one of the ad hoc collections (43%). These multi-gigabyte data sets became the de facto standard on which many new ideas were tested.

3.2 Classic TREC Ad hoc Measures

TREC was not only influential on test collections used by researchers but also on the effectiveness measures used. Although many measures were calculated by TREC organizers, three, MAP, R-precision, and MRR are commonly used in the broader community. Precision measured at a fixed rank, $P(n)$, although used before TREC was another measure adopted by the evaluation exercise and an important result comparing the properties of MAP and $P(n)$ was described for the first

[1] http://trec.nist.gov/data/reljudge_eng.html (accessed April 2010).

time at TREC meetings. The measures are detailed here. Following the chronological ordering of this review, more recent ad hoc measures are described later in Section 4.2.

3.2.1 Average Precision

In the first year of the exercise, TREC organizers calculated 11-point average precision using Keen's interpolation function. However, perhaps because weak orderings of rankings were less common by the early 1990s, this was soon replaced by a non-interpolated version. The first reference to this measure was in Harman [99], where the measure was called *non-interpolated average precision* (AP). It is defined as follows:

$$AP = \frac{\sum_{rn=1}^{N} (P(rn) \times rel(rn))}{R}.$$

Here, N is the number of documents retrieved, rn is the rank number; *rel(rn)* returns either 1 or 0 depending on the relevance of the document at rn; *P(rn)* is the precision measured at rank rn; and R is the total number of relevant documents for this particular topic. Simply, the measure calculates precision at the rank position of each relevant document and takes the average. Note, by summing over the N and dividing by R, in effect, precision is measured as being zero for any unretrieved relevant documents. This measure is similar to normalized precision [210, p. 290].

If one calculates AP for each of a set of topics and takes the mean of those average precision values, the resulting calculation is known as *Mean Average Precision* (MAP). Harman's original definition was published with a mistake, replacing the denominator R, with the number of relevant documents retrieved; a journal version of the paper contained the same error [102]. Voorhees appeared to be the first to describe the measure as mean average precision [267], though it took several years for MAP to become its universally accepted name. MAP became one of the primary measures used in many evaluation exercises as well as a large quantity of published IR research.

Note that, interpolated average precision (described in Section 2.3.1.) was often in older literature referred to as average precision or AP, which can cause confusion for the modern reader. Occasionally,

more recent papers and books appear to use interpolated AP where it would appear that the authors were unaware of the existence of the more established non-interpolated version.

3.2.2 Measuring Precision at a Fixed Ranking

A common option for measuring precision is to decide that a user will choose only to examine a fixed number of retrieved results and calculate precision at that rank position.

$$P(n) = \frac{r(n)}{n},$$

where $r(n)$ is the number of relevant items retrieved in the top n. The choice of n is often influenced by the manner of their display, P(10) being the commonest version. Sometimes the measure P(1) is used and referred to as the *Winner Takes All* (WTA) measure. Precision measured at a fixed rank has long been described in the literature. Both Salton [210], and Van Rijsbergen [262] mentioned calculating precision at a fixed rank. However, both described the measure in the context of producing graphs of precision over a range of ranks. Neither described the measure in the way it was used later on: a single value measured at one point in the ranking.

Note P(n) ignores the rank position of relevant documents retrieved above the cutoff and ignores all relevant below. Also if a topic has fewer than n relevant documents in the collection being searched, P(n) for that topic will always be < 1. However, there is little evidence these features of the measures are problematic.

However, there is one feature that is worth noting, the importance of which was described in one of the later years of TREC. Computing precision at a fixed rank ignores the total number of relevant documents there are for a topic. This number can affect the expected value of precision calculated at a fixed rank n. To illustrate, if we calculate P(10) on the ranks resulting from retrieval on two topics, one with 1,000 relevant documents, the other with 10. For both topics, the balance of relevant and non-relevant documents will change across the resulting ranks. However, in the first topic, it should be relatively straightforward for a retrieval system to place a great many relevant documents in the

top 10. For the second topic, the chances are that there will be fewer easy to retrieve relevant documents; consequently, it will be harder for a retrieval system to fill the top 10 with just relevant. In other words, IR systems are likely to get a high P(10) on the first topic and a low P(10) on the second. An improved system finding one more relevant documents for the first topic will score the same increase in P(10) as another system that finds one more relevant for the second topic, even though locating the extra relevant for the second topic was most likely algorithmically harder to achieve.

When evaluating within a particular test collection there does not appear to be any published evidence that this feature of the measure causes problems. However, there have been evaluations across two test collections, where differences in measurement arose. The effect was highlighted during the running of the *Very Large Collection* (VLC) track of TREC-6 [115]. Participating groups applied their retrieval systems to a 20 GB ad hoc collection and a 2 GB subset. Effectiveness was measured using P(20). Across all seven participating groups, P(20) was higher for searches on the 20 GB collection than on the subset; on average 39% higher. Hawking and Thistlewaite noted the increase, but did not study it. The following year the track was re-run, using a larger 100 GB collection along with 1% and 10% subsets, a similar increase in P(20) was noted [112].

Reasons for the increase were discussed at the TREC meeting. Consequently, Hawking and Robertson [114] studied the results in detail, postulating a number of hypotheses. They concluded that the core reason was that searching on a larger collection resulted in there being more relevant documents per topic and a consequent increase in the number of such documents that could be highly ranked. As a means of final confirmation, Hawking and Robertson examined the effectiveness of the systems participating in the VLC track using MAP, which measures precision across all retrieved and un-retrieved relevant documents. With this measure, no increase in effectiveness was observed.

Given that P(20) behaved completely differently from MAP, one might ask, is one measure better than the other? Hull discussed the qualities of the two approaches [125]. He pointed out that the answer to which is the best, depends on how users are going to use a search

engine. If the user is (like most web searchers) focused on obtaining a few relevant documents and examined only the first page of results, then a fixed rank version of precision seems more appropriate. If a search engine was able to increase the size of the collection it retrieved over, such users would view the resulting increase in relevant documents in the first page of a search engine as a clear improvement.

The situation would be different if the users of the system were, for example, patent searchers whose goal was to locate every relevant document in the collection. When the collection being searched was increased in size, such searchers would probably value the growth in the total number of relevant documents, but they might not view the engine as having improved because across all the relevant documents viewed by such a thorough searcher, the proportion of non-relevant to relevant would be unchanged.

Here, the rank cutoff version of precision appears to be the better choice in most situations. As will be seen later in Section 6.4, it would be rash to assume one version is always better than another: when the measures are compared in other contexts or used for alternate purposes, different conclusions are often drawn.

3.2.3 R-precision

Instead of calculating precision over the same fixed rank for a set of topics, one could use a different cut off for each topic; R-precision uses this approach. It is calculated as P(R), where R is the total number of relevant documents in the collection for a particular topic. Precision is calculated at the rank position where a perfect retrieval system would have retrieved all known relevant documents, a more consistent recall level than a fixed rank. The measure was first used in TREC-2 ([99], Appendix A).

Note that at the point R, the number of relevant documents ranked below R will always equal the number of non-relevant documents ranked above R, which has led others to refer to R as the *equivalence number* and called R-precision *missed@equivalent* [189]. Note also that all forms of AP and, as pointed out by Aslam et al. [14], R-precision approximates the area under a recall precision graph.

3.2.4 Searching for a Single Relevant Document

Known item search describes retrieval based on topics that have one relevant document in the collection being searched. It was first described in Thorne's original test collection paper [257]. Mean Reciprocal Rank (MRR) was created by Kantor and Voorhees[2] [143] to assess such retrieval. The measure calculates the reciprocal of the rank of the first relevant document in a ranking.

The Reciprocal Rank (RR) calculated over the four example rankings shown in Figure 3.2, is respectively, 1, 0.5, 0.5, and 0. Note how the measure is particularly sensitive to small changes in the location of the relevant document at top ranks: the RR for the second example is half of that of the first. Conversely, the measure is insensitive to large difference in low rank. Because any other relevant documents in a ranking are ignored, the RR is the same for the second and third examples. The MRR has been used in some evaluations, for example, it was used in a known-item search task in TREC [279].

3.2.5 Standardizing Measure Calculation

The organizers of TREC recognized that another important role it could play was to act as a supplier of a standard tool to calculate the effectiveness of a retrieval run. Thus it did with the public release of *trec_eval*: an application that, given a run and a set of qrels, calculates an extensive range of effectiveness measures over the run. The tool is used by many research groups and is generally viewed as holding

Rank	**Rel**	**Rank**	**Rel**	**Rank**	**Rel**	**Rank**	**Rel**
1	1	1	0	1	0	1	0
2	0	2	1	2	1	2	0
3	0	3	1	3	0	3	0
4	0	4	0	4	0	4	0
5	0	5	1	5	0	5	0

Fig. 3.2 Four example documents ranks.

[2] See also Kantor and Voorhees [144] for a more complete description of the work.

the definitive definition of many of the measures used by the IR community.[3]

3.3 The Other TREC Tracks and Uses of TREC Collections

Although the test collections and other associated data resulting from TREC ad hoc is the strongest legacy of TREC in the 1990s, test collections for many other documents types were developed at the same time. Voorhees and Harman [280, pp. 8–9] detailed both the names and nature of them in their book, some of the more significant tracks focused on:

- categorizing and/or retrieving streamed text, as addressed in the routing and filtering tracks [200];
- medical scholarly articles in the TREC-based genomics collection where matching to variants of gene names became a part of the search task [119];
- search across languages with English queries retrieving Spanish and Chinese documents, as covered in the cross language search tracks [230]; and
- retrieval of noisy channel data, output by OCR and speech recognizer systems, addressed in the Confusion [144] and Spoken Document Retrieval tracks [92].
- Later, within the field of distributed IR, groupings of TREC ad hoc collections were established into a set of commonly used collections by that research community. Shokouhi and Zobel [229] detail six such collated collections and the researchers who initially built them.

Some tracks established their own evaluation measures or methods — see the measures in the filtering track overview [200] or the differing methods of the interactive track [186]. However, across the majority of the TREC tracks in this period, the type of search task established in the ad hoc track was highly influential on the

[3] There does not appear to be any paper or technical report describing trec_eval. It can be downloaded from the following URL: http://trec.nist.gov/trec_eval/ (accessed April 2010).

topic design, definition of relevance, evaluation measures, and pooling methodologies used.

3.4 Other Evaluation Exercises

The success of TREC inspired many others to start similar evaluation exercises:

- CLEF[4] — The annual Cross Language Evaluation Forum focuses on search across European languages; though in recent years it diversified into other languages including Persian and some of the languages of the Indian sub-continent [33]. Search of other objects such as images has also been addressed: imageCLEF [64].
- NTCIR — The NII Test Collection for IR Systems is an evaluation exercise held every 18 months in Japan. NTCIR has focused on cross-language search for Asian languages such as Japanese, Chinese, and Korean. A particular focus was on patent search [141]. The first NTCIR evaluation exercise used a collection of the title and abstracts of several hundred thousand scholarly articles [142].
- TDT — Topic Detection and Tracking was an exercise examining the automatic detection and tracking of important emerging stories in streaming text [5].
- INEX — The INitiative for the Evaluation of XML Retrieval examines the retrieval of semi-structured documents, in particular focusing on retrieval of document fragments [157].
- TRECVID — an evaluation exercise focused on video retrieval [232].
- A number of other smaller and/or newer evaluation exercises were created, a number of which presented their work at the First International Workshop on Evaluating Information Access [219].

[4] Pronounced "clay", from the French word for key.

3.5 TREC's Run Collection

In addition to the test collections, topics and qrel sets generated each year by TREC, the runs (ranked lists of the documents retrieved for each topic) submitted by participating groups for each track were also archived. In the ad hoc track for TRECs 2–8, nearly 500 runs were archived. As will be seen in Section 6, this archive opened up new opportunities for research examining the impact of different evaluation measures and for exploring the effectiveness of test collection formation methodologies. Other evaluation exercises also archived their run data, see Sakai [206] for use of NTCIR runs and Sanderson et al. [221] for an example of use of runs from CLEF.

3.6 TREC Ad Hoc: A Great Success with Some Qualifications

TREC and its spin-off evaluations had a profoundly positive impact: providing large-scale test collections, a pooling method, evaluation measures, and other data sets to a research community that up to the formation of TREC did not appear to have the appetite or resources to build its own. The collections, particularly those from the ad hoc track, are extensively used. Ten years after the track stopped, it is still common to see the collections exploited in high impact research. The inaugural running in 2008 of the Indian languages evaluation exercise, FIRE (Forum for IR Evaluation) used collections and a topic design strongly influenced by the TREC ad hoc paradigm [170]. Because the ad hoc collections and methodologies continue to be widely used, it is worth reviewing some of the methods employed in those early years, focusing on collections, topics, and relevance.

3.6.1 Collections

The documents of test collections in this period were commonly newspaper articles. This tradition started with Salton's TIME collection, but was carried on by TREC and later other evaluation campaigns. TREC started in 1992. In late 1993, public web search engines were being created [154, p. 152]; by the summer of 1994, a large number were

in existence including the relatively well-known Lycos and Excite. However, TREC and the broader academic IR research community maintained their focus on search from almost nothing but newspaper and government documents for much of the 1990s. Using this form of content had a strong influence on the type of topics that were created for the test collections, which in itself limited the range of search tasks that were addressed by IR researchers in the 1990s. Some have criticized TREC organizers for not moving more quickly to the study of web search.

However, it is worth considering at least one of the contributing factors to this situation. The building of web and particularly enterprise collections for use in a large-scale evaluation is a difficult legal and privacy challenge. A web crawler has no automatic way of knowing the copyright status of the pages it is downloading. Just because an item is placed on a freely accessible web page does not mean the creator of that page is giving permission to others to copy and redistribute it. Although now, precedent has established that crawling and storing most web content is unlikely to be a copyright violation, in the 1990s this was not as clear. TREC took a cautious legal approach to such matters, which was undoubtedly a factor in the delay in adapting to Web tasks.[5]

3.6.2 Topic Selection

As a general rule one would expect the components of a test collection to be a sample of the documents and queries typically submitted to an IR system. When creating their original testing environment, Thorne [257] suggested sampling from a log of queries for example. The processes used for topic selection in the ad hoc track of TREC were

[5] One might view such caution as an inhibition to research. However, ignoring such concerns is not without risk: in a personal communication with the author, the head of a research group who were regularly crawling blogs (respecting conventions and protocols) hit problems when a blog owner went to the press claiming the researchers were spying, causing the head significant work in placating his University and the blog owner. A better-known example was the release of query logs from AOL in 2006. Although the logs were anonymized, sufficient session information was preserved to allow some people to be identified [21]. The consequence of this privacy failure was the sacking of two employees and the resignation of a company executive [292].

intended to obtain a representative sample. However, there was no log of existing searches on the document collections being built, therefore, topics had to be created. As described in Section 3.1, certain topics with too few or too many relevant documents were avoided; the later because of concerns that relevance assessors would be overwhelmed with documents to assess. However, by focusing only on topics that had a middling number of relevant documents associated with them, there was a danger of introducing bias into the topic sets.

Although TREC topics had a title, which mimicked a short user query, groups commonly submitted runs that were based on a combination of the topic's title and description fields. However, Rose and Stevens [204] described research based on query logs of web search engines showing that 53% of queries consisted of just one word; far shorter than the TREC topic titles of the time. As a reaction to this, the length of topic titles became shorter in subsequent years of TREC and participating groups were encouraged to submit runs based only on titles; as detailed by Voorhees and Harman [280, p. 39]. Nevertheless many still used the full-text of the topics in experiments despite their apparent lack of realism.

It is notable that while the length of topics was addressed, ambiguity was consciously avoided: in the TREC-5 overview, it was stated that "*Candidate topics were ... rejected if they seemed ambiguous*" and this approach persisted for many years, not just by TREC, but by most other evaluation campaigns. Attempts at building such a collection in the interactive track of TREC were made in the 1990s [187], however, the methodologies used there were not picked up by others in TREC or the wider search evaluation community until much later.

3.6.3 Binary Relevance

For many of TRECs early years, the focus was on topics locating as much information as possible even to the point of seeking documents with only a single relevant sentence. Sormunen [240] re-examined TREC relevance judgments for 38 topics from TREC-7 and 8 using three relevance grades: not relevant, marginally relevant and highly relevant. He reported that about 50% of the relevant documents were

marginally relevant and questioned if it was right for these commonly used test collections to be so strongly composed of this form of relevant document. While giving a talk about test collections and the TREC definition of relevance in 2000, a member of the audience who, at the time, was working for part of the UK intelligence community claimed (to me) that TREC's definition was the same as the one used by intelligence analysts in his organization. It would appear that this liberal definition of relevance was an appropriate definition for the information analyst task that TREC ad hoc was originally created for. Whether it was an appropriate definition for many of the information-seeking tasks carried out by other users is perhaps open to question.

3.6.4 Summing Up

It is worth reiterating that these qualifications are in the context of a highly successful on-going evaluation exercise. It is in many ways because of the success of TREC that the issues are highlighted here. TREC test collections particularly ad hoc were not only used but also imitated. Although as will be seen in the next section, TREC organizers and many others moved to address the problems listed here, there is a danger that other test collection creators sometimes too faithfully reproduced the early TREC approach. The user of any test collection would do well to examine overview documentation to understand fully the way the collection they intend to experiment on was built.

3.7 Conclusion

In this section, the initial development of large-scale test collections using a pooling approach for building qrels was described and the measures used to assess effectiveness were described. The innovation primarily from TREC of keeping run data and encouraging its use for a new form of experimentation was also described, before finally detailing some of the concerns over ad hoc test collection design.

4

Post Ad Hoc Collections and Measures

Although ad hoc-style test collections continue to be used and created, from the early 2000s, development of new test collections started with different document content, addressing new tasks, and beginning to employ novel evaluation measures. One of the motivations for this was a realization that TREC's ad hoc design had limitations.

An example of this appeared when some unusual results emerged from an early TREC web test collection: [110, 111]. Although the collection content was different from classic ad hoc collections, the topics were similar to those used in the past. One striking result from these collections was that link-based methods, such as PageRank, appeared to provide little or no value. Broder pointed out that many web queries were different from the classic information seeking view of search [34]. So-called navigational queries where users seek a home page did benefit from link-based methods, but they weren't present in the existing TREC web collections. Consequently, the organizers introduced different types of topics into subsequent collections, focusing particularly on locating named home pages. This was later generalized into the so-called *topic distillation* task: finding a series of home pages relevant to a particular topic.

It was now clear that testing different searching applications required different document collections and different types of searching task. In this section, the tasks and collections that were introduced in this post ad hoc period are described, followed by details of the new measures.

4.1 New Tasks, New Collections

At the same time that the TREC web track was being developed, the *Question Answering* (QA) track was created. Here the search task was for a QA system to interpret an inputted question and locate passages within documents that answered the question [269]. Searching for passages within documents was also explored in other QA tracks run in other evaluation exercises [169] as well as the novelty track of TREC [103]. Another evaluation exercise to explore search of document parts was INEX (INitiative for the Evaluation of XML Retrieval), where search of different collections of XML data was examined [90]. Here the focus was on evaluating searching systems for not only their ability to retrieve relevant structured documents, but also to locate the best point in a document's structure where a user could start reading the relevant part of the document.

Search of new web-based content was examined in the blog track of TREC, where several hundred thousand blog feeds were crawled to form collections composed of millions of postings. Here the search tasks examined a form of topic distillation to locate relevant feeds addressing a particular topic. Blog topics were sampled from a blog search engine log. In addition, detection of the opinions expressed in blogs became part of the search task [185].

Search tasks associated with organizations were explored in the Enterprise track, which examined search of email discussion threads as well as using a multi-faceted collection of documents from an organization to create a search task for location of experts in particular topics [70]. Multimodal search tasks started to be explored first off in the video track of TREC — TRECVid [233, 231] — and then in the image searching track of CLEF — ImageCLEF [64]. In both cases, topics were specified as a combination of text and examples of the media sought.

New types of high recall search were also examined. In 2002, NTCIR started a long-term examination of patent retrieval [129], considering a range of different search tasks. In this domain, location of all relevant past material was important. Aspects of patent search were more recently taken up by CLEF evaluation exercise [202]. Search supporting e-discovery was examined in the TREC legal track [22], another area of IR where the users (e.g., lawyers) wish to find all relevant items.

4.1.1 Moving Beyond Binary Relevance

From the start of test collection development, there were collections with multiple levels of relevance, Cleverdon's Cranfield I collection [58] had ternary judgments: relevant, partially relevant or not relevant. However, the qrels of most ad hoc style test collections used binary relevance judgments; see Kando et al. [142] and Oard et al. [184], as notable exceptions. This situation changed as researchers started to report that retrieval techniques that worked well for retrieving highly relevant documents were different from the methods that worked well at retrieving all known relevant documents [271].

There was also a realization that degrees of relevance were commonly being used in the internal test collections of web search companies. To illustrate, White and Morris [284, p. 256] mentioned a form of test collection within Microsoft with relevance judgments "*assigned on a six-point scale by trained human judges*". Carterette and Jones [50] described a collection within Yahoo! used for advertising retrieval with five levels of relevance ("*Perfect, Excellent, Good, Fair, and Bad*"); and Huffman and Hochster [124] described work in Google where assessors judged relevance of retrieved documents on a "*graduated scale*". Note, although the dates of these publications are more recent, they are the best examples found of the companies revealing some aspects of their testing work; it is thought very likely that this approach to relevance has been used for sometime in search companies.

Consequently, degrees of relevance become more common in test collections. For example, a ternary scheme was used in the web track of TREC [71]; degrees of relevance were used in TREC's Enterprise track as well as the Blog track.

4.1.2 Diversity

As described at the start of Section 2, the origins of IR test collections lay in methods developed for measuring the effectiveness of library cataloguing systems, where users wrote detailed information requests for librarians who would do the actual searching. However, Verhoeff et al. [264], Fairthorne [83] and Goffman [95], respectively, pointed out that users' definitions of relevance were diverse, queries were commonly ambiguous, and that the relevance of a document to a query could be strongly influenced by the documents already retrieved. However, test collections topics continued to follow the tradition of being detailed unambiguous statements of an information need for which one view of relevance was defined and the relevance of a document was assessed independent of other documents.

An early attempt to create topics with multiple distinct notions of relevance was the interactive track of TREC, which over several years, built a collection composed of 20 topics, each with relevance judgments that addressed multiple aspects [187]. This collection was widely used in diversity research. Following on from this, the novelty track of TREC [103] and the QA track [272] both encouraged the building of systems that retrieved fragments of documents (sentences for the novelty track, so-called *nuggets* for QA) that were both relevant and had not previously been seen. More recently, a collection addressing search of ambiguous person names was created [11]. Both Clarke et al. and Sanderson et al. described re-using existing test collections for diversity research: Clarke re-using a TREC QA collection [55]; Sanderson, an image search collection [221]. Liu et al. [167] described building a web test collection of ambiguous queries, where they examined search engine effectiveness against different levels of ambiguity in the queries.

This relatively small number of test collections addressing diversity is likely to change in the coming years, as web, blog, and image collections were used in the 2009 runs of TREC and CLEF.

4.2 Post Ad hoc Measures

With the new collections, tasks and relevance judgments, came a need for new effectiveness measures. Note, only the prominent measures are

described here, others exist, see Demartini and Mizzaro for a tabulation of a great many [77].

4.2.1 Grades of Relevance

Measures such as Precision can only be used with binary judgments, though using a threshold, one can map n-ary judgments to a binary scheme. Cleverdon used such a mapping in his early test collection work (the two sets of points in the graph in Figure 2.1 reflect results calculated with different thresholds). Researchers were aware that commonly used evaluation measures failed to consider the degrees of relevance and suggested alternatives: Pollock [191] conducted notable early work in this area, aspects of which are described later in this sub-section.

Assuming that one can transform relevance judgments on documents to numerical values, Järvelin and Kekäläinen [132] proposed a suite of measures that evaluated the effectiveness of a retrieval system regardless of the number of levels of relevance. Their simplest measure, *Cumulative Gain* (CG), is the sum of relevance values (*rel*) measured in the top n retrieved documents. Note, Cooper's ESL measure [65] operated in a similar way, though the number of non-relevant documents, instead of relevant, was counted.

$$CG(n) = \sum_{i=1}^{n} rel(i).$$

Examining some example rankings: in the left hand rank in Figure 4.1, CG(5) = 6. Because CG ignores the rank of documents we see that this is also the value of CG in the poorer rank on the right in Figure 4.1. However, *Discounted Cumulative Gain(DCG)*, where the relevance values are discounted progressively as one moves down the document ranking, used a log-based discount function to simulate users valuing highly ranked relevant documents over the lower ranked.

$$DCG(n) = rel(1) + \sum_{i=2}^{n} \frac{rel(i)}{\log_b(i)}$$

Järvelin and Kekäläinen suggested setting b to 2. The ranks in Figure 4.2 show the values of the discount function (in the denominator)

Rank	Rel	Rank	Rel
1	2	1	1
2	1	2	0
3	2	3	2
4	0	4	1
5	1	5	2

Fig. 4.1 Two document rankings.

Rank	Rel	Disc	Rel/Disc	DCG	Rank	Rel	Disc	Rel/Disc	DCG
1	2	1.00	2.0	2.0	1	1	1.00	1.0	1.0
2	1	1.00	1.0	3.0	2	0	1.00	0.0	1.0
3	2	1.58	1.3	4.3	3	2	1.58	1.3	2.3
4	0	2.00	0.0	4.3	4	1	2.00	0.5	2.8
5	1	2.32	0.4	4.7	5	2	2.32	0.9	3.6

Fig. 4.2 Document rankings with discount values.

Rank	Rel	Disc	Rel/Disc	IDCG
1	2	1.00	2.0	2.0
2	2	1.00	2.0	4.0
3	1	1.58	0.6	4.6
4	1	2.00	0.5	5.1
5	0	2.32	0.0	5.1

Fig. 4.3 Perfect ordering of relevant documents.

and the DCG scores for each rank position. We see that DCG(5) of the left hand rank in Figure 4.2 is 4.7 and 3.6 in the right hand poorer rank. In a follow-up paper Järvelin and Kekäläinen [133] added a third measure, *normalized* DCG (nDCG). Here DCG was normalized against an ideal ordering of the relevant documents, IDCG, see Figure 4.3.

$$nDCG(n) = \frac{DCG(n)}{IDCG(n)}.$$

The value of nDCG ranges between 0 and 1. The nDCG(5) of the left and right rankings in Figure 4.2 are $4.7/5.1 = 0.92$ and $3.6/5.1 = 0.71$. Note, Pollock worked on graded relevance, he proposed a measure that computed normalized cumulative gain [191].

Al-Maskari et al. [3] pointed out that in certain circumstances nDCG can produce unexpected results. To illustrate, for both rankings in Figure 4.4, there are only three known relevant documents, though the topic on the left has three highly relevant documents, the other has three partially relevant documents. For both topics the rankings

Rank	Rel	Disc	Rel/Disc	DCG	IDCG	Rank	Rel	Disc	Rel/Disc	DCG	IDCG
1	2	1.00	2.0	6.7	6.7	1	1	1.00	1.0	4.6	4.6
2	2	1.00	2.0	8.7	8.7	2	1	1.00	1.0	5.6	5.6
3	2	1.58	1.3	10.0	10.0	3	1	1.58	0.6	6.3	6.3
4	0	2.00	0.0	10.0	10.0	4	0	2.00	0.0	6.3	6.3
5	0	2.32	0.0	10.0	10.0	5	0	2.32	0.0	6.3	6.3

Fig. 4.4 Rankings from two different topics that result in the same *nDCG*.

are ideal, so the nDCG (DCG ÷ IDCG) in both cases is 1, which is perhaps a counterintuitive result.

Burges et al. [40] described a version of nDCG, for which the DCG component more strongly emphasized the high ranking of the most relevant documents:

$$DCG(n) = \sum_{i=1}^{n} \frac{2^{rel(i)} - 1}{\log(1+i)}.$$

A series of other graded relevance measures were proposed in recent times: Sakai [205] detailed and compared the properties of a number of them including Average Weighted Precision (AWP) and Q-measure. The measures were used by NTCIR organizers as graded relevance was a common feature of the test collections produced by that evaluation exercise. Järvelin and Kekäläinen reviewed many other proposed measures [133].

Moffat and Zobel [179] argued that the log-based discount function in DCG was not the best model of users' behavior when browsing a ranked list of documents. They constructed a model based on the probability p that a user progresses from one document in the ranking to the next. A high p models a persistent searcher; a low p models a fleeting one. The probability was incorporated into a geometric discount function forming the Rank-Biased Precision (RBP) measure

$$RBP = (1-p) \cdot \sum_{i=1}^{d} rel(i) \cdot p^{i-1},$$

where d was the depth of rank one wished to compute the measure over.

Although it is rarely discussed in the literature, when one has grades of relevance in a set of qrels, one could view measuring the effectiveness

of a retrieval system as a comparison of ranked lists: the retrieved ranking compared with the qrels ranked by their relevance. This idea was suggested by Pollack [191]. Joachims implemented the idea using Kendall's τ [153] to measure the correlation between the two ranks so as to obtain a measure of effectiveness [137].

4.2.2 Managing Unjudged Documents

When pooling approaches were introduced into test collection formation, the relevance judgments that accompanied the collections were composed of lists of documents that were judged relevant or not relevant. There was, however a third class of document: the unjudged, documents not in the pool. An early mention of unjudged documents can be found in Hersh et al. [118], where the researchers stated that they chose to ignore such documents when calculating effectiveness measures. A more common approach was to assume that unjudged documents were not relevant. Büttcher et al. [43] pointed out that in many situations this is a sensible approach.

However, Buckley and Voorhees [38] stated that there are other situations where unjudged documents were a potential problem. They were concerned that the size of pools relative to the size of collections was reducing as test collections grew. A related concern was that some test collections were created from sets of documents that were subsequently updated. If no new judging was done after an update, the pool would effectively reduce in size relative to the collection. Therefore, Buckley and Voorhees considered if a new evaluation measure could be devised that better estimated the effectiveness of a system when there were a large number of unjudged documents. They devised BPref, so-called as it "*uses binary relevance judgments to define the preference relation*". It is defined as follows:

$$BPref = \frac{1}{R}\sum_{r}\left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)}\right),$$

where R is the number of documents judged relevant for a particular topic; N is the number of documents judged not relevant; r is a relevant retrieved document; and n is a member of the first R irrelevant retrieved

documents. The measure considers a bounded number of judged non-relevant documents, determining the fraction of these documents that are ranked higher than r. The numerator captures relevance in terms of preference (n ranked higher than r). Although not the first measure to consider preference — see Frei and Schäuble's usefulness measure [88] — BPref is the first commonly used measure to which preference judgments could be easily applied.

In their 2004 paper, Buckley and Voorhees stated that the formulation of BPref was arrived at empirically after a number of experiments. Note the definition shown is from Soboroff's paper [237]. It supersedes the original definition for BPref and its variation BPref-10 and is the default definition of BPref used in recent years by TREC and its trec_eval system (from version 8 onwards). However, such is the popularity of the earlier paper the previous definitions persist in the research community.

Buckley and Voorhees devised a series of simulated experiments comparing the *stability* of BPref with P(10), R-precision, or MAP. They stated that BPref's greater stability was due to its ignoring the increasing numbers of unjudged documents, when the other measures treated these documents as not relevant.

Soon after, Yilmaz and Aslam [289] described a number of alternative effectiveness measures also built to handle un-judged documents including *induced AP* (indAP) and *inferred AP* (infAP). In Bpref, Buckley and Voorhees's aim was to create a measure that mimicked MAP as closely as possible, which was also Yilmaz and Aslam's aim. Unlike BPref, however, their two AP measures were more directly related to the formulation of MAP and in the presence of complete relevance information, resulted in the same score as MAP. Their second measure, infAP is the more widely used of the two and is described here in more detail.

Yilmaz and Aslam split the unjudged documents of a run into two sets: based on whether the documents would or would not have contributed to the test collection's pool had the run been used to build the collection. For the later set, the unjudged documents were assumed to be non-relevant; for the former set, infAP calculated the proportion of judged relevant and non-relevant documents in the document ranking

for that topic and assumed that this proportion was the probability that unjudged documents were relevant. For example, in most of the TREC test collections, pools were formed from the top 100 documents of each submitted run. For such a collection, infAP measured at rank position k would be formulated as follows:

$$infAP(k) = \frac{1}{\mathrm{R}} \sum_{r} \left[\frac{1}{k} + \frac{(k-1)}{k} \left(\frac{|d100|}{k-1} \bullet \frac{|rel| + \varepsilon}{(|rel| + |nonrel| + 2\varepsilon)} \right) \right].$$

Here the definitions of R and r are the same as BPref; $|d100|$ is the number of judged documents found above rank k plus the number of unjudged documents above rank k that would have contributed to the document pool; $|rel|$ is the number of documents above rank k that are judged relevant; and $|nonrel|$ is the number above k that were judged not relevant and ε is a smoothing constant.

In similar stability experiments to those conducted by Buckley and Voorhees, Yilmaz and Aslam showed that all of their new measures, in particular infAP, were substantially more stable than BPref. A number of evaluation campaigns adopted infAP using it in conjunction with pool sampling to streamline their relevance assessment process: TRECVID started in 2006 [188] as did the TREC Terabyte track [42]. For more on pool sampling see Section 6.3.1.

These were not the only example of such effectiveness measures, a tranche of similar measures were proposed and further analyses conducted. Sakai [207] tested a number of alternatives including one that considered graded judgments (RPref). Büttcher et al. [43] described their measure, RankEff, which inferred the relevance of an unjudged document based on its textual similarity to judged documents. A number of variants directly inspired by infAP were described in the literature, statAP, which is used in the Million Query track of TREC is probably the best known [51]. Bompada et al. [27] compared BPref, infAP, and nDCG under a wide range of situations where qrels were incomplete. They found nDCG (which simply ignores unjudged documents) to be the most stable measure. See also Sakai and Kando [209] for another detailed comparison of these types of measures.

Probably because it was the first such measure, BPref started to be used quite widely in the IR research community, however, given more

recent research questioning its stability compared to alternatives, this popularity may be brief. There is not yet a sufficiency of definitive work on which alternative is best.

A different approach to dealing with unjudged documents was suggested by a number of researchers: assessing potential error in a measure. Baillie et al. [20] proposed that the number of unjudged documents retrieved should be considered when making comparisons between runs. They found that if the number of unjudged between such runs was different, there was a danger that comparisons were unreliable. Moffat and Zobel similarly examined error rates when comparing systems on collections with unjudged documents [179].

4.2.3 Relevance Applied to Parts of Documents

Most IR evaluation focuses on retrieval of whole documents. It is to those whole units that judgments of relevance or non-relevance are applied. In early IR research, the "documents" being retrieved tended to be at most abstract-sized texts. However, as full-sized documents started to be retrieved, passage-based retrieval was increasingly studied and effective means of its evaluation was considered. Initial work [117, 212] focused on using passage retrieval to improve document retrieval, which meant that existing document test collections could be used unaltered. In passage retrieval, the aim was to identify accurately the passages of a document that were relevant to a user's request. Consequently, an adapted form of test collection was built with more detailed information in the qrels on the location of relevant passages.

Passage retrieval was part of the tasks in the TREC HARD track and was an integral part of the INEX evaluations of XML retrieval. A broad range of evaluation measures were developed within HARD [281] and INEX [149, 148, 140]. Many of the measures were extensions of existing approaches to evaluation of document retrieval, such as precision, recall, MAP, and DCG. Many of the measures were task specific and no single measure emerged that is used more than others or used beyond the evaluation exercises that created it. Here we illustrate the working of one of these measures taken from Kamps et al. [140]. Here,

document passages (called parts in INEX) p_r were ranked at positions r to collectively form a ranking retrieved in response to a query q. Precision at rank r was defined as:

$$P(r) = \frac{\sum_{i=1}^{r} rsize(p_i)}{\sum_{i=1}^{r} size(p_i)},$$

where $rsize(p_i)$ was the length of any segment of the document part that was judged relevant and $size(p_i)$ was the total number of characters in p_i. Recall at rank r was:

$$R(r) = \frac{\sum_{i=1}^{r} rsize(p_i)}{Trel(q)},$$

where $Trel(q)$ was the total quantity of relevant text in the collection for the query q. Kamps et al. [140] went on to describe an interpolated version of P to allow precision recall graphs to be plotted and an MAP-like measure MAiP to be calculated.

4.2.4 Dependence and Diversity in Rankings

All the evaluation measures described so far assumed that users judged the relevance of a document independent of all other documents. Many measures discounted the importance of documents retrieved lower down the ranking, but the level of discount was always determined by rank and not the quantity of relevant documents already retrieved. Measures that took a dependent view of relevance were developed in the context of diversity and novelty. For diversity, coverage of different aspects of relevance in rankings and individual documents was a priority. For novelty, the goal was to score higher IR systems that prevented repetition of the same relevant content in a ranking.

The initial work in measuring the effectiveness of diverse retrieval systems appears to be from the TREC-6 interactive track [186]. Documents relevant to the collection topics were expected to be relevant to one or more distinct *aspects*,[1] which Over stated were "*roughly one of*

[1] In later TRECs aspects were called *instances* [121].

many possible answers to a question which the topic in effect poses". When assessing runs submitted to the track, *aspectual precision* and *aspectual recall* were calculated. Respectively, Over defined them as "*the fraction of the submitted documents which contain one or more aspects*" and "*the fraction of total aspects ... for the topic that are covered by the submitted documents*". As defined, aspectual precision appears to be the same as precision.

Zhai et al. [293] defined a diversity-specific version of precision and formalized Over's definition of aspectual recall, choosing to instead to call it sub-topic recall (S-recall). Considering a topic with n_A subtopics and a ranking of documents, $d_1, \ldots, d_m$, S-recall calculated the percentage of subtopics retrieved by the documents up to rank position K:

$$S\text{-}recall(K) = \frac{\left|\bigcup_{i=1}^{K} s(d_i)\right|}{n_A}.$$

Here, $s(d_i)$ was the number of subtopics covered in d_i. The measure gave a higher score to runs that covered the largest number of subtopics. Several years later, Chen and Karger [53] proposed the measure k-call(n), which counted if at least one relevant document was retrieved by rank n. By measuring effectiveness in this way, IR system designers looking to optimize for this measure would be motivated to produce systems with diverse rankings.

Both measures ignored the rank of retrieved documents. Clarke et al. [55] proposed an adaptation of nDCG (Section 4.2.1) called α-nDCG, which included this aspect in a diversity measure. The researchers re-defined the function *rel(i)* from nDCG as:

$$rel(i) = \sum_{k=1}^{m} J(d_i, k)(1-\alpha)^{r_{k,i-1}},$$

where m is the number of distinct *nuggets* (the researchers' term for aspects or subtopics), $n_1, \ldots, n_m$, relevant to a particular topic; $J(d_i, k) = 1$ if an assessor judged that document d_i contained nugget n_k;

$$r_{k,i-1} = \sum_{j=1}^{i-1} J(d_j, k)$$

was the number of documents ranked before document d_i that were judged to contain nugget n_k; and the constant α represented the probability that the user of the retrieval system observed prior relevant documents. Note, if α was set to zero and the number of distinct nuggets $m = 1$, the measure reverted to standard DCG. Clarke et al. [56] also created the NRBP diversity metric based on Rank-Biased Precision (RBP). Agrawal et al. [2] pointed out that some of the sub-topics of a query could be more popular or important than others. They found sources of information to estimate a user's probable intended meaning when entering ambiguous topics. To assess their system, they adapted a number of conventional measures (nDCG, MAP, MRR) to handle diversity and to be *intent aware*.

Chapelle et al. [52] described Expected Reciprocal Rank (ERR). While a version of the measure that deals with diversity was described, ERR could also be used simply to promote novelty in search. The measure was defined as follows:

$$ERR = \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r,$$

where n was the number of documents in the ranking and R_i was the probability that the document at rank i was relevant. At each rank position, r, the probability that a user missed each of the relevant documents retrieved higher up the ranking $(1 - R_i)$ was used as a discount on the impact that R_r had on the final score.

The α-nDCG and intent aware precision measures were used in a recent TREC diversity evaluation track [54]; and cluster recall was used in ImageCLEF evaluations [10]. A consensus on a common diversity effectiveness measure is yet to emerge.

4.3 Are All Topics Equal?

Commonly, when an evaluation measure is defined in the literature, its formula is presented as a calculation over a single topic. It is assumed that when summarizing the values computed across a set Q of test collection topics, the arithmetic mean of the values is taken.

Alternatives have been discussed and occasionally tried. Cooper [65] suggested the use of the geometric mean and a weighted average when aggregating scores across queries for his ESL measure, but settled on the arithmetic mean. Later, Voorhees described GMAP, which uses the geometric mean of AP scores [273]. Robertson described the formulation of GMAP [198] and suggested a more computationally tractable version of the formula, which took the arithmetic mean of the log values of AP, Robertson referred to this as AL, this formula produced the same ranking of runs as GMAP though with different values. The two approaches to calculating geometric mean (GMAP, AL) of average precision (AP) values computed over a set of topics Q are as follows:

$$GMAP(Q) = \sqrt[|Q|]{\prod_{k=1}^{|Q|}[AP(Q_k) + \varepsilon]} \quad AL(Q) = \frac{1}{|Q|}\sum_{k=1}^{|Q|}\ln(AP(Q_k) + \varepsilon).$$

Note, if $AP = 0$ for any topic, GMAP becomes zero and AL undefined, therefore, Robertson added a small value ε to avoid such problems. Robertson discussed this measure in some detail pointing out that using geometric mean emphasized improvements in topics that had a low AP score. As Robertson stated...

> "*GMAP treats a change in AP from 0.05 to 0.1 as having the same value as a change from 0.25 to 0.5. MAP would equate the former with a change from 0.25 to 0.3, regarding a change from 0.25 to 0.5 as five times larger.*"

This property of GMAP caused it to be created for use in the Robust Track of TREC [273], where there was a particular focus on so-called poorly performing topics. Beyond the robust track, it is little used. Whether the method is a more effective averaging approach than the arithmetic mean is yet to be determined.

The quality of GMAP to emphasize some changes in topics over others was explored using alternate approaches. Given a large historical set of run scores for the topics of a given test collection, one can compute the score of a new run in relation to the historical scores, determining

on a per topic basis if the new run is better or worse than the past runs. Webber et al. [282] proposed such an approach, employing a methodology used in human testing called *score standardization* also known as *z-score standardization*. For each topic (t) in a collection, the score of a new run s is computed as m_{st}. The mean (μ_t) and standard deviation (σ_t) of the scores for the topic is computed from the historical data and a standardized score (m'_{st}) for s is computed as follows:

$$m'_{st} = \frac{m_{st} - \mu_t}{\sigma_t}.$$

It was long assumed that the topics of a test collection contributed equally to the measuring of the effectiveness of a search engine. Bodoff and Li [26] used Classical Test Theory (CTT) to examine how TREC ad hoc topics ranked the runs submitted to particular years of TREC. They showed how CTT identified potential outlier topics that ranked runs differently from the majority in a collection. The implication from Bodoff and Li's work was that these outliers were potentially, introducing noise into the test collection. Whether such topics were noise or important outliers was not examined by them. The same year, Mizzaro and Robertson [177] reported a study on TREC ad hoc run data looking to find redundant topics in test collections: topics that ranked runs similarly. Mizzaro and Robertson stated that such topics could be eliminated from a test collection and that one "*could do reliable system evaluation on a much smaller set of topics*". However, they only found this small set of topics through an exhaustive search of all possible combination of topics using the run data from TREC.

4.4 Summing Up

In the decade following the development of ad hoc test collections, IR evaluation research explored an increasingly wide range of collection types and search tasks. There was a re-discovery of evaluation ideas and practices described in the past, including use of logs to source topics, gathering, and measurement of graded relevance judgments and increasing consideration of diversity in search results. There was also an exploration of relatively new topics such as management of unjudged documents. As shown in this section, these topics produced an extensive

body of work. However, this was only one strand of evaluation research; work in studying statistical significance was addressed; as well as a more introspective examination of the methods that underpin the creation and use of test collections also became a major part of evaluation research. These two topics are described next.

5

Beyond the Mean: Comparison and Significance

Whichever evaluation measure one uses, the effectiveness of one run will almost always be compared to another, often considering the absolute or relative difference between the runs. However, simply considering the Δ between two values can hide important detail, which is illustrated with an examination of three runs: a, b, c.[1] The MAP for the three runs is 0.281, 0.324, and 0.373, respectively. With similar sized gaps between the runs, one might view the comparisons to be revealing similar differences. However, if one graphs a & b and b & c — plotting the AP scores across each of the 50 topics used in the collection — a more complex picture is revealed; see Figures 5.1 and 5.2. The order of topics in both graphs is sorted by the topic effectiveness scores of the b run.

It can be seen that there is great variation in AP ranging from 0.01 to 0.87. In Figure 5.1, the effectiveness of a follows a similar pattern and has a similar range of scores. Harman and Buckley [106] reported on a detailed study of run comparisons and stated that for most runs, the relatively similar performance on topics shown here is typical. Having

[1] The runs are from TREC-8, INQ604, ok8alx, and CL99XT. The later is a manual run, which probably explains the more erratic difference between it and the others in Figure 5.2.

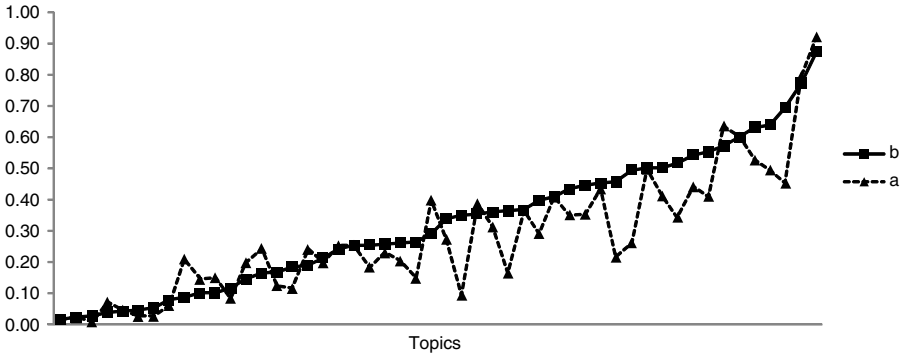


Fig. 5.1 Topic-by-topic comparison of two TREC-8 runs based on average precision scores.

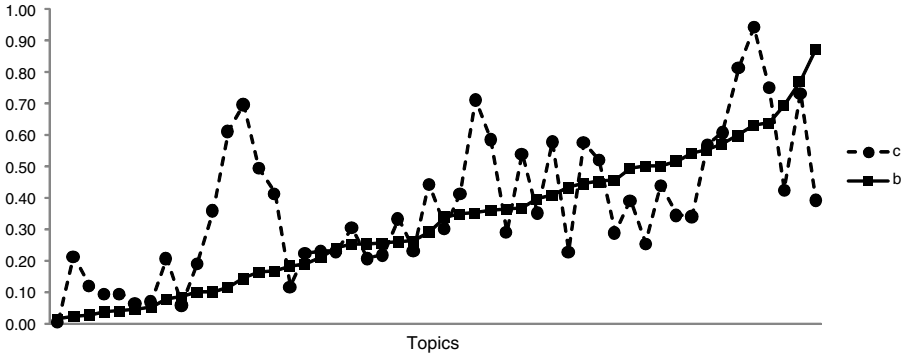


Fig. 5.2 Topic-by-topic comparison of two TREC-8 runs based on average precision scores – taken from Harman's work [104].

the worse run, in this case a, being the same or a little better for some topics (16 of the 50 topics here) is also common. The absolute difference between a and b is 4.3% and relative difference is 15.4%. Both the difference in scores and an examination of the graph in Figure 5.1, would lead most to agree that in this case system b is better.

Harman [104] highlighted the comparison shown in Figure 5.2 where it is arguably harder to determine the better run, yet the absolute and relative differences between b and c are 4.9% and 15.1% respectively, similar to the two runs above. While Harman and Buckley's work [106] showed that most run comparisons were like the case in Figure 5.1, situations such as those found in Figure 5.2 are not exceptional. Therefore,

it is necessary to examine more than the single value effectiveness measure calculated for a particular run.

5.1 Significance Tests

A common approach to more fully understanding a difference Δ measured between two runs is to use one or more significance tests. The tests estimate the probability p of observing a difference at least as large as Δ given that a so-called *null hypothesis* (H_0) is true. In the context of IR experiments, H_0 states that the systems producing the two runs under examination have effectively the same retrieval characteristics and that any difference between the runs occurred by random chance. The convention when using such tests is that if it is found that p is below a certain threshold — typically either 0.05 or 0.01 — it is concluded that H_0 is unlikely and consequently should be rejected. Although it is not universally agreed upon, the common interpretation of rejecting H_0 is to conclude that an alternate hypothesis, H_1 is true. This hypothesis states that the two IR systems have different retrieval characteristics, leading the experimenter to conclude that a significant difference was observed. The exact nature of H_1 depends on whether a *one-* or a *two-tailed* test is chosen. This topic is discussed below in Section 5.1.2.

The tests are not infallible and can make errors, which have been classified into Type I and Type II errors. Type I errors are false positives: leading the experimenter to incorrectly reject H_0 and conclude that H_1 is true; Type II errors are false negatives: leading the experimenter to incorrectly conclude that they cannot reject the null hypothesis. In IR parlance, Type I measures the precision of the test, Type II measures its recall. Different significance tests tend to produce a different balance between these two errors. For example, the sign test is known for its high number of Type II errors whereas the t-test is known for producing Type I.

The creators of the tests make assumptions on the underlying data being examined and it is important for experimenters to be aware of a test's assumptions before applying it. Tests that have fewer assumptions tend to generate more Type II errors and are said to have less *power*. So-called *non-parametric tests*, fit this profile, the best known in

IR research are the *Wilcoxon signed-ranks test* and the *Sign test*. More powerful tests generate fewer Type II errors but make more assumptions about the data being tested. If such tests, known as *parametric tests*, are applied to data in violation of such assumptions Type I errors can result. The best-known parametric test used in IR is the *t-test*.

Assuming each run is conducted on the same topics in a test collection, the significance test is usually operated as a paired test. If, less commonly, the runs being compared use different sets of topics an independent or two-sample version of the tests can be used. Although there is a wide range of tests available to the IR researcher, the three mentioned so far are the most often used. Lesk [162] discussed use of the sign and the t-test in IR experiments and Ide used the Wilcoxon and t-test to examine the significance of retrieval results [127].

At the same time, Saracevic urged caution on the use of a wide range of statistical tests. He found that no parametric statistical test could be used with confidence on data emanating from effectiveness evaluations because such "*data does not satisfy the rigid assumptions under which such tests are run*" [223, p. 13]. In addition, he pointed out that conditions set for use with non-parametric tests were also likely violated by the data output from test collection evaluations. Saracevic did not suggest that these problems should be viewed as being the end of the matter, instead he called for studies on the applicability of statistical tests in IR evaluation. Later Van Rijsbergen detailed the problems with using these test [262, Section 7]. For the t-test, results needed to be drawn from a "*normally distributed population*", which Van Rijsbergen stated did not occur with the output of test collection based retrieval experiments. For Wilcoxon and Sign, he pointed out these tests could only be used if data was drawn from continuous distributions; yet retrieval experiments produce discrete distributions. Despite the violations, Van Rijsbergen suggested "*conservative*" use of the Sign test.

In later work, Hull [125] suggested that such prudence was most likely excessive. Hull argued that with "*sufficient data, discrete measures are often well-approximated by continuous distributions*". Further, he stated that "*the t-test assumes that the error follows the normal distribution, but it often performs well even when this assumption is violated*". Hull went on to discuss the properties of a range of

significance tests, including the three examined by Van Rijsbergen as well as variations of the ANOVA and Friedman tests. Robertson had earlier conducted a theoretical analysis of a set of tests [196], estimating the topic set sizes needed in order to obtain statistical significance in retrieval experiments. He examined the Mann–Whitney U-test, the Chi-squared and the t-test. However, neither Hull nor Robertson empirically tested their comparisons.

Empirical examinations of the tests appeared to have been first conducted by Keen [152] who described a small-scale comparison of the Sign and Wilcoxon tests, the results of which showed the tests gave "*a similar picture*". Keen stated that Wilcoxon tended to indicate significance more readily. Zobel described a study of which of three significance tests was the best suited for IR experiments [295]. Splitting the topics of a test collection in half: one-half used as a mini test collection; the other as a simulation of an operational setting, he tested the paired t-test, the paired Wilcoxon's signed rank test and ANOVA (his method is described in detail in Section 6.4). Zobel concluded that use of all three tests resulted in accurate prediction of which was the best run, though he expressed a preference for the Wilcoxon test "*given its reliability and greater power*".

Later, Sanderson and Zobel [222] using Zobel's [295] methodology examined the t-test, Wilcoxon and Sign tests. Their conclusions were that use of the t-test and the Wilcoxon test allowed for more accurate prediction over the sign test of which run was better. However, the differences between the t-test and Wilcoxon test were small. They also pointed out that even if one observed a significant difference between two runs based on a small number of topics (≤ 25), one should not be confident that the same observed ordering of runs will be seen in the operational setting. Using a variant methodology, Cormack and Lynam [68] reported significant differences resulting from the t-test failing to be observed in an operational setting particularly for topics with a small number of relevant documents (≤ 5).

Savoy [226] proposed use of the *bootstrap* test,[2] as assumptions of normality, or continuous distributions are not required with this test. Despite Savoy's promotion, the test was little used by the IR

[2] Savoy cites Efron and Tibshirani [81, 82]; Léger et al. [161] as the originators of the test.

community until more recent times, when Cormack and Lynam [68], Sakai [206] and Jensen [134] applied the test. Sakai [208] provided a detailed explanation of how two forms of the test can be used.

A comparison of a number of such newer significance tests was conducted by Smucker et al. [235] who compared the Wilcoxon, sign, and t-tests with the bootstrap and the *randomization* or *permutation* test. Like the bootstrap, almost no properties of a data set must hold before the randomization test can be applied. Smucker et al. compared the values of p obtained across every possible pairing of runs submitted to 6 years of TREC's ad hoc track. They found the t-test, bootstrap and randomization produced similar p values across the pairs, with the Wilcoxon and sign tests producing quite different p values. Smucker et al. stated that the Wilcoxon and sign tests were simpler versions of the randomization test and so argued that in comparison with the randomization test, the different sets of p-values were indicative of errors in the two non-parametric tests, leading the researchers to argue that use of the Wilcoxon and sign should cease. Although their experiments did not show any difference between the remaining three tests, they argued from references to past work that the randomization test was likely to be the best to use in test collection experiments.

5.1.1 Not Using Significance

Spärck Jones [241] suggested a simple approach to determining the significance of comparisons stating that "*in the absence of significance tests, performance differences of less than 5% must be disregarded*". Voorhees and Buckley [276] later clarified that Spärck Jones was referring to was an absolute percentage difference. Given the example of the two TREC runs graphed above, with an absolute difference of 4.9%, one might be tempted to agree with this view. She went on to state that she would "*broadly characterize performance differences, assumed significant, as noticeable if the difference is of the order of 5–10%, and as material if it is more than 10%*". Use of simple tests like this are rarely reported in the research literature.

Voorhees and Buckley [276] when comparing effectiveness measures (see Section 6.4) chose to treat Spärck Jones's threshold as a form of simple significance test: requiring a 5% difference between runs. The

question that Voorhees and Buckley addressed was how many topics were needed in a test collection in order for a 5% difference (measured in the collection) to accurately predict which of a pair of runs was the better in the operational setting. They found that around 50 topics were needed before one could be confident that a 5% absolute difference measured between two runs on a test collection would predict which was the better run in an operational setting. Their result was in stark contrast to Zobel's 1998 work who found that using 25 topics in conjunction with the common significance tests was more than sufficient [295]. Given that the methodologies and data sets used between the two works were similar, it would suggest that, just measuring the magnitude of difference in effectiveness scores has limited utility.

Another work of note is that of Frei and Schäuble [88], who proposed a new evaluation measure that as part of its calculation, computed an error probability to indicate the stability of the value measured. The measure, usefulness, was never widely adopted, although, another feature of the work, that it used relative relevance judgments, proved to be influential on others later.

5.1.2 One or Two-Tail Tests?

So far H_0 has been described, but H_1 has not. There are two types of hypothesis that can be chosen for H_1, which correspond to different types of test: a one- and a two-*tailed* test (also known as a one- or two-*sided* test). In a two-tailed test, H_1 states that the two systems under examination are not equal, e.g., from the runs in Figure 5.1, H_1 would state that system a does not have the same retrieval characteristics as system b. Comparing a and b, a two-tailed t-test of H_0 returns $p = 0.002$; the Wilcoxon signed-rank test returns $p = 0.004$; and the sign test $p = 0.015$. Assuming a 5% threshold, regardless of which test was used, the experimenter would reject the null hypothesis and consider the difference between a and b to be significant.

Since IR experiments are often concerned with determining if a new type of IR system is better than an existing baseline, experimenters sometimes use a form of significance test that focuses only on the question of difference in one direction between two runs: this

is the one-tailed test. Here the experimenter predicts before conducting the experiment that one of the systems will be better than the other and sets H_1 to reflect that prediction. Taking this time the comparison from Figure 5.2, if system b is a baseline and system c is a new system under test, the experimenter sets H_1 to predict that $c > b$. A one-tailed t-test returns $p = 0.036$; the Wilcoxon $p = 0.026$; and the sign test, $p = 0.102$.[3] Despite the lack of significance in the sign test, most experimenters would consider the improvement of c over b as significant. The one-tailed test is recommended for use in IR experiments by Van Rijsbergen [262, Section 7] and more recently by Croft et al. [74], however, it is worth noting that in some areas of experimental science the one-tailed test is viewed as almost always inappropriate [8, p. 171].

The one-tailed version of a significance test has a p value that is half that of the two-tailed version, which makes it a tempting choice for experimenters as its use doubles the chance of finding significance. Note for example that all of the two-tailed tests comparing b and c would have failed to reject H_0. However, if using the one-tailed test, it is important to understand what its use entails. If from Figure 5.1 an experimenter had incorrectly predicted that system $a >$ baseline b and chose to use a one-tailed test; and upon discovering that a $< b$, the experimenter would *have* to conclude that they had failed to reject H_0. In other words, the experimenter would be obliged to report that a and b had the same retrieval characteristics, no matter how much worse a was compared to b; to many a strange conclusion to draw. The experimenter could of course conduct the one-tailed test in the opposite direction, but this second test could *only* be conducted on a new data set.

Recalling the example in Figure 5.2, an abuse of significance tests would arise if an experimenter decided to use a two-tailed test, when comparing c and b, found no significance and so switched to a one-tailed test in the favorable direction in order to search out significance.

[3] It is also worth noting in the c and b comparison how the Wilcoxon and t-tests produced p values below the 0.05 threshold but the sign test did not. The former tests were more influenced by the substantial improvements of c over b in some topics. The sign test ignored the size of a difference; considering only the sign of the difference.

The choice of a one- or two-tailed test needs to be made before analyzing the data of an experiment and not after. If you are not sure of the direction of difference you wish to test for when comparing two systems, a two-tailed test is the appropriate choice. If you are certain that you only wish to test for a difference in one pre-selected direction, the one-tailed test can be used. It is important that the experimenter always states which "tailed version" they used when describing their work.

5.1.3 Consider the Data in More Detail

It is also always worth remembering that although the use of significance tests can help the IR researcher better understand the difference Δ between two runs, the tests are not oracles, they are merely a generic statistical tool constructed for the purpose of estimating the probability p of observing at least Δ if the null hypothesis, H_0, is true. The value of p is calculated on the sample of topics, documents, and relevance judgments in the test collection. If that test set is a representative sample of the broader population of queries, documents and judgments made by users in an operational setting, then the conclusions on whether H_0 can be rejected or not should apply to the population. If, however, as is often the case, the sample is not representative, then conclusions drawn may be unreliable. Sanderson and Zobel [222] showed a number of examples where a range of significance tests produced p values ≤ 0.05 for a pair of runs on one sample test collection, but for the pair using a different sample collection produced p values > 0.05. Voorhees, using larger topic sets [275] went further, occasionally finding examples where one system was better than another, but on a different topic set, the system ordering was swapped and in both cases the differences were significant.

Even if the experimenter compareds two runs using the collection and found $p \leq 0.05$, there remains the question, is the result *practically significant*? Comparing the runs in Figure 5.2, if b was a baseline system already installed at an organization, even though c is significantly better (according to a one-tailed test), the manager of the existing baseline system might argue it is questionable if users would welcome the new system c given that it is notably worse than b on 10 of the 50 topics

(20%). In a different setting, a manager might conclude that c is worth installing because it appears to improve substantially on topics that b performed very poorly on, but only reduces somewhat b's top performing topics and those reductions would be acceptable to his/her users. Such issues, which could be critical in deciding the value of one system over another, are not addressed by significance tests and can only be answered by a more detailed understanding of the uses and users of an IR system.

It is also important to consider the magnitude of difference between two systems: i.e., if a significant difference is substantial enough. In an operational setting, as Cooper pointed out [67], a better system might require more compute resource than the baseline and the benefits of the new might not outweigh the disadvantages of the additional resource needed. Alternatively, experimentation comparing a new system with a baseline might fail to reject H_0. However, if the new system uses fewer resources than the baseline, the new system could still be the better choice. The question of what constitutes a sufficiently large improvement in retrieval effectiveness continues to be examined in some detail by the IR community, details of which can be found in Section 6.5.

A significance test provides a binary decision on whether there is something of note in data or not. On finding significance, researchers may not feel it is necessary to examine their data further. Perhaps of more concern is researchers who fail to find significance, may not look further at their data, which may prevent them from learning what had gone wrong and/or how to fix any problem. Webber et al. explored this situation, pointing out that one possible explanation for failing to reject H_0 is that the experimental setup did not have sufficient *statistical power* for such a difference to be reliably observed: i.e., a Type II error occurred [283]. The researchers described the means of measuring such power in test collections and detailed a method for incrementally adding topics to a test collection until the required power was achieved in order to avoid such errors.

While significance tests are without doubt a popular statistical data analysis method, it is worth remembering that many statisticians feel the tests are overused and that their use discourages researchers from examining their data in more detail. As pointed out by Gigerenzer,

a wealth of other statistical analysis methods exists to allow different forms of analysis to be conducted [94]. One popular alternative is the *confidence interval* (CI) which can be used to compute an interval around a value, commonly displayed in graphs using an error bar. If when comparing two values, the error bars don't overlap, a researcher can state that the difference between the values is of note. In some scientific fields, confidence intervals have replaced significance tests to become the default method for analyzing experimental data. It would appear they were chosen because their use encouraged more analysis of the properties of the data, than significance testing does. Confidence intervals are sometimes used in IR literature; in describing statMAP, Carterette et al. defined how to compute CIs over that measure [48]. Cormack and Lynam [68] described how to calculate an interval on MAP.

6

Examining the Test Collection Methodologies and Measures

Cleverdon's report of his design and use of the Cranfield test collection [58] — with its set of documents, topics, and qrels — concluded a decade of preliminary work in testing IR systems and established a methodology for evaluation that, now approaching its 6^{th} decade, continues to be widely used. The changes in computer technology in that time have been profound, causing IR systems to transform from slow searchers of limited collections to engines capable of searching billions of documents across different media, genres, and languages. With such enormous change, it is striking that the test collection methodology has altered little over that time.

In the decade of research conducted after the development of ad hoc test collections, there was a wide ranging examination of all aspects of test collection methodology, helped greatly by the run data sets produced by TREC and NTCIR. Exploitation of these sets prompted a re-examination of the impact of assessor consistency on measurements made on test collections; an exploration of pooling including consideration of effective use of relevance assessors' time; determining which were the best topics to use in a test collection; establishing which is the best evaluation measure; and perhaps most importantly, determining

if test collections actually predict how users will use an IR system. The work is now described.

6.1 Re-checking Assessor Consistency

As highlighted in Section 2.5, an early concern of some researchers was the potential of inconsistent relevance assessments resulting in poor quality qrel sets, which could mean that measurements made on test collections could be inaccurate. Early on, both Lesk and Salton as well as Cleverdon tested the potential for such error and found that variations in assessments although high did not affect the relative ranking of runs. Voorhees conducted a larger scale study using TRECs 4 and 6 run data [268], for which multiple relevance assessments, qrels, were available.

Voorhees established a method of correlating ranks of runs that became widely used by many other IR researchers. The method involved ranking the runs submitted to the two TRECs using different qrel sets. Voorhees measured the correlation between a rank of runs using one qrel set and the same runs ranked using a different qrel set. A high degree of correlation meant the qrel sets ranked IR systems similarly; and low correlation implied that each qrel set was ranking runs differently.

In order to determine the similarity between ranks, Voorhees used Kendall's Tau (τ) [153]. For any two rankings of the same n set of items, τ is a linear function of the number of pairs of items which are in different orders in the two rankings. This function is constrained such that $\tau = 1$ if the two rankings are in identical order, and $\tau = -1$ if the order is reversed. There are a number of τ variations, we illustrate the one originally defined by Kendall. It is as follows:

$$\tau = \frac{2(n_C - n_D)}{n(n-1)}.$$

Here n_C is the number of pairs in the two rankings that are in the same order (i.e., concordant) and n_D is the number of pairs that are in different order (i.e., discordant). Note that every possible pair of items in the rankings (i.e., there are $n(n-1)$ such pairings) is compared when calculating n_C and n_D. If one was using τ to measure the correlation

between two rankings of ten runs from TREC and the two rankings were identical, then $n_C = 45$, $n_D = 0$, and $\tau = 1.0$. If there was a swap anywhere in one ranking of two adjacent items, then $n_C = 44$, $n_D = 1$, and $\tau = 0.96$. If the swap was between the first item and the last item of one ranking, then $n_C = 28$, $n_D = 17$, and $\tau = 0.24$.

When considering τ calculated over a set of rankings, a common question that is asked is at what level of correlation can one view two rankings as effectively equivalent? Voorhees suggested $\tau \geq 0.9$ as such a threshold [270], which was adopted by a number of subsequent researchers.[1] Using Kendall's τ Voorhees found the rankings of runs, though not identical, were very similar, leading Voorhees to conclude that variations in assessment did not impact noticeably on retrieval effectiveness. The second part of Lesk and Salton's work on examining the consistency of judgments against the rank of the documents being judged was also repeated in later years: both Sanderson [216] and later Voorhees [270] using different TREC data sets, showed that inter-assessor agreement was higher for top-ranked documents.

In the experiments conducted up to this point, the assessors used to generate different qrels were all assumed to be capable of judging the relevance of the documents. In later work based on TREC Enterprise track data, Bailey et al. [19] drew from different sets of assessors based on their knowledge about the test collection topics. The assessors were classed as gold, silver, and bronze judges. Gold and silver were subject experts with the gold judges having a more intimate knowledge of the data set being searched. The bronze judgments were made by the participants in the TREC track: presumably motivated, but non-subject experts. Like the previous results, Bailey et al. showed that the

[1] Sanderson and Soboroff [220] pointed out that the items in a ranking are sorted by a score and the range of the scores of the items in a list, impacts on the value of τ. They showed that if the range is large, there is a greater likelihood of finding high τ correlation scores. This quality is common to all rank correlation measures and makes use of absolute thresholds difficult. Another criticism of τ is that it measures correlation equally across a ranking and for many IR tasks, correlation in the top part of a ranking (i.e., the runs of the best performing systems) is generally more important than the bottom. Yilmaz et al. [290] produced a new correlation coefficient τ_{ap} that addresses this failing. They also cite a number of other proposed τ variants. Carterette has also described an alternative to τ [46]. See also Melucci [175] on concerns about use of τ in IR experiments.

qrels from the gold and silver judges produced similar rankings of runs. However, the rankings from gold and bronze judges were different.

These works show that while test collections are more resilient to assessor variation than was originally feared, there are limits to this resilience and the appropriateness of the assessors used to make judgments needs to be carefully considered when forming qrels.

It is often thought that differences in assessment are an indication of some sort of human error. However, Chen and Karger [53] showed that one can view the differences as simply two distinct, but valid, interpretations of what constitutes a relevant document. They used the TREC-4 and 6 multiple assessments to test a retrieval system that returned diverse search results. Viewing the two sets of assessments as representing two legitimate interpretations of relevance for the topics in TRECs 4 and 6, Chen and Karger showed that supporting diversity ensured that more documents were retrieved that satisfied at least one of the assessors. For more on diversity evaluation, see Section 4.1.2.

6.2 Does Pooling Build an Unbiased Sample?

The aim of pooling is to locate an unbiased sample of the relevant documents in a large test collection, as made clear by Spärck Jones [243]. She was confident in the validity of pooling in part due to earlier work by Harman [102] and later Zobel [295] who tried to estimate the quantity and impact of relevant documents missing from TREC pools. In the early years of the TREC ad hoc collections, pools were formed from the union of the top 100 documents retrieved by each submitted run for each topic. Harman examined a pool formed by the documents in ranks 101–200 for a sample of the runs and topics in TREC-2 and all the runs and topics in TREC-3. She reported that a further 11% of relevant documents were discovered in the TREC-2 pool and a further 21% in TREC-3. Harman stated that "*These levels of completeness* [in the pools] *are quite acceptable for this type of evaluation*". Zobel examined the relationship of the number of relevant documents found (n) to the depth of rank used to form a pool (p) and found the relationship to match well to the following power law distribution:

$$n = Cp^s - 1,$$

where C and s were constants. So strong was the fit that Zobel felt confident to extrapolate the curve beyond the depth 100 pools of TREC. Extrapolating what n would be when $p = 500$, he stated that the number of extra relevant documents would be double that found when $p = 100$. The prospect of so many unfound relevant documents caused Zobel to consider if it was possible for runs to retrieve the majority of the unknown relevant and consequently receive an unfairly low effectiveness score. He, therefore, explored how the contributing runs to TREC would have been ranked if they had not contributed to the formation of the pool. This was achieved by, in-turn, removing from the pool the relevant documents unique to a particular run, re-forming the reduced qrels and then comparing a ranking of all runs with the original run rank. Zobel found relatively small changes in the way that left out runs were ranked, the pools appeared to be an unbiased sample. Zobel stated the result boded well for the reusability of test collections.

More recently, Büttcher et al. [43] explored two adaptations of Zobel's "*leave one run out*" experiment. They pointed out that it was common for a research group to submit multiple runs to TREC, leaving a single run out, as Zobel had done, was perhaps not the best of simulations as other runs from the same group would have been left in. Therefore, borrowing a technique described by Voorhees and Harman [277] the researchers used a "*leave one group out*" approach. Testing their work on the TREC 2006 Terabyte track runs and qrels, the researchers found that removing a whole group made a relatively small difference to the way runs from the held out group were ranked. In a second test, the researchers held out relevant documents uniquely found in manual runs from the pools. The manual runs are generally the richest source of relevant documents. By leaving them out Büttcher et al. were attempting to simulate a situation where after a test collection was formed, a new substantially better run was tested on the collection. Examining how well these runs were ranked by the reduced pools, Büttcher et al. reported that the runs were ranked somewhat differently compared to when the full TREC pools were used.

Such a result was undoubtedly concerning. However, the extent that it represented a problem for experiments is yet to be fully determined. There would not appear to be much in the way of evidence that users of

test collections have found new retrieval systems poorly scored, though admittedly few researchers actually analyze for such potential problems. There appears in the literature to be just one example of a run that if it had not contributed to the pool of a test collection, it would have been poorly scored on that collection. So unusual was this particular run that it was studied in some depth by Buckley et al. [36]. They found that within the pool of runs used to form the test collection was a particular bias: most relevant documents contained at least one word from the title of the test collection topics. However, the errant run contained many documents that were both relevant and did not contain words in the topic title.

Buckley et al. studied this unusual run and the properties of the collection to try to better understand the causes of this anomaly. It would appear that the size of the collection was an important factor. It is normal for many of the relevant documents in a test collection to contain words from the title of a topic and it is also to be expected that such documents would be highly ranked. The number of such documents is finite. In the smaller test collections it would appear that the size of the pools assessed was larger than the number of relevant documents containing terms from the topic title. However, in the collection under study, the pool size was not large enough to encompass all the relevant documents containing a topic title word and to also find enough of the relevant documents without.

Whether this one example was an outlier or an indicator of a broader problem with current approaches to pooling in large test collections is yet to be determined.

Examining a different aspect of potential bias in pooling, Azzopardi and Vinay [17] studied if within large collections there are documents that are almost never retrieved by any search engine. Loading large collections into a conventional ranked retrieval system, they ran hundreds of thousands of queries on the collection. The queries were either single word terms that occurred >5 times in the collection, and bigrams that occurred >20 times. The researchers' aim was to understand if through all those queries there were documents in the collections that persistently failed to be highly ranked. Their conclusions were that a notable number of such documents existed in these collections. Such a

conclusion could be of concern since pools are built through querying. However, Azzopardi et al. did not examine the relevance of the poorly retrieved documents. It is unclear at the moment if this probable bias is important with respect to locating a representative sample of relevant documents.

It would appear that despite concerns of some in the IR community that pooling risks the creation of test collections a biased sample of qrels, studies have largely shown such concerns are unfounded. However, the effort required to build pools is substantial and as described in Section 6.3, attempts are being made to produce smaller pools, which might introduce new forms of bias. This is a topic, therefore, that is likely to be returned to in the future.

6.3 Building Pools Efficiently

Reflecting on the first eight years of TREC, Voorhees and Harman [278] detailed that the average number of documents assessed per topic in the ad hoc tracks of TREC was 1,464 (averaged from data in Table 4 of that paper). With 50 topics per year, approximately 73,000 judgments were made each year. At a rate of two judgments per minute — Voorhees and Harman [279, Section 2.1.3] estimate — 8 hours of work per day, judging the TREC ad hoc pool each year took just over 75 person days. This was the limit of human resource TREC organizers were able to supply; similar limits applied to other large evaluation exercises such as CLEF or NTCIR.

Assessors in evaluation exercises tend to be used in a relatively straightforward and similar manner. Here we highlight the way that assessors were used in TREC ad hoc. When groups submitted a run to TREC, the top 100 documents for each topic in the run were extracted and merged into a pool for each topic and sorted by document id. The assessor for a topic was generally the person who created that topic. They examined every document from every system in the pool for that topic. This straightforward approach to examining a pool was justified by Voorhees and Harman [278] and later by Soboroff and Robertson [239] by stating that it was important for the pools to contain a set of documents that were not biased in any way toward one particular

retrieval strategy or a particular type of document. This was judged vital so that the qrels could be used not only to fairly determine the relative effectiveness of runs submitted to TREC, but also that later users of the test collections could be confident that the effectiveness of a new retrieval strategy will be accurately and fairly measured.

A number of researchers examined other ways of selecting or sampling parts of a pool to judge so as to use fewer human assessments. The approaches are grouped here into examinations of the way that pools are scanned; research assessing if pool depth or topic breadth is more important; approaches to solve the assessment resource problem by distributing assessment; an examination of an approach that opened the possibility of avoiding use of assessors at all; and finally exploitation of existing data to simulate assessments.

6.3.1 Scanning the Pools in Different Ways

Zobel [295] pointed out that some topics in a test collection will have more relevant documents in them than others and suggested that as topics were being assessed for pools, the number of relevant found so far could be noted and for those topics richer in relevant documents, more assessor effort could be focused on examining a larger pool for them.

In the same year that Zobel's made his suggestion, Cormack et al. [69] proposed a number of alternate strategies. In a similar vein to Zobel's ideas of focusing assessor effort on the richest sources of relevant documents, they pointed out that the runs from certain retrieval systems contributing to the pool contained many more relevant documents than others. They proposed focusing assessor effort onto those runs richer in relevant documents. This approach they called local move to front (MTF) pooling. The researchers also tested an approach that included Zobel's ideas: prioritizing assessment of both the most fruitful runs and the most fruitful topics, which they called global MTF pooling. The authors tested the approaches and showed that they could assess 10% of the full TREC pool and produce a qrel set that ranked runs in an almost identical manner to the ranking achieved using the TREC baseline pool. Global MTF appeared to be more effective than local MTF. See also Moffat et al. [178] for related work.

Cormack et al.'s paper contained one other approach to building qrels, which they called *Interactive Search and Judge* (ISJ). Here they proposed an alternative role for the relevance assessor, instead of judging a long list of documents, the assessor would search the test collection, issuing multiple queries, noting down relevant documents found and searching until they could find no more relevant for a particular topic. Cormack et al. reported that a group of ISJ assessors was given the task of locating relevant documents for one of the years of TREC. In just over 13 person days (compared to TREC's 75 person days) a qrel set was formed that was shown (through experimentation) to be of comparable quality to that produced by TREC.

The work was a re-examination of the approach to pooling used by Katzer et al. [146] and earlier still by Lancaster [158]. TREC implicitly uses ISJ through its encouragement for manual runs to be submitted to its tracks [217]. It has also been used explicitly by a number of evaluation campaigns such as CLEF [63, 93] and NTCIR [156]. Recognizing that TREC assessors preferred to assess rather than search, Soboroff and Robertson [239] described an alternative approach to ISJ where relevant documents located by assessors were used as positive examples for a relevance feedback system to locate more items for assessment. Soboroff reported that the approach had worked well in reducing assessor time and effort. Oard et al. [184], detailed using this technique, which called they *search-guided assessment*, to build a test collection.

Carterette et al. [47] pointed out that certain documents in the pool were better at distinguishing runs from each other and these should be targeted for assessment. For example, if one compared two runs using P(10) and all that one wished to determine was which run was better, only the top ten documents needed to be examined and any documents in common could be left unjudged. Through such targeting, the researchers took eight runs from six different searching systems retrieving across 60 topics and in 6 hours of assessor time were able to produce a ranking of those runs that with 90% confidence was the same as the ranking produced by a baseline TREC top 100 pooling approach.

Another approach to reducing assessment is sampling of pools, an analysis of which was first described by Spärck Jones and Bates [244,

pp. 20–21]. Aslam et al. [13] described experiments with pool sampling showing that one could sample as little as 4% of a TREC ad hoc pool and still produce accurate results. Lewis [165] used stratified sampling to the pools of the TREC filtering track. More recently this approach was exploited in the TREC million query track [51] and the legal track [22]. When sampling pools, the number of unjudged documents is likely to increase, consequently, the measures described in Section 4.2.2, tend to be used, particularly, *infAP* and *statAP*.

When compared to the original approach used to form qrels for the TREC ad hoc collections, it would appear that at least some of the methods described here could be used with confidence. For some of the newer approaches involving more extreme forms of sampling, although experimental results have shown the methods to be reliable when tested on historical data, there is still the question of how re-useable such collections will be in experiments run in the future.

6.3.2 Narrow and Deep, or Wide and Shallow?

One question considered by test collection creators is should assessors be focused on judging deeply the pools of a small number of topics, or the shallow pools of a larger number of topics? The convention was to limit the number of topics and examine their pools in some detail. For many years in TREC, the top 100 retrieved documents of submitted runs were assessed, though in recent years this was reduced to the top 50. Sanderson and Zobel [222] speculated on the number of topics that could be assessed if a smaller part of a run was drawn into a pool. They estimated that if only the top 10 of each run was assessed, a test collection with between 357 and 454 topics could be created using the same amount of assessor effort as with 50 topics examined to depth 100. They also pointed out that the top part of runs generally have a greater density of relevant documents, consequently such a strategy would in all likelihood find between 1.7 and 3.6 times more relevant documents than a conventional pooling approach.

Bodoff and Li [26] pointed out that sources of score variation in test collection based evaluation can be attributed to different IR systems, different topics, and to interactions between systems and topics. The

researchers analyzed TREC run data using Generalizability Theory to identify where the main source of variation was, concluding that topics were the highest source. This led the researchers to conclude that building test collections with more topics was a priority. Webber et al. [283] applied their analysis of statistical power in test collection based experiments to also study this question. They found that greater statistical power would result from a wide and shallow approach to pooling.

At around this time, publications from the commercial search engine community were produced showing that internal test collections had substantially larger numbers of topics than existed in publically available ones: White and Morris [284] mentioned a collection at Microsoft with 10,680 "*query statements*"; Carterette and Jones [49] described a collection in Yahoo! with 2,021 queries; at the same company Chapelle et al. [52] mentioned several internal collections, one with over 16,000 queries, judged to depth five.

In reaction to this, the academic community looked to build its own collections with many more topics. Carterette et al. [51] described work on the so-called Million Query Track, which, using both Carterette et al.'s just-in-time approach and the pool sampling methods associated with *statAP*, created a test collection with 1,755 topics. With their larger data set Carterette et al. were able to empirically test the proposal by Sanderson and Zobel that "wide and shallow" was better that "deep and narrow". In their data set Carterette et al. found that 250 topics with 20 judgments per topic were the most cost-effective in terms of minimizing assessor effort and maximizing accuracy in ranking runs. Voorhees expressed concern that wide and shallow test collections might not be as reusable as ones built using the deep and narrow approach [274]. However, Carterette [45] provided evidence that such collections were more reusable than was perhaps previously thought.

6.3.3 Distributing Assessment

The relevance assessors are generally paid by the organizers of evaluation campaigns. There have been attempts to reduce or remove such

costs. Both INEX and the enterprise track of TREC explored making relevance assessment a necessary part of groups being able to participate in the campaign. There is relatively little research on the accuracy of such "coerced" judgments. However, Bailey et al.'s work [19] on gold, silver, and bronze judgments suggested that such an approach to gathering judgments was not without its risks.

Another potentially large source of human assessors can be found through crowd sourcing. Alonso et al. [7] described their use of the Amazon Mechanical Turk system to obtain relevance judgments. Mechanical Turk is a market place where workers are paid small amounts of money (typically ranging from 1 cent to $2) to conduct short run tasks, called *HITS*. The tasks on offer in the market place include writing short reviews, adding metadata to images, and judging documents for relevance. Because the workers on systems like Mechanical Turk are anonymous, it is hard to know the motivation of those conducting the task. It is reasonable to assume that some workers will attempt to earn money for little or no work. Alonso et al. described their attempts to ensure that the anonymous workers chosen were motivated and appropriate for the task.

Work in this area is still relatively novel and the success of such approaches requires more study.

6.3.4 Absolute vs. Relative/Preference Judgments

Virtually every test collection built has gathered its qrels using absolute judgments: asking an assessor to determine a document's relevance relative to a topic independent of other documents seen. Researchers have sometimes asked if these so-called *absolute* judgments are the most reliable approach to gathering qrels, suggesting instead that *relative* or *preference* judgments made between pairs of documents are sought instead. Some research addressed this question. In the context of concern about consistency of assessor judgments, Rees and Schultz [193] stated "*It is evident that the Judgmental Groups agree almost perfectly with respect to the relative ordering of the Documents.*" (p. 185). In contrast, Katter reported on an experiment the results of which showed that absolute judgments were more reliable [145].

A concern with relative judgments was that gathering a complete set required showing assessors every possible pairing of documents under consideration, an $O(n^2)$ problem. Rorvig conducted experiments examining the tractability of building test collections with relative judgments [203]. His results indicated that collections could be built from such judgments as preference judgments appeared to be transitive, which meant that some preferences could be reliably inferred, substantially cutting the number of judgments needing to be assessed. He proposed a methodology for building a test collection. More recently, Carterette et al. [48] showed that relative judgments drawn from users produced more reliable results than absolute. They also found that 99% of judgments gathered were transitive and went on to build on Rorvig's methods for reducing the number of preference judgments that needed to be made.

There does not as yet appear to have been a public test collection built from relative judgments. As will be seen in Section 7.2, however, deriving relevance judgments using preference from query logs is increasingly common.

6.3.5 Don't Use Assessors at All?

Soboroff et al. [238] examined the possibility of not using human input of any kind when creating relevance assessments. They hypothesized that judgments could be generated simply by randomly selecting commonly retrieved documents from the pool of runs used to form a test collection. The researchers examined the way in which runs submitted to several years of TREC ad hoc were ranked using such automatically generated qrels and compared the ranking using the standard qrels of TREC. They found that the two rankings were correlated relatively well. However, the most effective runs were ranked as poor or middle performing runs by the automated qrels. Consequently, the approach was judged to not work.

Aslam et al. [12] later pointed out that the method was ranking runs by their similarity to each other; those runs retrieving the most popular documents amongst the set of runs were ranked highest. The best runs retrieved relevant documents that few other runs found, such

documents were not in the automatic qrels, consequently the best runs were scored poorly. Soboroff et al.'s method was explored further, Wu and Crestani [287]; Shang and Li [227]; Can et al. [44] and later Nuray and Can [183]. As yet, no success has been found in fixing the important failing in Soboroff et al.'s approach.

Others suggested automatically creating both qrels and topics. A recent exploration of this area was described by Azzopardi et al. [16] who examined means of creating known item topics for a range of collections and languages; they were inspired by earlier work from Tague and Nelson [252]. The work demonstrated the potential for this approach, but the authors acknowledged that more investigation was needed.

6.3.6 Exploiting Structure or Other Data Sets

Although pooling was the predominant means of forming qrels, alternative approaches to seeking relevant documents were tried. What follows is a list of some proposals.

- Sheridan et al. [228] described building a small spoken document test collection of broadcast news items. To make relevance assessments easier, queries are referred to events that occurred on a specific date. This allowed the researchers to concentrate assessment to items broadcast on or after the date of the event.
- Using pre-existing manual organization of documents has been used on a number of occasions. Harmandas et al. [108] described the building of a web image test collection where assessors were encouraged to use the topical classifications present on many websites to locate relevant items. Haveliwala et al. [109] used a similar approach using the topical grouping of the Open Directory website to locate relevant related documents.
- Cleverdon building his Cranfield II collection composed of scientific papers used references in papers to identify potentially relevant material. This approach was further developed by Ritchie et al. [194].

- When working in the enterprise web search domain, Hawking et al. [113] described using the sitemaps of a website (a page that maps out for users the location of important pages on a website) as a source of relevance judgments for known item searching. The known item being the identified page, the query for that item being a title extracted from the sitemap page.
- Amitay et al. [9] proposed *trels*. For each topic in a test collection it was proposed to manually form a set of words that defined what was and was not a relevant document. Once a stable set of trels was formed, unjudged documents were assessed against the trels to determine relevance. Amitay and her collaborators showed trels to be successful in a simulation on TREC data. This approach was also used to build reusable question answering test collections, see Lin and Katz [166].
- Jensen et al. [135], in the context of web search, tested the combining of manual relevance judgments with judgments mined from a website taxonomy, such as DMOZ. They were able to show that the additional judgments improved evaluation accuracy.
- In the field of personalized search, use of bookmark or URL tagging data has been used as an approximation to relevance judgments in a personalized searching system. See for example, Xu et al. [288].
- An ever increasing body of work has examined the use of search engine logs to help determine relevance of items. Morita and Shinoda [182] explored using the time that a retrieved item was viewed as a way to infer the relevance of the item. More common was the use of click data in logs to determine relevance, e.g., Fox et al. [87]. So much log data is being generated particularly within large web search engines, that there is extensive research in analyzing log data and exploiting it. A description of the work in this area is described in Section 7.

With the exception of using query logs, none of the methods described has been as thoroughly tested as pooling.

6.4 Which is the Best Effectiveness Measure?

Perhaps surprisingly, for a research field that so values evaluation, it would appear that for many decades there was no quantitative research into the relative merits of different effectiveness measures. This was rectified in recent years through two forms of study: calculating the correlation between evaluation measures and assessing the stability of measures.

6.4.1 Correlating Measures

Tague-Sutcliffe and Blustein [254] were the first to quantitatively compare evaluation measures, establishing a methodology that became the standard for most subsequent research. Taking archived runs TREC, Tague-Sutcliffe and Blustein used different precision-based evaluation measures to each rank the runs. Correlations measured between the ranks showed strong similarities across the measures. The researchers concluded that there was little value in calculating different precision-based measures. However, more recent investigations, e.g., Buckley and Voorhees [39] and Thom and Scholer [255], showed that high precision measures, such as P(10) and P(1), correlated less well with measures such as MAP or R-precision.

6.4.2 Measuring Measure Stability

Zobel [295] devised a method to test the predictive power of evaluation measures. The core role of a test collection is to determine which retrieved method will produce the best effectiveness when used in an operational setting. Zobel simulated this testing and operational setup by splitting the topics of a test collection in half: one-half was treated as a mini test collection, the other half was a simulation of the operational setting. Using TREC-5 run data, Zobel took pairs of runs and determined which was the best on the mini collection and then measured

if the winning run was still the best in the operational setting. If it was, then a correct prediction was made using the reduced collection, if the pairs had swapped order, the result measured on the collection was a mistake. Using this *swap method*, Zobel determined which of four precision based the measures produced better predictions. He reported that although P(10) and $P(100)$ were worse at predicting than 11 point interpolated AP, in his judgment, the difference between the measures was too small to be of concern.

Using an alternative method, Buckley and Voorhees exploited the TREC-8 query track test collection [37], which had 21 so-called *query sets*: manually generated variations of each of the 50 topics in the collection. Each of the sets was run against a range of different retrieval systems resulting in 9 runs for each of the 21 sets. Buckley and Voorhees sought an evaluation measure that ranked the runs consistently over the query sets that also produced the smallest number of ties. They reported that measures such as MAP, R-precision, and P(1000) were the most stable; P(10) and P(1) the least. Buckley and Voorhees judged MAP to have the best balance between high stability and few ties.

In a separate study Voorhees and Buckley [276] applied Zobel's swap method to a wide range of TREC test collections and again confirmed that rank cutoff measures like P(10) were less accurate at predicting the effectiveness of runs than measures like MAP. One possible reason for the difference between Zobel's ambivalent and Voorhees and Buckley's more emphatic conclusions about a measure like P(10) was that Zobel used his measures in conjunction with a significance test, Voorhees and Buckley did not.

Sanderson and Zobel [222] pointed out that when comparing two measures, such as, MAP and P(10), the effort required to judge the relevant documents for MAP was substantially higher than that required to assess P(10); where only the top 10 documents from each run need be examined. Analyzing nine years of TREC data, Sanderson and Zobel showed that P(10) required between 11% and 14% of the assessor effort required to calculate MAP. The researchers concluded that P(10) was far more stable than MAP per equal quantity of assessor effort. If the qrels of a test collection already exist, then Sanderson and Zobel's point on the value of P(10) over MAP was not important. If one was

evaluating retrieval systems without a test collection, where assessors still had to judge the relevance of documents, then consideration of assessor effort was critical.

In contrast to most papers suggesting that MAP produces stable ranks of runs, Soboroff [236] used the swap method on a test collection with a small number of relevant documents per topic: the TREC 2003 topic distillation collection. He found that P(10) was noticeably more stable than R-precision and MAP. Soboroff also showed that MRR can be stable when used in a collection with a large number of topics (≥ 100). Further means of testing stability were described by Bodoff and Li [26] using Cronbach's alpha (a statistic that measures co-variance); and Sakai [208] who used the bootstrap test to count statistical significance between pairs of runs when measured with a particular evaluation measure.

It is worth remembering that the work on measure stability, while valuable, has its limitations. An "evaluation measure" could be created that ranks the runs of different retrieval systems by an alphabetical sorting of the run's name: e.g., a run labeled "Inquery" would be ranked higher than a run labeled "Okapi", which would be ranked higher than "Terrier". Under every stability test described here, this useless measure is perfectly stable; Sakai's significance count methodology would result in the maximum number of observable significant differences, and the Cronbach's alpha approach would show perfect co-variance.

Ignoring questions of stability, Aslam et al. [15] used a maximum entropy-based method to explore the degree to which an evaluation measure predicted the distribution of relevant and non-relevant documents across a retrieved list. Essentially, in this work the aim was to understand how well the single value from an evaluation measure summarized the distribution of relevant and non-relevant documents. Aslam et al. found that average precision was the better measure compared to R-precision and precision measured at fixed rank.

In this section, the assessment of measures was achieved by comparing a relatively simple property of each measure against some ideal. In the following section, the outputs of test collections and evaluation measures were compared with models of user searching behavior.

6.5 Do Test Collections or Measures Predict User Behavior?

A series of experiments were conducted to measure how well predictions made using test collections or evaluation measures correlated with a range of user behaviors when searching on systems under test. Results from this work are contradictory; the research described here is broken into those that concluded that little or no correlation existed, those that showed some link and those that showed a stronger link. Finally the apparent contradictions between these sets of work are discussed.

6.5.1 Little Correlation Found

In testing the impact of a searching system on user behavior, one can choose to measure effectiveness scores of users searching on an operational system and look for correlations between the scores and some aspect of user behavior or an outcome from the search. A number of studies took this approach. Tagliacozzo [251] showed that 18% of ∼900 surveyed MEDLINE (a medical literature search engine) users did not appear to be satisfied with search results despite them containing a large number of relevant retrieved documents. Su [247] attempted to correlate many measures of IR performance with user satisfaction. She found that precision did not correlate significantly with satisfaction and examined this issue in more detail later [248]. Hersh et al. [120] examined medical students' ability to answer clinical questions after searching on MEDLINE. Expert assessors were used to calculate recall and precision of the students' search outputs looking for correlations between these measures and the scores students attained for the questions. The researchers reported no correlation. Similar work was conducted by Huuskonen and Vakkari [126] producing similar negative results.

Hersh et al. [122] were the first to try to correlate test collection-based results with user behavior. They examined a pair of IR systems measured as significantly different on a small test collection; when subjects used one of the pair of systems, no significant difference in user behavior was observed. This experiment was repeated on another small collection with the same perhaps surprising conclusion [258]. See also

a more detailed examination of the experiments [259]. Using a method of artificially creating ranked document lists each with a different level of MAP, Turpin and Scholer [260] described a larger experiment that showed some small significant differences in user behavior when there were large differences in MAP between the artificial ranks.

Smith and Kautor [234] engaged 36 users to each search 12 information gathering topics on two versions of a web search engine: one the normal searching system, the other a version of the engine which displayed results starting from rank 300, presumably much worse. No significant difference in user success in finding relevant items was observed. Smith et al. reported that users adapted to the poorer system by issuing more queries; this change appeared to mitigate the smaller number of relevant documents retrieved in each search.

To many researchers, the totality of this work highlighted the artificiality of test collections. Ingwersen and Järvelin [128, p. 234] provided a detailed survey of past work that outlined the limitations of what an experimental result on a test collection can tell the researcher. The collective results from these works were viewed by some as strong evidence that there was a problem with the test collection methodology.

6.5.2 Some Correlation Found

Allan et al. [6] studied the problem of locating relevant text fragments, called *facets*. The researchers created artificial document rankings displaying fragments and links to full document texts. The rankings were formed starting by randomly degrading a perfect ranking. Users were asked to identify within the rankings, sections of documents that were relevant to a topic. Subjects were given many hours to complete the task. Allan et al. measured the time users took to complete their task, their error rate, and their facet recall. Unlike previous work, the researchers found a correlation between user behavior and test collection-based evaluation measures.

Huffman and Hochster [124] addressed the question of how effectively a test collection can be used to predict user satisfaction. They described getting two sets of assessors to judge the search results of 200 queries: the first assessors judged the relevance of the top three results;

and the second set of assessors judged user satisfaction with the overall results. The researchers reported finding a correlation between DCG measured on the relevance judgments and user satisfaction.

Al-Maskari et al. [3] conducted a small study measuring correlations between user satisfaction measures and different evaluation measures based on examinations of Google searches. She showed that there was a strong correlation between user rankings of results and the ranking produced by the evaluation measures she tested. She found that Cumulative Gain (CG) correlated better with user measures than P(10), DCG, and nDCG. Later, she and others used a test collection to select a pair of retrieval systems that had noticeably different effectiveness scores on a particular topic [4]. They then measured how well groups of users performed on those two systems for that topic. Fifty-six users searched from a selection of 56 TREC topics. The researchers showed a correlation between test collection experiments and user behavior, though they noted that user satisfaction was harder to predict than more objective measures such as the number of relevant documents saved.

6.5.3 Strong Correlation Found

When conducting an analysis of click log data, Joachims claimed that "*It appears that users click on the (relatively) most promising links ... independent of their absolute relevance*" [136]. He described experimental results showing that users, given different versions of an IR system, clicked at almost the same average rank position, despite there being differences in the effectiveness of the three versions. Joachims highlighted Rees and Schultz's [193] past work on relative relevance judgments and proposed an alternative approach for measuring user interaction with different systems. His suggestion was to interleave the outputs of the different systems into a single ranking and observe if users tended to click more on results from one ranking over another. The results of this *preference*-based experiment showed that users chose the results from the better ranking in a statistically significantly measurable way. This work was repeated later by Radlinski et al. [192], showing the same results.

Inspired by Joachims, Thomas and Hawking [256] presented a different preference methodology that allowed users to express a preference for not only the ranking of a retrieval system but potentially its interface as well. In their methodology, two versions of a search engine result were presented side-by-side to users. Users could query the two engines and interact with them as normal. Thomas et al. presented in the two panels, the top 10 results of Google and the presumably worse Google results in ranks 21–30. The researchers observed a clear statistically significant preference for the results from the top ranks over the lower-ranked results.

6.5.4 Discussion

The work showing little correlation might lead some to question the value of test collections; however, it is notable that many of the studies in the opening sub-section failed to show statistical significance in the user-based tests. A lack of significance can mean that there is no measurable difference or it can mean that the experiment was not powerful enough to allow such differences to be measured (see Section 5). The challenge of accurately measuring users was pointed out by Voorhees [274] who suggested that the experiments, such as those from Hersh and Turpin et al., concluding failure in test collections may in fact have failed to measure their users' behavior accurately enough. Perhaps the strongest conclusions to draw from these collective works is that faced with a poor search, or worse a poor IR system, users either make do with the documents they are shown or they work around the system to manage to achieve their search task successfully. The last tranche of studies contrasts with the former as user's performance was assessed in a relative instead of an absolute way. From that work, it would appear that given a choice between two systems, users prefer to use the better system as a source of retrieval results.

6.6 Conclusions

This section examined in some detail the range of research that tested many aspects of the test collection method. Assessor consistency was

re-examined and was generally found to be un-problematic. Pooling was found to produce a sample of relevant documents to effectively rank runs. Means of building collections more efficiently were proposed and a number of those methods adopted. Evaluation measures were examined in detail and the importance of selecting the right measure for the right task was highlighted. Finally, the consistency with which test collection results predicted user behavior on operational system was examined. Perhaps the simplest conclusion to draw here is that measuring users accurately requires care.

7

Alternate Needs and Data Sources for Evaluation

As shown in Section 6, over the past decade, a detailed examination of the construction and use of test collections was conducted that by and large found the long-standing evaluation methodology to be a valid approach to measuring the effectiveness of IR systems. However, during that period, the needs of at least part of the IR research community changed and at the same time, new potential sources of information about the relevance of documents became more accessible to researchers. In this section, the new need is described and the data sets created for it are outlined. Also two new evaluation data sources are introduced. As much of this work is beyond the scope of a test collection review article, it is described here briefly.

7.1 Learning to Rank

Test collections and evaluation measures are commonly used for the purposes of comparison: deciding if one approach or one retrieval system is better than another. However, there is a related use, retrieval function optimization. The ranking functions of IR systems are increasingly complex, containing a wide range of parameters for

which optimal settings need to be found. A common approach to finding such values is to use a machine learning approach known as *Learning To Rank* (LTR). The study of LTR has its origins in the late 1980s [89]; see [168] for other early LTR papers. Although a resurgence of interest started around 2005, from the point of view of evaluation, work is still in its infancy. There are two key evaluation areas to consider: data sets and evaluation measures.

7.1.1 Data Sets

As with any machine learning approach, data is needed to train an LTR retrieval function, which is then tested on a separate data set. In LTR, the data are generally composed of the classic components of a test collection: documents, topics, and qrels. It is notable that in his pioneering work, Fuhr stated that a key concern was the approach "*needs large samples of relevance feedback data for its application*", by which he meant training and testing data. Fuhr used a test collection with > 240 usable topics [91]. The first shared LTR data set was the LETOR benchmark [168]. It was composed of two existing IR test collections: OHSUMED and TREC web, which together had a similar number of topics to Fuhr's earlier collection. A series of features were extracted from all relevant and top-ranked documents in relation to each topic in the data set. Machine learning groups who were not interested in extracting such features from the documents could simply apply the features to their learning algorithms.

The collection quickly became a standard for use in LTR experiments. However, it is relatively new and recent publications have suggested that certain biases exist within it [176]. It is likely that adjustments to LETOR to correct these biases will arise as will the creation of new LTR collections. However, it is not clear if adapting existing IR test collections will produce large enough data sets for the LTR community. Web search companies, such as Yahoo!, have released custom built LTR data sets[1]; exploiting sources such as query logs to build data sets are an active area of research, which are described in Section 7.2.

[1] http://learningtorankchallenge.yahoo.com/ (accessed April 26, 2010).

7.1.2 Evaluation Measures

An LTR function is trained with respect to a particular evaluation measure. Liu et al. [168] described training using a series of common measures: P(n), nDCG, and MAP. It was assumed that the measure to use when optimizing was the one that reflected, most accurately, a model of the user in the operational setting one is optimizing for. Recent research, however, from Yilmaz and Robertson [291] showed that measures that make the greatest use of available training data can in fact be the better measure to employ. For example, although one might argue that P(10) is a more accurate model of a typical casual search engine user. If one optimizes on that measure, relevance information from only the top 10 documents will be used. If instead, one optimizes on MAP, relevance information from across the document ranking will be used. Yilmaz and Robertson showed that LTR systems trained on MAP and tested on P(10) produced better rank optimization than systems trained and tested on P(10).

Because the range of parameters in a retrieval functions can be very large, it is impossible to exhaustively explore every possible combination. In order to optimize an LTR system effectively, techniques drawn from the machine learning community, such as gradient ascent, are used. However, Robertson and Zaragoza [201] showed that the current suite of existing evaluation measures are not ideal for use with gradient ascent and related learning techniques. In their paper they argued that new measures need to be built to ensure that optimization can be achieved more successfully. This is likely to be an area that will come to factor more significantly in future evaluation surveys.

7.2 Query Logs — Modeling Users

For as long as automated searching systems existed, logs of activities on those systems were gathered and studied. An early example is Meister and Sullivan [174], who, studying the NASA/RECON citation search engine, examined both the volume of searches and the number of retrieved items that were viewed. Inductive studies of user behavior as recorded in *query logs* continued from that time, growing considerably

with the introduction of web search engines and the selective release of large public data sets from them; see [130] for an overview of that research. Such use of logs in this way was influential in IR researchers' understanding of user behavior, from the shortness of query length to the prevalence of spelling mistakes.

Section 6.3.6 briefly mentioned research exploiting data in logs to help generate conventional test collections: Fox et al. [87] showed that it was possible to use clicks as indicators of relevance. However, more recent research showed that in order to use such data, noise and bias needed to be removed. Noise was introduced to logs by automated programs repeatedly querying a search engine either to gather information or to try to deliberately spam the search engine in some way. Simple methods for identifying the activities of information gathering systems were found to be relatively straightforward: Jansen et al., for example, removed search sessions that had > 100 queries [131]. Detecting spam data, which will be engineered to be as similar to user interactions as possible, is harder to spot. Description of that work is beyond the scope of this monograph.

Bias in the query logs arises from the way that users interact with search engines. Joachims et al. [138] identified two forms of user bias, what they called *trust bias* (in other publications, this was called *presentation bias*) and *quality bias*. Trust bias was given its name due to users' willingness to trust the search engine to find the most relevant item in the top ranks. Joachims demonstrated the strength of this bias by manipulating search results, deliberately placing non-relevant documents in top ranked positions and showing that users still commonly clicked on the top position. With the second form of bias, Joachims showed that when the overall quality of search results was poor, users appeared willing to click on less relevant documents.

Joachims et al.'s conclusions were that extracting absolute relevance judgments from a query log was hard and as an alternative, proposed that relative or *preference judgments* should be extracted. For example, if a user clicked on the item in the 2^{nd} rank position but not the 1^{st}, one would infer that the item at rank 2 was more relevant than the item at rank 1. Joachims later showed that how such preference judgments were used in LTR [139], as did Zheng et al. [294].

Agichtein et al. [1] used query logs to learn how to customize search results for individual users. They removed trust bias from query logs by building a model of the typical bias toward certain rank positions and then subtracted that bias from the query and click log data of the user under study. This work was notable as it was one of the first to use the technique of *click prediction*. Here the researchers split the query log into two parts. They trained their system to a particular user (in this case using the first 75% of the log) and then used the system to predict which result that user would click on for the queries they submitted in the remaining part of the log, thus determining if the user model was accurate See also Piwowarski et al. [190] for further work in this area.

Joachim's observations of bias in user clicks were an initial attempt to model user behavior when examining search results. A series of models were subsequently proposed and tested on extensive collections of search log data often using click prediction. Craswell et al. [72] showed how modeling behavior simply based on document rank was not ideal. They introduced what they referred to as a *cascade model* where the probability of a click on a search result was dependent on the probability of the current result being relevant and of the higher ranked results not being relevant. See Dupret and Piwowaeski [80] and Chapelle et al. [52] for further extension to and testing of the cascade model.

Query logs were also used to validate evaluation measures. Chapelle et al. [52] compared a range of evaluation measures, using a combination of assessor-based relevance judgments and click data from a large query log.

7.3 Live Labs

The involvement of users in evaluation of IR systems has long been advocated and conducted as is recorded and promoted in the works of Ingwersen and Järvelin [128], Saracevic [225], and Borlund [30, 31]. A key limiting factor in the experimental methods promoted by such researchers is the challenge of finding a sufficiency of users. Given there are now very large numbers of people who have high speed access to the Internet, new forms of search evaluation are possible, using what

has sometimes been called *live labs*. This rather broad term covers a range of experimental methodologies, which we outline here.

An early example was the work of Dumais et al. [79] who as part of the testing of their desktop search engine, Stuff I've Seen, deployed a working version of the system, which was installed by 234 users (employees of a large organization) who used the system as their desktop search tool. The search engine was instrumented to log certain information about user interaction, which enabled the researchers to understand how the system was used and how often. Unbeknownst to the employees, the researchers randomly deployed different versions of the search interface and using the logs were able to determine how the versions affected searcher behavior. This approach of deploying software to willing volunteers/users was used by others, e.g., [75].

When working with services accessed over a network, such as a search engine, it is possible to make changes to the searching system at regular intervals without the users of the engine to have to install any updates. Such activity was described at a conference panel by Cutting [163] where he stated that several updates to commercial search engines in a single day was not an unusual occurrence. After each change, search logs could be examined to observe any change in user behavior. Joachims appeared to be the first to publish on this topic (see Section 6.5.3), describing a methodology for measuring user preferences for two different versions of a search engine. A key part of Joachims' approach was that users were unaware they were being given a choice between two different searching systems.[2] In a later paper working with Radlinski et al. [192], Joachims, deployed this methodology to a popular academic paper searching system for a month and was able to observe user behavior for over 20,000 queries.

A number of IR researchers were inspired by von Ahn's ESP game [266], where users label images as part of their activities while playing a multi-user game. Clough et al., [62] keen to study cross language searching created an image finding system built on top of Flickr and

[2] Cooper [67] proposed an evaluation methodology where experimenters would go to the site where an IR system was being used and observe a random sample of users conducting their search tasks on either an existing system or a new trial system. The users would not know which system they were being shown.

through user interactions with the game were able to study interaction. Kazai et al. [150] created a game that involved players making relevance judgments on documents.

While enticing as approaches to creating large-scale evaluations or data sets for evaluation, all three methods are challenging to implement. The first requires the software being deployed to be of a high standard before users will willingly engage with it for a long period of time. The second method requires the experimenter to have access to a popular search engine so as to manipulate its results. The third requires high-quality software to be developed where the game play is enticing enough for a sufficient number of people to participate.

There is, however, another approach, as mentioned in Section 6.3.3, it is possible, using services like Mechanical Turk, to pay people to conduct short-run tasks. The small amount of money they are willing to work for means many people can be employed. The example task described in the earlier section was that of judging the relevance of a document for the purposes of building a test collection, however, the potential range of tasks is broader than this: annotating corpora, seeking user opinion of search interfaces, and comparing result rankings are just some of the possibilities. Exploiting systems like Mechanical Turk for IR research is relatively new with little research to review as yet. There are challenges to using such services, but nevertheless, the service is likely to be increasingly used.

8

Conclusions

This monograph presented a brief history of the development of test collection-based evaluation from the earliest works through to the highly influential TREC exercises. Next a series of prominent evaluation measures were described and research testing the properties of those measures was detailed. The need for and use of significance tests in IR experiments was outlined next. One can see that the IR community still uses the model for evaluation initiated by the pioneering work of Thorne, Cleverdon, and Gull in the early 1950s and consolidated by Cleverdon's Cranfield collections of the early 1960s. Most of the evaluation measures used by the community are closely related to the measures created by Gull and Kent et al. in the 1950s. The commonest significance tests used in research papers today are the same as those used by IR researchers in the late 1960s.

One might expect for the research community to discover flaws in such a long-standing methodology. A great deal of research conducted in the past decade has tried specifically to determine if such flaws exist. However, the results of the research are some new evaluation measures; some useful alternatives to the means by which test collections are built; but ultimately the research has validated the test collection

approach. The components of a test collection — a set of documents, a set of topics, and a list of qrels — while a somewhat artificial construct remains at the core of experimental validation of new methodologies in IR. It is clear that query logs offer a means of constructing noisy though vast testing sets that are particularly helpful in new lines of IR research such as LTR. However, it is likely that this approach will not be a replacement, instead offering a complementary methodology to the long standing and proven approach of measuring the effectiveness of an IR system on a test collection.

Acknowledgments

I am most grateful to Paul Clough and Peter Willett for their helpful comments while preparing this monograph and two of my students Azzah Al-Maskari and Shahram Sedghi, both of whom were working on evaluation topics and whose work and conversations were particularly stimulating while writing. In addition, Evangelos Kanoulas, Ian Soboroff, Chris Buckley, and Ellen Voorhees at TREC were continually helpful in charting out more recent developments in evaluation and being willing to discuss some of the finer (and admittedly quite nerdy) points of IR evaluation. Stefan Rüger, Peter Bath, and Andrew Holmes were a tremendous help in guiding my understanding of significance tests. Finally, I wish to thank the reviewers for their detailed and invaluable comments after examining the earlier versions of this monograph.

Obtaining primary sources in the early history of IR evaluation research would have been much harder had it not been for the invaluable and timely help of Donna Harman, whose formation of a digital library of the key IR texts made access to the old Cornell and Cranfield reports trivially easy. In addition Bill Maron and Keith van Rijsbergen helped me obtain copies of the early reports from Maron, Kuhns,

and Ray. Tefko Saracevic kindly scanned in some pages from one of his early IR reports. Peter Willett and Micheline Beaulieu's personal collection of books, preprints, and technical reports was an invaluable resource also. Finally, my efforts to locate old articles were helped by the long tradition of IR research conducted in my department in Sheffield, which meant that in a dark, lower basement corner of the University of Sheffield Western Bank Library, a wealth of 1940s, 1950s, and 1960s journals, books, proceedings and reports lay in wait for me and my photocopy card.

References

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning user interaction models for predicting web search result preferences," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–10, New York, NY, USA: ACM, 2006.

[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 5–14, ACM, 2009.

[3] A. Al-Maskari, M. Sanderson, and P. Clough, "The relationship between IR effectiveness measures and user satisfaction," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 773–774, New York, NY, USA: ACM Press, 2007.

[4] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio, "The good and the bad system: Does the test collection predict users' effectiveness?," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–66, New York, NY, USA: ACM, 2008.

[5] J. Allan, *Topic Detection and Tracking: Event-based Information Organization,* (The Kluwer International Series on Information Retrieval, vol. 12). Springer, 1st ed., 2002.

[6] J. Allan, B. Carterette, and J. Lewis, "When will information retrieval be "good enough"?," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 433–440, New York, NY, USA: ACM, 2005.

[7] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *ACM SIGIR Forum*, vol. 42, no. 2, pp. 9–15, 2008.

[8] D. G. Altman, *Practical Statistics for Medical Research*. Chapman & Hall/ CRC, 1st ed., 1990.

[9] E. Amitay, D. Carmel, R. Lempel, and A. Soffer, "Scaling IR-system Evaluation using Term Relevance Sets," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, pp. 10–17, New York, NY, USA: ACM, 2004.

[10] T. Arni, P. Clough, M. Sanderson, and M. Grubinger, "Overview of the ImageCLEFphoto 2008 photographic retrieval task," *Evaluating Systems for Multilingual and Multimodal Information Access*, Lecture Notes in Computer Science, *5706/2009*, 500-511. doi:10.1007/978-3-642-04447-2_62, 2009.

[11] J. Artiles, S. Sekine, and J. Gonzalo, "Web people search: Results of the first evaluation and the plan for the second," in *Proceedings of the 17th International Conference on World Wide Web*, pp. 1071–1072, New York, NY, USA: ACM Press, 2008.

[12] J. A. Aslam, V. Pavlu, and R. Savell, "A unified model for metasearch, pooling, and system evaluation," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 484–491, New York, NY, USA: ACM, 2003.

[13] J. A. Aslam, V. Pavlu, and E. Yilmaz, "A statistical method for system evaluation using incomplete judgments," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 541–548, ACM, 2006.

[14] J. A. Aslam, E. Yilmaz, and V. Pavlu, "A geometric interpretation of r-precision and its correlation with average precision," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 573–574, New York, NY, USA: ACM, 2005.

[15] J. A. Aslam, E. Yilmaz, and V. Pavlu, "The maximum entropy method for analyzing retrieval measures," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 27–34, New York, NY, USA: ACM, 2005.

[16] L. Azzopardi, M. de Rijke, and K. Balog, "Building simulated queries for known-item topics: An analysis using six European languages," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 455–462, New York, NY, USA: ACM, 2007.

[17] L. Azzopardi and V. Vinay, "Retrievability: An evaluation measure for higher order information access tasks," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 561–570, ACM Press: New York, NY, USA, 2008.

[18] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999.

[19] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz, "Relevance assessment: Are judges exchangeable and does it matter," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 667–674, New York, NY, USA: ACM, 2008.

[20] M. Baillie, L. Azzopardi, and I. Ruthven, "A retrieval evaluation methodology for incomplete relevance assessments," in *Advances in Information Retrieval*, Lecture Notes in Computer Science, vol. 4425, pp. 271–282, 2007.

[21] M. Barbaro and T. Zeller Jr, "A face is exposed for AOL Searcher No. 4417749," *The New York Times*. Retrieved from http://www.nytimes.com/2006/08/09/technology/09aol.html, August 9 2006.

[22] J. R. Baron, D. D. Lewis, and D. W. Oard, "TREC-2006 legal track overview," in *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, NIST Special Publication, vol. 500, pp. 79–98, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2006.

[23] S. Bjørner and S. C. Ardito, "Online before the internet, Part 1: Early pioneers tell their stories," *Searcher: The Magazine for Database Professionals*, vol. 11, no. 6, 2003.

[24] D. C. Blair, "STAIRS redux: Thoughts on the STAIRS evaluation, ten years after," *Journal of the American Society for Information Science*, vol. 47, no. 1, pp. 4–22, 1996.

[25] D. C. Blair and M. E. Maron, "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Communications of the ACM*, vol. 28, no. 3, pp. 289–299, doi:10.1145/3166.3197, 1985.

[26] D. Bodoff and P. Li, "Test theory for assessing IR test collections," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 367–374, New York, NY, USA: ACM, 2007.

[27] T. Bompada, C. C. Chang, J. Chen, R. Kumar, and R. Shenoy, "On the robustness of relevance measures with incomplete judgments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 359–366, New York, NY, USA: ACM Press, 2007.

[28] A. Bookstein, "When the most "pertinent" document should not be retrieved — An analysis of the Swets model," *Information Processing & Management*, vol. 13, no. 6, pp. 377–383, 1977.

[29] H. Borko, *Evaluating The: Effectiveness of Information Retrieval Systems* (No. Sp-909/000/00). Santa Monica, California: Systems Development Corporation, 1962.

[30] P. Borlund, "The concept of relevance in IR," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 10, pp. 913–925, 2003.

[31] P. Borlund and P. Ingwersen, "The development of a method for the evaluation of interactive information retrieval systems," *Journal of Documentation*, vol. 53, pp. 225–250, 1997.

[32] H. Bornstein, "A paradigm for a retrieval effectiveness experiment," *American Documentation*, vol. 12, no. 4, pp. 254–259, doi:10.1002/asi.5090120403, 1961.

[33] M. Braschler and C. Peters, "Cross-language evaluation forum: Objectives, results, achievements," *Information Retrieval*, vol. 7, no. 1–2, pp. 7–31, 2004.

[34] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, doi:10.1145/792550.792552, 2002.

[35] E. C. Bryant, "Progress towards evaluation of information retrieval systems," in *Information Retrieval Among Examining Patent Offices: 4th Annual*

Meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices (ICIREPAT), pp. 362–377, Spartan Books, Macmillan, 1966.

[36] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees, "Bias and the limits of pooling for large collections," *Information Retrieval*, vol. 10, no. 6, pp. 491–508, 2007.

[37] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, New York, NY, USA: ACM, 2000.

[38] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–32, New York, NY, USA: ACM, 2004.

[39] C. Buckley and E. M. Voorhees, "Retrieval system evaluation," in *TREC: Experiment and Evaluation in Information Retrieval*, pp. 53–75, MIT Press, 2005.

[40] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, Bonn, Germany, 2005.

[41] R. Burgin, "Variations in relevance judgments and the evaluation of retrieval performance," *Information Processing & Management*, vol. 28, no. 5, pp. 619–627, doi:10.1016/0306-4573(92)90031-T, 1992.

[42] S. Büttcher, C. L. A. Clarke, and I. Soboroff, "The TREC 2006 terabyte track," in *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006)*, vol. 500, pp. 128–141, Maryland, USA: Gaithersburg, 2006.

[43] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff, "Reliable information retrieval evaluation with incomplete and biased judgements," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 63–70, New York, NY, USA: ACM Press, 2007.

[44] F. Can, R. Nuray, and A. B. Sevdik, "Automatic performance evaluation of Web search engines," *Information Processing and Management*, vol. 40, no. 3, pp. 495–514, 2004.

[45] B. Carterette, "Robust test collections for retrieval evaluation," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 55–62, New York, NY, USA: ACM Press, 2007.

[46] B. Carterette, "On rank correlation and the distance between rankings," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 436–443, ACM, 2009.

[47] B. Carterette, J. Allan, and R. Sitaraman, "Minimal test collections for retrieval evaluation," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 268–275, New York, NY, USA: ACM, 2006.

[48] B. Carterette, P. Bennett, D. Chickering, and S. Dumais, "Here or There," in *Advances in Information Retrieval*, pp. 16–27, 2008. Retrieved from http://dx.doi.org/10.1007/978-3-540-78646-7_5.

[49] B. Carterette and R. Jones, "Evaluating search engines by modeling the relationship between relevance and clicks," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 217–224, 2007.

[50] B. Carterette and R. Jones, "Evaluating search engines by modeling the relationship between relevance and clicks," *Advances in Neural Information Processing Systems*, vol. 20, pp. 217–224, 2008.

[51] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan, "Evaluation over thousands of queries," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 651–658, New York, NY, USA: ACM, 2008.

[52] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 621–630, New York, NY, USA: ACM Press, 2009.

[53] H. Chen and D. R. Karger, "Less is more: Probabilistic models for retrieving fewer relevant documents," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 429–436, New York, NY, USA: ACM, 2006.

[54] C. L. A. Clarke, N. Craswell, and I. Soboroff, "Preliminary report on the TREC 2009 Web track," *Working notes of the proceedings of TREC 2009*, 2009.

[55] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 659–666, New York, NY, USA: ACM, 2008.

[56] C. L. A. Clarke, M. Kolla, and O. Vechtomova, "An effectiveness measure for ambiguous and underspecified queries," in *Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10–12, 2009 Proceedings*, pp. 188–199, New York Inc: Springer-Verlag, 2009.

[57] C. W. Cleverdon, "The evaluation of systems used in information retrieval (1958: Washington)," in *Proceedings of the International Conference on Scientific Information — Two Volumes*, pp. 687–698, Washington: National Academy of Sciences, National Research Council, 1959.

[58] C. W. Cleverdon, *Report on the Testing and Analysis of an Investigation Into the Comparative Efficiency of Indexing Systems*. ASLIB Cranfield Research Project. Cranfield, UK, 1962.

[59] C. W. Cleverdon, *The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages*, (Cranfield Library Report No. 3). Cranfield Institute of Technology, 1970.

[60] C. W. Cleverdon, "The significance of the cranfield tests on index languages," in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12,

Chicago, Illinois, United States: ACM Press New York, NY, USA, 1991. doi:10.1145/122860.122861.

[61] C. W. Cleverdon and M. Keen, "Factors Affecting the Performance of Indexing Systems," Vol 2. *ASLIB, Cranfield Research Project. Bedford, UK: C. Cleverdon*, pp. 37–59, 1966.

[62] P. Clough, J. Gonzalo, J. Karlgren, E. Barker, J. Artiles, and V. Peinado, "Large-scale interactive evaluation of multilingual information access systems — The iCLEF flickr challenge," in *Proceedings of Workshop on Novel Methodologies for Evaluation in Information Retrieval*, pp. 33–38, Glasgow, UK, 2008.

[63] P. Clough, H. Muller, T. Deselaers, M. Grubinger, T. M. Lehmann, J. Jensen, and W. Hersh, "The CLEF 2005 cross-language image retrieval track," in *Accessing Multilingual Information Repositories*, Lecture Notes in Computer Science, vol. 4022, pp. 535–557, 2006.

[64] P. Clough, M. Sanderson, and H. Muller, "The CLEF cross language image retrieval track (ImageCLEF) 2004," *Lecture notes in Computer Science*, pp. 243–251, 2004.

[65] W. S. Cooper, "Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems," *American Documentation*, vol. 19, no. 1, pp. 30–41, doi:10.1002/asi.5090190108, 1968.

[66] W. S. Cooper, "A definition of relevance for information retrieval," *Information storage and retrieval*, vol. 7, no. 1, pp. 19–37, 1971.

[67] W. S. Cooper, "On selecting a measure of retrieval effectiveness," *Journal of the American Society for Information Science*, vol. 24, no. 2, 1973.

[68] G. V. Cormack and T. R. Lynam, "Statistical precision of information retrieval evaluation," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 533–540, New York, NY, USA: ACM, 2006.

[69] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke, "Efficient construction of large test collections," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 282–289, New York, NY, USA: ACM, 1998.

[70] N. Craswell, A. de Vries, and I. Soboroff, "Overview of the trec-2005 enterprise track," in *Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005)*, Gaithersburg, Maryland, USA, 2005.

[71] N. Craswell and D. Hawking, "Overview of the TREC 2002 Web track," in *The Eleventh Text Retrieval Conference (TREC-2002)*, NIST Special Publication 500-251, pp. 86–95, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2003.

[72] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proceedings of the international conference on Web search and web data mining*, pp. 87–94, ACM, 2008.

[73] W. B. Croft, "A file organization for cluster-based retrieval," in *Proceedings of the 1st Annual International ACM SIGIR Conference on Information Storage and Retrieval*, pp. 65–82, New York, NY, USA: ACM, 1978.

[74] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Addison Wesley, 1st ed., 2009.

[75] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin, "Fast, flexible filtering with phlat," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 261–270, New York, NY, USA: ACM, 2006.

[76] A. Davies, *A Document Test Collection for Use in Information Retrieval Research*, (Dissertation). Department of Information Studies. University of Sheffield, 1983.

[77] G. Demartini and S. Mizzaro, "A classification of IR effectiveness metrics," in *Advances in Information Retrieval*, Lecture Notes in Computer Science, vol. 3936, pp. 488–491, 2006.

[78] B. K. Dennis, J. J. Brady, and J. A. Dovel, "Index manipulation and abstract retrieval by computer," *Journal of Chemical Documentation*, vol. 2, no. 4, pp. 234–242, 1962.

[79] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff I've seen: A system for personal information retrieval and re-use," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72–79, ACM, 2003.

[80] G. E. Dupret and B. Piwowarski, "A user browsing model to predict search engine click data from past observations," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 331–338, ACM, 2008.

[81] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical Science*, vol. 1, no. 1, pp. 54–77, 1986.

[82] B. Efron and R. J. Tibshirani, "An introduction to the bootstrap," *Monographs on Statistics and Applied Probability*, vol. 57, pp. 1–177, 1993.

[83] R. A. Fairthorne, "Implications of test procedures," in *Information Retrieval in Action*, pp. 109–113, Cleveland, Ohio, USA: Western Reserve UP, 1963.

[84] E. M. Fels, "Evaluation of the performance of an information-retrieval system by modified Mooers plan," *American Documentation*, vol. 14, no. 1, pp. 28–34, doi:10.1002/asi.5090140105, 1963.

[85] E. A. Fox, *Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts*, (Computer Science Technical Reports). Cornell University. Retrieved from http://techreports.library.cornell.edu:8081/Dienst/UI/1.0/Display/cul.cs/TR83-561, 1983.

[86] E. A. Fox, *Virginia Disc One*. Blacksburg, VA, USA: Produced by Nimbus Records, 1990.

[87] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, "Evaluating implicit measures to improve web search," *ACM Transactions on Information Systems (TOIS)*, vol. 23, no. 2, pp. 147–168, 2005.

[88] H. P. Frei and P. Schäuble, "Determining the effectiveness of retrieval algorithms," *Information Processing and Management: An International Journal*, vol. 27, no. 2–3, pp. 153–164, 1991.

[89] N. Fuhr, "Optimum polynomial retrieval functions based on the probability ranking principle," *ACM Transactions on Information Systems*, vol. 7, no. 3, pp. 183–204, 1989.

[90] N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, "INEX: INitiative for the Evaluation of XML retrieval," in *Proceedings of the SIGIR 2002 Workshop on XML and Information Retrieval*, 2002.

[91] N. Fuhr and G. E. Knorz, "Retrieval test evaluation of a rule based automatic indexing (AIR/PHYS)," in *Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 391–408, UK: British Computer Society Swindon, 1984.

[92] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 107–130, Gaithersburg, Maryland, USA, 2000.

[93] F. Gey, R. Larson, M. Sanderson, H. Joho, P. Clough, and V. Petras, "GeoCLEF: The CLEF 2005 cross-language geographic information retrieval track overview," in *Accessing Multilingual Information Repositories*, Lecture Notes in Computer Science, vol. 4022, pp. 908–919, 2006.

[94] G. Gigerenzer, "Mindless statistics," *Journal of Socio-Economics*, vol. 33, no. 5, pp. 587–606, 2004.

[95] W. Goffman, "On relevance as a measure," *Information Storage and Retrieval*, vol. 2, pp. 201–203, 1964.

[96] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A brief history," in *Proceedings of the 16th Conference on Computational Linguistics — vol. 1*, pp. 466–471, Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from http://portal.acm.org/citation.cfm?id=992628.992709, 1996.

[97] C. D. Gull, "Seven years of work on the organization of materials in the special library," *American Documentation*, vol. 7, no. 4, pp. 320–329, doi:10.1002/asi.5090070408, 1956.

[98] D. K. Harman, "Evaluation issues in information retrieval," *Information Processing & Management*, vol. 28, no. 4, pp. 439–440, 1992.

[99] D. K. Harman, "Overview of the second text retrieval conference (TREC-2)," in *NIST Special Publication. Presented at the Second Text Retrieval Conference (TREC 2)*, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1993.

[100] D. K. Harman, "Overview of the third text retrieval conference (TREC-3)," in *The Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA*, NIST Special Publication, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1994.

[101] D. K. Harman, "Overview of the fourth text retrieval conference (TREC-4)," in *The Forth Text Retrieval Conference (TREC-4), Gaithersburg, MD, USA*, NIST Special Publication, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1995.

[102] D. K. Harman, "Overview of the second text retrieval conference (TREC-2)," *Information Processing and Management*, vol. 31, no. 3, pp. 271–289, 1995.

[103] D. K. Harman, "Overview of the TREC 2002 novelty track," in *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*, pp. 46–56, Gaithersburg, Maryland, USA, 2002.

[104] D. K. Harman, "Some Interesting Unsolved Problems in Information Retrieval," Presented at the Center for Language and Speech Processing, Workshop 2002, The Johns Hopkins University 3400 North Charles Street, Barton Hall Baltimore, MD 21218. Retrieved from http://www.clsp.jhu.edu/ws02/preworkshop/lecture_harman.shtml, July 2 2002.

[105] D. K. Harman, "The TREC ad hoc experiments," in *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*, pp. 79–98, MIT Press, 2005.

[106] D. K. Harman and C. Buckley, "The NRRC reliable information access (RIA) workshop," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 528–529, New York, NY, USA: ACM, 2004.

[107] D. K. Harman and G. Candela, "Retrieving records from a gigabyte of text on a minicomputer using statistical ranking," *Journal of the American Society for Information Science*, vol. 41, no. 8, pp. 581–589, 1990.

[108] V. Harmandas, M. Sanderson, and M. D. Dunlop, "Image retrieval by hypertext links," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 296–303, New York, NY, USA: ACM, 1997.

[109] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk, "Evaluating strategies for similarity search on the web," in *Proceedings of the 11th International Conference on World Wide Web*, pp. 432–442, New York, NY, USA: ACM Press, 2002.

[110] D. Hawking, "Overview of the TREC-9 Web track," in *NIST Special Publication*, pp. 87–102, 2001. Presented at the Ninth Text Retrieval Conference (TREC-9), Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[111] D. Hawking, P. Bailey, and N. Craswell, "ACSys TREC-8 Experiments," in *NIST Special Publication*, pp. 307–316, 2000. Presented at the Eighth Text Retrieval Conference (TREC-8), Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[112] D. Hawking, N. Craswell, and P. Thistlewaite, "Overview of TREC-7 very large collection track," in *The Seventh Text Retrieval Conference (TREC-7)*, pp. 91–104, NIST Special Publication, 1998. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[113] D. Hawking, F. Crimmins, and N. Craswell, "How valuable is external link evidence when searching enterprise Webs?," in *Proceedings of the 15th Australasian database conference*, vol. 27, pp. 77–84, Darlinghurst, Australia: Australian Computer Society, Inc, 2004.

[114] D. Hawking and S. E. Robertson, "On collection size and retrieval effectiveness," *Information Retrieval*, vol. 6, no. 1, pp. 99–105, 2003.

[115] D. Hawking and P. Thistlewaite, "Overview of TREC-6 very large collection track," in *The Sixth Text Retrieval Conference (TREC-6)*, pp. 93–106, NIST Special Publication, 1997. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[116] M. A. Hearst, *Search User Interfaces*. Cambridge University Press, 1st ed., 2009.

[117] M. A. Hearst and C. Plaunt, "Subtopic structuring for full-length document access," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–68, New York, NY, USA: ACM, 1993.

[118] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "OHSUMED: An interactive retrieval evaluation and new large test collection for research," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 192–201, New York, NY, USA: Springer-Verlag New York, Inc, 1994.

[119] W. Hersh, A. M. Cohen, P. Roberts, and H. K. Rekapalli, "TREC 2006 genomics track overview," in *The Fifteenth Text Retrieval Conference*, pp. 52–78, Gaithersburg, Maryland, USA, 2006.

[120] W. Hersh, M. K. Crabtree, D. H. Hickam, L. Sacherek, C. P. Friedman, P. Tidmarsh, and C. Mosbaek et al., "Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions," *Journal of American Medical Informatics Association*, vol. 9, 2002.

[121] W. Hersh and P. Over, "TREC-9 Interactive Track Report," in *proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pp. 41–50, Gaithersburg, Maryland: NTIS, 2000.

[122] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson, "Do batch and user evaluations give the same results?," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 17–24, New York, NY, USA: ACM Press, 2000.

[123] J. E. Holmstrom, "Section III. Opening plenary session," in *The Royal Society Scientific Information Conference, 21 June–2 July 1948: Report and papers submitted*, London: Royal Society, 1948.

[124] S. B. Huffman and M. Hochster, "How well does result relevance predict session satisfaction?," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 567–574, New York, NY, USA: ACM Press, 2007.

[125] D. Hull, "Using statistical testing in the evaluation of retrieval experiments," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329–338, New York, NY, USA: ACM, 1993.

[126] S. Huuskonen and P. Vakkari, "Students' search process and outcome in Medline in writing an essay for a class on evidence-based medicine," *Journal of Documentation*, vol. 64, no. 2, pp. 287–303, 2008.

[127] E. Ide, "New experiments in relevance feedback," in *Report ISR-14 to the National Science Foundation*, Cornell University, Department of Computer Science, 1968.

[128] P. Ingwersen and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, 2005.

[129] M. Iwayama, A. Fujii, N. Kando, and A. Takano, "Overview of patent retrieval task at NTCIR-3," in *Proceedings of the third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[130] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing and Management*, vol. 42, no. 1, pp. 248–263, 2006.

[131] B. J. Jansen, A. Spink, and S. Koshman, "Web searcher interaction with the Dogpile.com metasearch engine," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 5, pp. 744–744, 2007.

[132] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, New York, NY, USA: ACM, 2000.

[133] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[134] E. Jensen, *Repeatable Evaluation of Information Retrieval Effectiveness In Dynamic Environments*. Illinois Institute of Technology. Retrieved from http://ir.iit.edu/∼ej/jensen_phd_thesis.pdf, May 2006.

[135] E. C. Jensen, S. M. Beitzel, A. Chowdhury, and O. Frieder, "Repeatable evaluation of search services in dynamic environments," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 1, p. 1, doi:10.1145/1292591.1292592, 2007.

[136] T. Joachims, "Evaluating retrieval performance using clickthrough data," in *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, pp. 12–15, 2002.

[137] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, New York, NY, USA: ACM Press, 2002.

[138] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 154–161, New York, NY, USA: ACM, 2005.

[139] T. Joachims and F. Radlinski, "Search engines that learn from implicit feedback," *Computer*, pp. 34–40, 2007.

[140] J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. E. Robertson, "INEX 2007 evaluation measures," pp. 24–33, Retrieved from http://dx.doi.org/10.1007/978-3-540-85902-4_2, 2008.

[141] N. Kando, "Evaluation of information access technologies at the NTCIR workshop," in *Comparative Evaluation of Multilingual Information Access Systems. 4th Workshop of the Cross-Language Evaluation Forum, CLEF*, pp. 29–43, Springer, 2003.

[142] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka, "Overview of IR tasks at the first NTCIR workshop," in *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 11–44, 1999.

[143] P. Kantor and E. M. Voorhees, "The TREC-5 Confusion Track," in *The Fifth Text Retrieval Conference (TREC-5)*, NIST Special Publication.

Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1996.

[144] P. B. Kantor and E. M. Voorhees, "The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text," vol. 2, no. 2, pp. 165–176, 2000.

[145] R. V. Katter, "The influence of scale form on relevance judgments," *Information Storage and Retrieval*, vol. 4, no. 1, pp. 1–11, 1968.

[146] J. Katzer, M. J. McGill, J. A. Tessier, W. Frakes, and P. DasGupta, "A study of the overlap among document representations," *Information Technology: Research and Development*, vol. 1, no. 4, pp. 261–274, 1982.

[147] J. Katzer, J. A. Tessier, W. Frakes, and P. DasGupta, *A Study of the Impact of Representations in Information Retrieval Systems*. Syracuse, New York: School of Information Studies, Syracuse University, 1981.

[148] G. Kazai and M. Lalmas, "INEX 2005 evaluation measures," in *Advances in XML Information Retrieval and Evaluation*, Lecture Notes in Computer Science, vol. 3977, pp. 16–29, 2006.

[149] G. Kazai, M. Lalmas, and A. P. de Vries, "The overlap problem in content-oriented XML retrieval evaluation," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72–79, New York, NY, USA: ACM, 2004.

[150] G. Kazai, N. Milic-Frayling, and J. Costello, "Towards methods for the collective gathering and quality control of relevance assessments," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 452–459, ACM, 2009.

[151] E. M. Keen, "Evaluation parameters," in *Report ISR-13 to the National Science Foundation*, Cornell University, Department of Computer Science, 1967.

[152] E. M. Keen, "Presenting results of experimental retrieval comparisons," *Information Processing and Management: An International Journal*, vol. 28, no. 4, pp. 491–502, 1992.

[153] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1–2, pp. 81–93, 1938.

[154] A. Kent, *Encyclopedia of Library and Information Science*. CRC Press, 2002.

[155] A. Kent, M. M. Berry, F. U. Luehrs Jr, and J. W. Perry, "Machine literature searching VIII. Operational criteria for designing information retrieval systems," *American Documentation*, vol. 6, no. 2, pp. 93–101, doi:10.1002/asi.5090060209, 1955.

[156] K. Kuriyama, N. Kando, T. Nozue, and K. Eguchi, "Pooling for a large-scale test collection: An analysis of the search results from the first NTCIR workshop," *Information Retrieval*, vol. 5, no. 1, pp. 41–59, 2002.

[157] M. Lalmas and A. Tombros, "Evaluating XML retrieval effectiveness at INEX," *SIGIR Forum*, vol. 41, no. 1, pp. 40–57, doi:10.1145/1273221.1273225, 2007.

[158] F. W. Lancaster, *Evaluation of the MEDLARS Demand Search Service*. (No. PB-178-660) (p. 278). Springfield, VA 22151: Clearinghouse for Federal Scientific and Technical Information, 1968.

[159] F. W. Lancaster, *Information Retrieval Systems Characteristics, Testing, and Evaluation*. John Wiley & Sons, Inc, 1968.

[160] R. Ledwith, "On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases," *Information Processing & Management*, vol. 28, no. 4, pp. 451–455, doi:10.1016/0306-4573(92)90003-I, 1992.

[161] C. Léger, J. P. Romano, and D. N. Politis, "Bootstrap technology and applications," *Technometrics*, vol. 34, no. 4, pp. 378–398, 1992.

[162] M. E. Lesk, "SIG — The significance programs for testing the evaluation output," in *Report ISR-12 to the National Science Foundation*, Cornell University, Department of Computer Science, 1966.

[163] M. E. Lesk, D. Cutting, J. Pedersen, T. Noreault, and M. Koll, "Real life information retrieval (panel): Commercial search engines," in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 333, New York, NY, USA: ACM, 1997.

[164] M. E. Lesk and G. Salton, "Relevance assessments and retrieval system evaluation*1," *Information Storage and Retrieval*, vol. 4, no. 4, pp. 343–359, doi:10.1016/0020-0271(68)90029-6, 1968.

[165] D. Lewis, "The TREC-5 filtering track," in *The Fifth Text Retrieval Conference (TREC-5)*, pp. 75–96, Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 1996.

[166] J. Lin and B. Katz, "Building a reusable test collection for question answering," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 7, pp. 851–861, 2006.

[167] H. Liu, R. Song, J. Y. Nie, and J. R. Wen, "Building a test collection for evaluating search result diversity: A preliminary study," in *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pp. 31–32, 2009.

[168] T. Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li, "Letor: Benchmark dataset for research on learning to rank for information retrieval," in *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pp. 3–10, 2007.

[169] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. De Rijke, and P. Rocha et al., "Overview of the CLEF 2004 multilingual question answering track," *Lecture notes in Computer Science*, vol. 3491, p. 371, 2005.

[170] P. Majumder, M. Mitra, D. Pal, A. Bandyopadhyay, S. Maiti, S. Mitra, and A. Sen et al., "Text collections for FIRE," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 699–700, ACM, 2008.

[171] R. Manmatha, T. Rath, and F. Feng, "Modeling score distributions for combining the outputs of search engines," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–275, New York, NY, USA: ACM Press, 2001.

[172] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[173] M. Maron, J. Kuhns, and L. Ray, *Probabilistic Indexing: A Statistical Technique for Document Identification and Retrieval* (Technical Memorandum No. 3) (p. 91), *Data Systems Project Office*. Los Angeles, California: Thompson Ramo Wooldridge Inc, 1959.

[174] D. Meister and D. Sullivan, "Evaluation of User Reactions to a Prototype On-line Information Retrieval System," Prepared under Contract No.

NASw-1369 by Bunker-Ramo Corporation, Canoga Park, CA. (No. NASA CR-918). NASA, 1967.

[175] M. Melucci, "On rank correlation in information retrieval evaluation," *ACM SIGIR Forum*, vol. 41, no. 1, pp. 18–33, 2007.

[176] T. Minka and S. E. Robertson, "Selection bias in the LETOR datasets," in *SIGIR Workshop on Learning to Rank for Information Retrieval*, pp. 48–51, 2008.

[177] S. Mizzaro and S. Robertson, "Hits hits TREC: exploring IR evaluation results with network analysis," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 479–486, New York, NY, USA: ACM, 2007.

[178] A. Moffat, W. Webber, and J. Zobel, "Strategic system comparisons via targeted relevance judgments," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 375–382, Amsterdam, The Netherlands, 2007.

[179] A. Moffat and J. Zobel, "Rank-biased precision for measurement of retrieval effectiveness," *ACM Transactions on Information Systems*, vol. 27, no. 1, p. Article No. 2, 2008.

[180] C. N. Mooers, *The Intensive Sample Test for the Objective Evaluation of the Performance of Information Retrieval System* (No. ZTB-132) (p. 20). Cambridge, Massachusetts: Zator Corporation, 1959.

[181] C. N. Mooers, "The next twenty years in information retrieval: Some goals and predictions," in *Papers Presented at the Western Joint Computer Conference*, pp. 81–86, ACM, 1959.

[182] M. Morita and Y. Shinoda, "Information filtering based on user behavior analysis and best match text retrieval," in *Proceedings of the 17th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 272–281, New York, NY, USA: Springer-Verlag New York, Inc, 1994.

[183] R. Nuray and F. Can, "Automatic ranking of information retrieval systems using data fusion," *Information Processing and Management*, vol. 42, no. 3, pp. 595–614, 2006.

[184] D. W. Oard, D. Soergel, D. Doermann, X. Huang, G. C. Murray, J. Wang, and B. Ramabhadran et al., "Building an information retrieval test collection for spontaneous conversational speech," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–48, ACM, 2004.

[185] I. Ounis, M. De Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the TREC-2006 blog track," in *Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006)*, pp. 17–31, Gaithersburg, Maryland, USA, 2006.

[186] P. Over, "TREC-6 interactive report," in *Proceedings of the sixth Text Retrieval Conference (TREC-6)*, NIST Special Publication, vol. 500, pp. 73–82, Gaithersburg, Maryland, USA, 1997.

[187] P. Over, "The TREC interactive track: An annotated bibliography," *Information Processing and Management*, vol. 37, no. 3, pp. 369–381, 2001.

[188] P. Over, T. Ianeva, W. Kraaij, A. F. Smeaton, and S. Valencia, "TRECVID 2006-An overview," in *Proceedings of the TREC Video Retrieval Evaluation Notebook Papers*, 2006.

[189] W. R. Pearson, "Comparison of methods for searching protein sequence databases," *Protein Science: A Publication of the Protein Society*, vol. 4, no. 6, p. 1145, 1995.

[190] B. Piwowarski, G. Dupret, and R. Jones, "Mining user web search activity with layered bayesian networks or how to capture a click in its context," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 162–171, New York, NY, USA: ACM, 2009.

[191] S. M. Pollock, "Measures for the comparison of information retrieval systems," *American Documentation*, vol. 19, no. 4, pp. 387–397, doi:10.1002/asi.5090190406, 1968.

[192] F. Radlinski, M. Kurup, and T. Joachims, "How does clickthrough data reflect retrieval quality?," in *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pp. 43–52, 2008.

[193] A. M. Rees and D. G. Schultz, "A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching," Final Report to the National Science Foundation. Volume II, Appendices. Clearinghouse for Federal Scientific and Technical Information, Springfield, VA. 22151 (PB-176-079), MF $0.65, HC $3.00), October 1967.

[194] A. Ritchie, S. Teufel, and S. Robertson, "Creating a test collection for citation-based IR experiments," in *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 391–398, Association for Computational Linguistics Morristown, NJ, USA, 2006.

[195] S. E. Robertson, "The parametric description of retrieval tests: Part II: Overall measures," *Journal of Documentation*, vol. 25, no. 2, pp. 93–107, 1969.

[196] S. E. Robertson, "On sample sizes for non-matched-pair IR experiments," *Information Processing and Management: An International Journal*, vol. 26, no. 6, pp. 739–753, 1990.

[197] S. E. Robertson, "Salton award lecture on theoretical argument in information retrieval," *SIGIR Forum*, vol. 34, no. 1, pp. 1–10, doi:10.1145/373593.373597, 2000.

[198] S. E. Robertson, "On GMAP: And other transformations," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 78–83, New York, NY, USA: ACM Press, 2006.

[199] S. E. Robertson, "On the history of evaluation in IR," *Journal of Information Science*, vol. 34, no. 4, pp. 439–456, doi:10.1177/0165551507086989, 2008.

[200] S. E. Robertson and D. A. Hull, "The TREC-9 filtering track final report," in *Proceedings of the Ninth Text REtrieval Conference (TREC-2001)*, pp. 25–40, Gaithersburg, Maryland, USA: NTIS, 2001.

[201] S. E. Robertson and H. Zaragoza, "On rank-based effectiveness measures and optimization," *Information Retrieval*, vol. 10, no. 3, pp. 321–339, 2007.

[202] G. Roda, J. Tait, F. Piroi, and V. Zenz, "CLEF-IP 2009: Retrieval experiments in the intellectual property domain," in *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

[203] M. E. Rorvig, "The simple scalability of documents," *Journal of the American Society for Information Science*, vol. 41, no. 8, pp. 590–598, doi: 10.1002/(SICI)1097-4571(199012)41:8<590::AID-ASI5>3.0.CO;2-T, 1990.

[204] D. Rose and C. Stevens, "V-twin: A lightweight engine for interactive use," in *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pp. 279–290, 1996.

[205] T. Sakai, "New performance metrics based on multigrade relevance: Their application to question answering," in *NTCIR-4 Proceedings*, 2004.

[206] T. Sakai, "Evaluating evaluation metrics based on the bootstrap," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 525–532, Seattle, Washington, USA: ACM Press New York, NY, USA, 2006. doi:10.1145/1148170.1148261.

[207] T. Sakai, "Alternatives to bpref," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 71–78, ACM, 2007.

[208] T. Sakai, "Evaluating information retrieval metrics based on bootstrap hypothesis tests," *Information and Media Technologies*, vol. 2, no. 4, pp. 1062–1079, 2007.

[209] T. Sakai and N. Kando, "On information retrieval metrics designed for evaluation with incomplete relevance assessments," *Information Retrieval*, vol. 11, no. 5, pp. 447–470, doi:10.1007/s10791-008-9059-7, 2008.

[210] G. Salton, *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

[211] G. Salton, *The Smart Retrieval System. Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[212] G. Salton, J. Allan, and C. Buckley, "Approaches to passage retrieval in full text information systems," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–58, New York, NY, USA: ACM, 1993.

[213] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Communications of the ACM*, vol. 26, no. 11, pp. 1022–1036, doi:10.1145/182.358466, 1983.

[214] G. Salton and M. E. Lesk, "Computer evaluation of indexing and text processing," *Journal of the ACM (JACM)*, vol. 15, no. 1, pp. 8–36, 1968.

[215] G. Salton and C. T. Yu, "On the construction of effective vocabularies for information retrieval," in *Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval Table of Contents*, pp. 48–60, New York, NY, USA: ACM, 1973.

[216] M. Sanderson, "Accurate user directed summarization from existing tools," in *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pp. 45–51, New York, NY, USA: ACM, 1998.

[217] M. Sanderson and H. Joho, "Forming test collections with no system pooling," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, New York, NY, USA: ACM, 2004.

[218] M. Sanderson and C. J. Rijsbergen, "NRT: News retrieval tool," *Electronic Publishing*, vol. 4, no. 4, pp. 205–217, 1991.

[219] M. Sanderson, T. Sakai, and N. Kando EVIA 2007: The First International Workshop on Evaluating Information Access, 2007.

[220] M. Sanderson and I. Soboroff, "Problems with Kendall's tau," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 839–840, New York, NY, USA: ACM, 2007.

[221] M. Sanderson, J. Tang, T. Arni, and P. Clough, "What else is there? Search diversity examined," in *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval*, pp. 562–569, Springer, 2009.

[222] M. Sanderson and J. Zobel, "Information retrieval system evaluation: Effort, sensitivity, and reliability," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162–169, New York, NY, USA: ACM, 2005.

[223] T. Saracevic, *An Inquiry into Testing of Information Retrieval Systems: Part II: Analysis of Results*. 1968. (No. CSL:TR-FINAL-II). Comparative Systems Laboratory: Final Technical Report. Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University.

[224] T. Saracevic, "RELEVANCE: A review of and a framework for the thinking on the notion in information science," *Journal of the American Society for Information Science*, vol. 26, no. 6, pp. 143–165, 1975.

[225] T. Saracevic, "Evaluation of evaluation in information retrieval," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 138–146, New York, NY, USA: ACM, 1995.

[226] J. Savoy, "Statistical inference in retrieval effectiveness evaluation," *Information Processing and Management*, vol. 33, no. 4, pp. 495–512, 1997.

[227] Y. Shang and L. Li, "Precision evaluation of search engines," *World Wide Web*, vol. 5, no. 2, pp. 159–173, doi:10.1023/A:1019679624079, 2002.

[228] P. Sheridan, M. Wechsler, and P. Schäuble, "Cross-language speech retrieval: Establishing a baseline performance," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 99–108, New York, NY, USA: ACM, 1997.

[229] M. Shokouhi and J. Zobel, "Robust result merging using sample-based score estimates," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 3, pp. 1–29, 2009.

[230] A. Smeaton and R. Wilkinson, "Spanish and Chinese document retrieval in TREC-5," in *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pp. 57–64, Gaithersburg, Maryland, USA, 1997.

[231] A. F. Smeaton, W. Kraaij, and P. Over, "TRECVID-An overview," in *Proceedings of the TRECVID 2003 Conference*, Gaithersburg, Maryland, USA. Retrieved from http://www-nlpir.nist.gov/projects/tvpubs/tvpapers03/tv3overview.pdf, 2003.

[232] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 321–330, New York, NY, USA: ACM, 2006.

[233] A. F. Smeaton, P. Over, and R. Taban, "The TREC-2001 video track report," in *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)*, pp. 52–60, 2001.

[234] C. L. Smith and P. B. Kantor, "User adaptation: Good results from poor systems," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 147–154, New York, NY, USA: ACM, 2008.

[235] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 623–632, New York, NY, USA: ACM, 2007.

[236] I. Soboroff, "On evaluating web search with very few relevant documents," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 530–531, 2004.

[237] I. Soboroff, "Dynamic test collections: Measuring search effectiveness on the live web," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 276–283, New York, NY, USA: ACM, 2006.

[238] I. Soboroff, C. Nicholas, and P. Cahan, "Ranking retrieval systems without relevance judgments," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 66–73, New Orleans, Louisiana, United States: ACM. doi:10.1145/383952.383961, 2001.

[239] I. Soboroff and S. E. Robertson, "Building a filtering test collection for TREC 2002," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 243–250, New York, NY, USA: ACM, 2003.

[240] E. Sormunen, "Liberal relevance criteria of TREC-: Counting on negligible documents?," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 324–330, New York, NY, USA: ACM, 2002.

[241] K. Spärck Jones, "Automatic indexing," *Journal of Documentation*, vol. 30, no. 4, pp. 393–432, 1974.

[242] K. Spärck Jones, *Information Retrieval Experiment*. Butterworth-Heinemann Ltd, 1981.

[243] K. Spärck Jones, "Letter to the editor," *Information Processing & Management*, vol. 39, no. 1, pp. 156–159, doi:10.1016/S0306-4573(02)00026-2, 2003.

[244] K. Spärck Jones and R. G. Bates, *Report on a design study for the 'ideal' information retrieval test collection* (British Library Research and Development Report No. 5428). Computer Laboratory, University of Cambridge, 1977.

[245] K. Spärck Jones and C. J. van Rijsbergen, *Report on the need for and the provision of an 'ideal' information retrieval test collection* (British Library Research and Development Report No. 5266) (p. 43). Computer Laboratory, University of Cambridge, 1975.

[246] K. Spärck Jones and C. J. van Rijsbergen, "Information retrieval test collections," *Journal of Documentation*, vol. 32, no. 1, pp. 59–75, doi:10.1108/eb026616, 1976.

[247] L. T. Su, "Evaluation measures for interactive information retrieval," *Information Processing & Management*, vol. 28, no. 4, pp. 503–516, 1992.

[248] L. T. Su, "The relevance of recall and precision in user evaluation," *Journal of the American Society for Information Science*, vol. 45, no. 3, pp. 207–217, 1994.

[249] J. A. Swets, "Information retrieval systems," *Science*, vol. 141, no. 3577, pp. 245–250, 1963.

[250] J. A. Swets, "Effectiveness of information retrieval methods," *American Documentation*, vol. 20, no. 1, pp. 72–89, doi:10.1002/asi.4630200110, 1969.

[251] R. Tagliacozzo, "Estimating the satisfaction of information users," *Bulletin of the Medical Library Association*, vol. 65, no. 2, pp. 243–249, 1977.

[252] J. M. Tague and M. J. Nelson, "Simulation of user judgments in bibliographic retrieval systems," in *Proceedings of the 4th Annual International ACM SIGIR Conference on Information Storage and Retrieval: Theoretical Issues in Information Retrieval*, pp. 66–71, New York, NY, USA: ACM, 1981.

[253] J. M. Tague-Sutcliffe, "Some perspectives on the evaluation of information retrieval systems," *Journal of the American Society for Information Science*, vol. 47, no. 1, pp. 1–3, doi:10.1002/(SICI)1097-4571(199601)47:1<1::AID-ASI1>3.0.CO;2-3, 1996.

[254] J. M. Tague-Sutcliffe and J. Blustein, "A statistical analysis of the TREC-3 data," in *The Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA*, pp. 385–398, NIST Special Publication, 1994. Department of Commerce, National Institute of Standards and Technology.

[255] J. A. Thom and F. Scholer, "A comparison of evaluation measures given how users perform on search tasks," *Presented at the Proceedings of the Twelfth Australasian Document Computing Symposium*, pp. 56–63, 2007.

[256] P. Thomas and D. Hawking, "Evaluation by comparing result sets in context," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 94–101, New York, NY, USA: ACM Press, 2006.

[257] R. Thorne, "The efficiency of subject catalogues and the cost of information searches," *Journal of Documentation*, vol. 11, pp. 130–148, 1955.

[258] A. Turpin and W. Hersh, "Why batch and user evaluations do not give the same results," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 225–231, New York, NY, USA: ACM, 2001.

[259] A. Turpin and W. Hersh, "User interface effects in past batch versus user experiments," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 431–432, New York, NY, USA: ACM, 2002.

[260] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 11–18, New York, NY, USA: ACM, 2006.

[261] C. J. van Rijsbergen, "Foundation of evaluation," *Journal of Documentation*, vol. 30, no. 4, pp. 365–373, 1974.

[262] C. J. van Rijsbergen, *Information Retrieval*. Butterworth-Heinemann Ltd, 2nd ed., 1979.

[263] P. K. T. Vaswani and J. B. Cameron, *The National Physical Laboratory Experiments in Statistical Word Associations and Their Use in Document Indexing And Retrieval.* National Physical Laboratory Computer Science Division-Publications; COM.SCI.42 (p. 171). National Physical Lab., Teddington (Great Britain), 1970.

[264] J. Verhoeff, W. Goffman, and J. Belzer, "Inefficiency of the use of Boolean functions for information retrieval systems," *Communications of the ACM*, vol. 4, no. 12, pp. 557–558, 1961.

[265] B. C. Vickery, *On Retrieval System Theory.* Butterworths, 2nd ed., 1965.

[266] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326, New York, NY, USA: ACM Press, 2004.

[267] E. M. Voorhees, "On expanding query vectors with lexically related words," in *The Second Text Retrieval Conference (TREC 2)*, NIST Special Publication 500-215, pp. 223–231, Department of Commerce, National Institute of Standards and Technology, 1993.

[268] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in information retrieval*, pp. 315–323, New York, NY, USA: ACM Press, 1998.

[269] E. M. Voorhees, "The TREC-8 question answering track report," in *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 77–82, Gaithersburg, Maryland, USA, 1999.

[270] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information Processing and Management*, vol. 36, no. 5, pp. 697–716, 2000.

[271] E. M. Voorhees, "Evaluation by highly relevant documents," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74–82, New Orleans, Louisiana, United States. New York, NY, USA: ACM Press, 2001.

[272] E. M. Voorhees, "Overview of the TREC 2003 question answering track," in *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003),* vol. 142, 2003.

[273] E. M. Voorhees, "Overview of the TREC 2004 robust retrieval track," in *The Thirteenth Text Retrieval Conference (TREC 2004)*, NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, 2005.

[274] E. M. Voorhees, "On test collections for adaptive information retrieval," *Information Processing and Management*, 2008.

[275] E. M. Voorhees, "Topic set size redux," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 806–807, ACM, 2009.

[276] E. M. Voorhees and C. Buckley, "The effect of topic set size on retrieval experiment error," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 316–323, New York, NY, USA: ACM, 2002.

[277] E. M. Voorhees and D. K. Harman, "Overview of the seventh text retrieval conference," in *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*, pp. 1–24, NIST Special Publication, 1998.

[278] E. M. Voorhees and D. K. Harman, "Overview of the eighth text retrieval conference (TREC-8)," in *The Eighth Text Retrieval Conference (TREC-8)*, pp. 1–24, NIST Special Publication, 1999. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

[279] E. M. Voorhees and D. K. Harman, "Overview of TREC 2001," in *NIST Special Publication 500-250*, pp. 1–15, Presented at the Tenth Text Retrieval Conference (TREC 2001), Government Printing Office, 2001.

[280] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, illustrated ed., 2005.

[281] C. Wade and J. Allan, *Passage Retrieval and Evaluation* (CIIR Technical Report No. IR-396). Amherst, MA, USA: University of Massachusetts, Amherst Center for Intelligent Information Retrieval, 2005.

[282] W. Webber, A. Moffat, and J. Zobel, "Score standardization for inter-collection comparison of retrieval systems," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 51–58, New York, NY, USA: ACM, 2008.

[283] W. Webber, A. Moffat, and J. Zobel, "Statistical power in retrieval experimentation," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp. 571–580, ACM, 2008.

[284] R. W. White and D. Morris, "Investigating the querying and browsing behavior of advanced search engine users," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 255–262, New York, NY, USA: ACM Press, 2007.

[285] D. Williamson, R. Williamson, and M. E. Lesk, "The Cornell implementation of the SMART system," in *The SMART Retrieval System: Experiments in Automatic Document Processing*, (G. Salton, ed.), p. 12, Englewood Cliffs, New Jersey: Prentice-Hall, Inc, 1971.

[286] I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigabytes*. Morgan Kaufmann, 1999.

[287] S. Wu and F. Crestani, "Methods for ranking information retrieval systems without relevance judgments," in *Proceedings of the 2003 ACM symposium on Applied computing*, pp. 811–816, New York, NY, USA: ACM, 2003.

[288] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 155–162, New York, NY, USA: ACM, 2008.

[289] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 102–111, New York, NY, USA: ACM Press, 2006.

[290] E. Yilmaz, J. A. Aslam, and S. E. Robertson, "A new rank correlation coefficient for information retrieval," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 587–594, New York, NY, USA: ACM, 2008.

[291] E. Yilmaz and S. E. Robertson, "On the choice of effectiveness measures for learning to rank," in *Learning to Rank for Information Retrieval. Workshop in Conjunction with the ACM SIGIR Conference on Information Retrieval*, Boston, MA, USA: ACM Press New York, NY, USA, 2009.

[292] T. Zeller Jr, "AOL Moves to Increase Privacy on Search Queries," *The New York Times*, Retrieved from http://www.nytimes.com/2006/08/22/technology/22aol.html, August 22 2006.

[293] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 10–17, New York, NY, USA: ACM Press, 2003.

[294] Z. Zheng, K. Chen, G. Sun, and H. Zha, "A regression framework for learning ranking functions using relative relevance judgments," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 287–294, ACM, 2007.

[295] J. Zobel, "How reliable are the results of large-scale information retrieval experiments?," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 307–314, New York, NY, USA: ACM Press, 1998.

Index

Foundations and Trends® in
Information Retrieval
Vol. 3, Nos. 1–2 (2009) 1–224

DOI: 10.1561/1500000012

now
the essence of knowledge

Methods for Evaluating Interactive Information Retrieval Systems with Users

By Diane Kelly

Contents

Foundations and Trends® in
Information Retrieval
Vol. 3, Nos. 1–2 (2009) 1–224

DOI: 10.1561/1500000012

now
the essence of knowledge

Methods for Evaluating Interactive Information Retrieval Systems with Users

Diane Kelly

School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, dianek@email.unc.edu

Abstract

This paper provides overview and instruction regarding the evaluation of interactive information retrieval systems with users. The primary goal of this article is to catalog and compile material related to this topic into a single source. This article (1) provides historical background on the development of user-centered approaches to the evaluation of interactive information retrieval systems; (2) describes the major components of interactive information retrieval system evaluation; (3) describes different experimental designs and sampling strategies; (4) presents core instruments and data collection techniques and measures; (5) explains basic data analysis techniques; and (4) reviews and discusses previous studies. This article also discusses validity and reliability issues with respect to both measures and methods, presents background information on research ethics and discusses some ethical issues which are specific to studies of interactive information retrieval (IIR). Finally, this article concludes with a discussion of outstanding challenges and future research directions.

1

Introduction

Information retrieval (IR) has experienced huge growth in the past decade as increasing numbers and types of information systems are being developed for end-users. The incorporation of users into IR system evaluation and the study of users' information search behaviors and interactions have been identified as important concerns for IR researchers [46]. While the study of IR systems has a prescribed and dominant evaluation method that can be traced back to the Cranfield studies [54], studies of users and their interactions with information systems do not have well-established methods. For those interested in evaluating interactive information retrieval systems with users, it can be difficult to determine how to proceed from a scan of the literature since guidelines for designing and conducting such studies are for the most part missing.

In interactive information retrieval (IIR), users are typically studied along with their interactions with systems and information. While classic IR studies abstract humans out of the evaluation model, IIR focuses on users' behaviors and experiences — including physical, cognitive and affective — and the interactions that occur between users and systems, and users and information. In simple terms, classic IR evaluation asks

the question, does this system retrieve relevant documents? IIR evaluation asks the question, can people use this system to retrieve relevant documents? IIR studies include both system evaluations as well as more focused studies of users' information search behaviors and their interactions with systems and information. IIR is informed by many fields including traditional IR, information and library science, psychology, and human–computer interaction (HCI). IIR has often been presented more generally as a combination of IR and HCI, or as a sub-area of HCI, but Ruthven [225] argues convincingly that IIR is a distinct research area. Recently, there has been interest in HCIR, or human computer information retrieval, but this looks similar to IIR and papers about this area have not established its uniqueness (e.g., [191]).

The proposition that IR systems are fundamentally interactive and should be evaluated from the perspective of users is not new. A review of IR literature reveals that many leaders in the field were writing about and studying interactive IR systems during the early years of IR research. For instance, Salton wrote a paper entitled "*Evaluation problems in interactive information retrieval*" which was published in 1970. In this paper, Salton [229] identified user effort measures as important components of IR evaluation, including the attitudes and perceptions of users. Cleverdon et al. [55] identified presentation issues and user effort as important evaluation measures for IR systems, along with recall and precision. Tague and Schultz [259] discuss the notion of user friendliness.

Some of the first types of IR interactions were associated with relevance feedback. Looking closely at this seemingly simple type of interaction, we see the difficulties inherent in IIR studies. Assuming that users are provided with information needs, each user is likely to enter a different query, which will lead to different search results and different opportunities for relevance feedback. Each user, in turn, will provide different amounts of feedback, which will create new lists of search results. Furthermore, causes and consequences of these interactions cannot be observed easily since much of this exists in the user's head. The actions that are available for observation — querying, saving a document, providing relevance feedback — are surrogates of cognitive activities. From such observable behaviors we must *infer* cognitive

activity; for instance, users who save a document may do so because it changes or adds to their understanding of their information needs.

User–system interactions are influenced by a number of other factors that are neither easily observable nor measurable. Each individual user has a different cognitive composition and behavioral disposition. Users vary according to all sorts of factors including how much they know about particular topics, how motivated they are to search, how much they know about searching, how much they know about the particular work or search task they need to complete, and even their expectations and perceptions of the IIR study [139, 194]. Individual variations in these factors mean that it is difficult to create an experimental situation that all people will experience the same, which in turn, makes it difficult to establish causal relationships. Moreover, measuring these factors is not always practical since there are likely a large number of factors and no established measurement practices.

The inclusion of users into any study necessarily makes IIR, in part, a behavioral science. As a result, appropriate methods for studying interactive IR systems must unite research traditions in two sciences which can be challenging. It is also the case that different systems, interfaces and use scenarios call for different methods and metrics, and studies of behavior and interaction suggest research designs that go beyond evaluation. For these reasons, there is no strong evaluation or experimental framework for IIR evaluations as there is for IR studies. IIR researchers are able to make many choices about how to design and conduct their evaluations, but there is little guidance about how to do this.

1.1 Purpose and Scope

There is a small body of research on evaluation models, methods, and metrics for IIR, but such studies are the exception rather than the rule (e.g., [34, 149]). In contrast to other disciplines where studies of methods and experimental design comprise an important portion of the literature, there are few, if any, research programs in IIR that investigate these issues and there is little formal guidance about how to conduct such studies, despite a long-standing call for such work

[231]. Tague's [260, 262] work and select chapters of the edited volume by Spärck-Jones [246] provide good starting points, but these writings are 15–20-years-old. While it might be argued that Spärck-Jones' book still describes the basic methodology behind traditional IR evaluations, Tague's work, which focuses on user-centered methods, needs updating given changes in search environments, tasks, users, and measures. It is also the case that Tague's work does not discuss data analysis. One might consult a statistics textbook for this type of information, but it can sometimes be difficult to develop a solid understanding of these topics unless they are discussed within the context of one's own area of study.

The purpose of this paper is to provide a foundation on which those new to IIR can make more informed choices about how to design and conduct IIR evaluations with human subjects.[1] The primary goal is to catalog and compile material related to the IIR evaluation method into a single source. This paper proposes some guidelines for conducting one basic type of IIR study — laboratory evaluations of experimental IIR systems. This is a particular kind of IIR study, but not the only kind. This paper is also focused more on quantitative methods, rather than qualitative. This is not a statement of value or importance, but a choice necessary to maintain a reasonable scope for this paper.

This article does not prescribe a step-by-step recipe for conducting IIR evaluations. The design of IIR studies is not a linear process and it would be imprudent to present the design process in this way. Typically, method design occurs iteratively, over time. Design decisions are interdependent; each choice impacts other choices. Understanding the possibilities and limitations of different design choices help one make better decisions, but there is no single method that is appropriate for all study situations. Part of the intellectual aspects of IIR is the method design itself. Prescriptive methods imply research can only be done in

[1] The terms *user* and *subject* are often used interchangeably in published IIR studies. A distinction between these terms will be made in Section 7. Since this paper focuses primarily on laboratory evaluations, the term *subject* will be used when discussing issues related to laboratory evaluations and *user* will be used when discussing general issues related to all IIR studies. *Subject* is used to indicate a person who has been sampled from the *user* population to be included in a study.

one way and often prevent researchers from discovering better ways of doing things.

The focus of this paper is on text retrieval systems. The basic methodological issues presented in this paper are relevant to other types of IIR systems, but each type of IIR system will likely introduce its own special considerations and issues. Additional attention is given to the study of different types of IIR systems in the final section of this paper. Digital libraries, a specific setting where IIR occurs, are also not discussed explicitly, but again, much of the material in this paper will be relevant to those working in this area [29].

Finally, this paper surveys some of the work that has been conducted in IIR. The survey is not intended to be comprehensive. Many of the studies that are cited are used to illustrate particular evaluation issues, rather than to reflect the state-of-the-art in IIR. For a current survey of research in IIR, see Ruthven [225]. For a more historic perspective, see Belkin and Vickery [23].

1.2 Sources and Recommended Readings

A number of papers about evaluation have been consulted in the creation of this paper and have otherwise greatly influenced the content of this paper. As mentioned earlier, the works of Tague [260, 262, 263, 264] and Tague and Schultz [259] are seminal pieces. The edited volume by Spärck-Jones [246] also formed a foundation for this paper.

Other research devoted to the study and development of individual components or models for IIR evaluation have also influenced this paper. Borlund [32, 34] has contributed much to IIR evaluation with her studies of simulated information needs and evaluation measures. Haas and Kraft [115] reviewed traditional experimental designs and related these to information science research. Ingwersen and Järvelin [139] present a general discussion of methods used in information seeking and retrieval research. Finally, the TREC Interactive Track [80] and all of the participants in this Track over the years have made significant contributions to the development of an IIR evaluation framework.

Review articles have been written about many topics discussed in this paper. These articles include Sugar's [255] review of user-centered

perspectives in IR and Turtle et al.'s [277] review of interactive IR research as well as Ruthven's [225] more recent version. The *Annual Review of Information Science and Technology* (*ARIST*) has also published many chapters on evaluation over its 40-year history including King's [173] article on the design and evaluation of information systems,[2] Kantor's [161] review of feedback and its evaluation in IR, Rorvig's [223] review of psychometric measurement in IR, Harter and Hert's [123] review of IR system evaluation, and Wang's [290] review of methodologies and methods for user behavior research.

Several special issues of journals about evaluation of IR and IIR systems are also worth mentioning. The most current is Borlund and Ruthven's [37] special issue of *IP&M* about evaluating IIR systems. Other special issues include Dunlop et al.'s [82] special issue of *Interacting with Computers* and Harman's [120] special issue of *IP&M*, which included Robertson and Hancock-Beaulieu's [221] discussion of changes in IR evaluation as a result of new understandings of relevance, interaction and information behavior. These articles, along with Savage-Knepshield and Belkin's [240] analysis of how IR interaction has changed over time, Saracevic's [233] assessment of evaluation in IR, and Ingwersen and Järvelin's [139] book on information seeking and retrieval are great background reading for those interested in the evolution of IIR systems and evaluation.

In addition to the sources from the IIR and IR literature, a number of sources related to experimental design and statistics were instrumental in the development of this paper: Babbie [13], Cohen [56], Gravetter and Wallnau [110], Myers and Well [200], Pedhazur and Schmelkin [208], and Williams [296].

1.3 Outline of Paper

The paper begins with a description of IIR and short discussion of its history. The next section reviews general approaches to studying IIR. Although this paper focuses on laboratory evaluations, other approaches are discussed briefly. Section 5 introduces

[2] Six articles were published in *ARIST* with the title, *Design and evaluation of information systems*, during the period 1968–1975.

research basics — research questions, theory, hypotheses, and variables. More advanced readers might want to skip this section, although the discussion of levels of measurement is particularly important for understanding the later material on statistics. Basic experimental designs are introduced in Section 6, followed by a discussion of sampling (Section 7). Instruments and data collection techniques are then presented in Section 8, followed by a discussion of some of the more common measures used in IIR evaluation (Section 10). A lengthy section on data analysis is in Section 11; although some instruction regarding qualitative data analysis is provided, this section primarily focuses on quantitative data analysis. This presentation starts with the basics of statistical data analysis, so advanced readers might want to skim parts of this section. Discussions of validity and reliability and research ethics are in Section 12. The paper concludes with future directions and challenges in Section 14.

2

What is IIR?

What is meant by *IIR*? An easy answer is that IIR is IR with users, but this does not really tell the whole story. One way to think about IIR is to place it in the middle of a continuum that is anchored by system focused studies and human focused studies (Figure 2.1). Studies situated at the system end of the spectrum are focused on developing and evaluating retrieval algorithms and indexing techniques. Studies such as those conducted in most TREC tracks would be examples of studies at this end of the continuum. There are no real users in these types of studies. Assessors may be used to create topics and evaluate documents, but they do not really function as users *per se*. Studies at the system end of the continuum can also be characterized by a lack of interaction — even if assessors are present, no searching takes place. Voorhees and Harman's [288] edited book describing TREC can be consulted for examples of these types of studies.

As we move along the continuum, the next type of study we observe are those that employ users to make relevance assessments of documents in relation to tasks. Users are basically used to build infrastructure so that a system-oriented study can be conducted. No searching is conducted and there is usually a lack of interest in users'

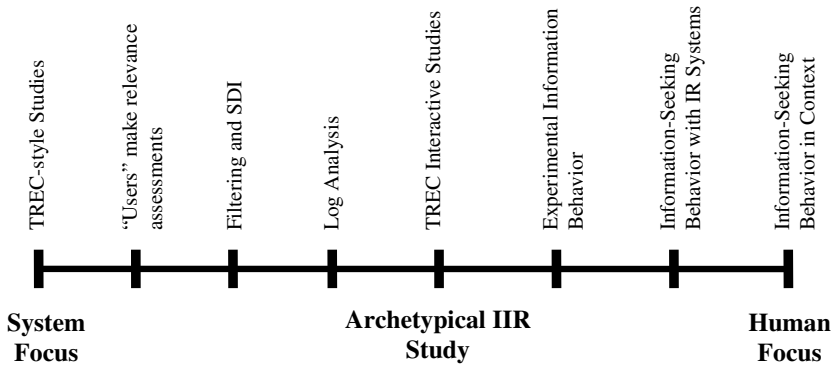


Fig. 2.1 Research continuum for conceptualizing IIR research.

search experiences and behaviors, and their interactions with systems. This type of study differs from a pure system-centered study because researchers recruit users to make assessments and build new infrastructure, rather than relying on the TREC infrastructure. This is often done because researchers are working on new problems or tasks that have not been addressed by TREC. For example, Teevan et al. [269] studied relevance feedback and personalization; this required the collection of queries, documents, and relevance assessments from users. Although it is possible to study the interaction between the user and the information need, or the user and documents, this is usually not the focus of this type of study.

Intent and purpose of the research are important in determining where a study belongs along the continuum. Consider a study where a system evaluation with users has been conducted but the researchers are primarily interested in demonstrating the goodness of the system, rather than understanding the user-system IR interaction; the user study is, in effect, an ancillary task rather than a central focus. In many ways, these types of studies undermine efforts to create a more solid foundation for IIR studies, since users are essentially treated as black boxes. Although it is not claimed that all IR studies should focus on users, an explicit mention of the focus of the study should be made so that readers can better distinguish between findings about IR systems, findings about interactive IR and findings about users. There is also

the question of whether it is even possible to claim that such a study is not at least in part a study of human behavior since users interact with the system and these interactions impact the system's responses. Although it is often claimed that the system is being evaluated and not the user, in practice this is difficult to do since a user is required for interaction.

As we move along the continuum, we come to studies where both the system and user are of interest, but there is still no interactive searching. These types of studies are less common today, but as the push towards personalization continues, we may see more studies of this type. A classic example of this type of study is evaluation of systems for the selective dissemination of information (SDI) from the 1970s, where the common evaluation model was to have users construct profiles and evaluate documents that were selected by the profiles [187]. Although users did not engage in searching, there was an attempt to understand how best to represent and update the user's information need and how best to present documents to users. The interaction, in these types of studies, was among the user, the information need and the documents. More current studies of proactive IR systems that observe users' behaviors while they do some activity (e.g., searching or writing) and fetch related articles are also in this class [41, 92].

Studies using transaction logs fall next on the spectrum. Although such studies have been around for quite some time, the availability of large search logs, in particular from search engine companies, have made this type of study very popular. Currently, the most popular type of analysis that is conducted with search engine logs looks at queries, search results and click-through behavior. These types of studies are primarily descriptive rather than explanatory, even though it is possible to model user behavior and interactions for certain situations. While many assumptions must be made about user intention, the shear amount of data that is available allows researchers to identify important regularities in Web search behavior. Furthermore, it is possible to manipulate some aspects of the search experience in a live test of a new interface feature or ranking algorithm and use the log data as a way to observe potential differences in performance [11]. Transaction log

analysis is often used in many types of studies, but studies represented at this place on the continuum use logs as the primary data source. Example studies are Agichtein et al. [3], Bilenko and White [28], and Jansen and Spink [146].

The next type of study is embodied by the TREC Interactive Track evaluation model. In this type of study, a system or interface feature is typically being evaluated. The design of such features is usually related directly to the human — whether it is human behavior and cognition, or information seeking context. The goals of such features are usually to better support the search process. This includes not only finding useful documents, but also supporting cognitive activities associated with search, such as sense-making. Studies located around this point on the continuum employ multiple methods of data collection, and usually include measures of system performance, interaction and usability. It is also common for studies around this point to use interviews to obtain qualitative feedback from users about their experiences. This point represents the classic or archetypical IIR study, which is the primary focus of this paper. Example studies are Joho and Jose [154] and White et al. [292].

As we move along the continuum, the next type of study focuses more on information behavior. In this type of study, the researcher controls aspects of the search process that are not typically controlled in the classic IIR study. For instance, the researcher might control what results are retrieved in response to a user's query or the order in which search results are presented to users. The point of such studies is to isolate and study individual aspects of the search process, rather than the entire process. One difficulty in studying the entire process is that each person experiences search differently; the goal of studies represented by this point on the continuum is to make these experiences as similar as possible so that causality can be studied with greater confidence. Studies represented by this point often use experimental methods that are commonly used in psychology. These studies focus on a slice of the search process and manipulation and control are often used. These studies are generally more interested in saying something specific about behavior, rather than on demonstrating the goodness of a particular IIR feature or system. Example studies are Arapakis and

Jose [12], Huang and Wang [135], Joachims et al. [152], Kelly et al. [168], Smith and Kantor [243], and Turpin and Scholer [276].

The next set of studies focus on general information search behavior in electronic environments. Most often there is no experimental system involved; researchers are instead interested in observing and documenting people's natural search behaviors and interactions. This might include studies of users' search tactics, studies of how users make relevance assessments, or studies of how users re-find information on the Web. Studies at this point also include those of searcher intermediaries and other professional searchers. Although these people are searching on behalf of others, they interact with the system to retrieve information. These studies differ from those mentioned above primarily in the methods they employ and the amount of the search process they examine. These studies often lead to better understandings of users' natural search behaviors, which is fundamental to the development of better IIR systems. However, these studies are not always driven by system concerns. Example studies are Byström [43], Ford et al. [99], Kellar et al. [163], and Kim and Allen [171].

Finally, the human-end of the spectrum can be characterized by research that focuses most exclusively on humans, their information needs and information behaviors. Researchers often insert themselves into a setting as an observer and gather data using qualitative techniques such as observation and interviews. In these studies, investigators explore the real information needs of users and their subsequent information seeking, within the particular context in which these needs arise without regard to a particular type of IR system. Solomon [244] provides a review of this type of research. Other work of this type can be found under the heading of everyday life information seeking [241]. Although results of these studies might inform the design of IR systems, this is usually not the primary goal. Example studies are Chatman [51] and Hirsh and Dinkelacker [131].

The primary focus of this paper is the archetypical IIR study, which represents IIR system evaluation with users. The archetypical IIR study represents a good entry point into IIR since it has a relatively long history, is most connected to traditional IR research, and is somewhat balanced with respect to computer and behavior sciences. Similar to

traditional IR evaluation, IIR evaluation has historically emphasized effectiveness, innovation, and application, sometimes at the cost of more basic scientific findings about interaction and behavior. Despite often being called *user-centered*, much of this research has primarily been about systems. As IIR evolves, it is hoped that equal emphasis will be placed on research that develops explanations of user behaviors and interactions that extend beyond an individual system.

Studies that span the spectrum from information seeking in electronic environments to log-based studies will also be discussed at some level because such studies form the core of IIR. Studies at either end of the spectrum will not be discussed since they are not considered as a core part of IIR, but rather as two different areas that frame IIR research. For a comprehensive discussion of theories of information seeking and methods for investigating information seeking behavior, readers are referred to Fisher et al. [93], Case [49] and Wildemuth [294]. For a discussion of common evaluation models used in systems-centered research, readers are referred to Spärck-Jones [246] and Robertson [220].

3

Background

3.1 Cognitive Viewpoint in IR

One important event which changed the way in which users were conceived and integrated into IR research was the *Workshop on the Cognitive Viewpoint* in Ghent, Belgium [70]. De Mey is noted as having first articulated the cognitive viewpoint at this workshop, although Ingwersen and Järvelin [139] note that this general viewpoint had been expressed by others at the time of De Mey's articulation. The cognitive viewpoint offered the first coherent alternative to what is known as the systems or algorithm viewpoint in IR [140]. The systems or algorithm viewpoint is the one most familiar to IR researchers — this viewpoint is embodied by the Cranfield and TREC evaluation models [288]. This viewpoint focuses on the system and makes a number of simplifying assumptions about users, and their needs and behaviors. Researchers adopting the system perspective are not blind to these assumptions, but maintain they are necessary to isolate and study individual aspects of the system. These assumptions have lead to the development of some strong evaluation norms for researchers working from the systems perspective.

The cognitive viewpoint embraces the complexity inherent in IR when users are involved and focuses attention on the cognitive activities that take place during information seeking and retrieval, and user-information, user-system interactions [138]. Ingwersen and Järvelin [139] identify five central and interrelated dimensions of the cognitive viewpoint[1]: (1) information processing takes place in senders *and* recipients of messages; (2) processing takes place at *different* levels; (3) during communication of information any actor is *influenced* by its past and present experiences (*time*) and its social, organizational and cultural environment; (4) individual actors *influence* the environment or domain; and (5) information is *situational* and *contextual*. While it is clear in viewing these dimensions that the cognitive viewpoint focuses on the user, Ingwersen and Järvelin [139] are carefully to point out that the cognitive viewpoint is not just about users' cognitive structures, but also about the numerous other cognitive structures represented in the IR system.[2] For instance, those represented by document authors and system developers. An examination of these five dimensions exemplifies some of the difficulties with conducting studies from this perspective. Specifically, that each individual user experiences IR in a different way, and that as soon as a user begins interacting with an IR system a series of cognitive changes take place which are unobservable, but will likely affect the user's subsequent interactions and behaviors. Similar to researchers working under the system perspective, researchers working under the cognitive perspective must also make simplifying assumptions and abstractions about some parts of the process.

Which perspective — the system or cognitive — is behind IIR research? Given the emphasis on the user in IIR it would appear that such research is guided by the cognitive perspective. However, a close look at discussions surrounding these two perspectives raises some questions. The systems or algorithm viewpoint is often referred to as the laboratory approach. If this is the case, then IIR surely falls under

[1] Emphasis is from Ingwersen and Järvelin [139].

[2] The cognitive viewpoint uses *system* in a more general sense to mean a combination of things or parts — this includes the technology, user, and environment.

the systems viewpoint since many IIR studies are in fact conducted in a laboratory.[3] Borlund [33] states that the systems viewpoint is concerned with achieving reliable results through control of experimental variables and repeatability of experiments, while the user-centered approach[4] is concerned with studying IR under real-life operational conditions. This would seem to preclude any controlled laboratory experiments from the cognitive viewpoint, another indication that perhaps IIR is part of the system perspective. Borlund [34] notes that a hybrid approach is needed for IIR that combines elements from the systems and cognitive perspectives and goes on to propose a framework for the evaluation of IIR systems. Ingwersen and Järvelin [139, p. 21] hint that IIR is a part of IR and represents the system perspective, perhaps because many IIR evaluations use standard TREC collections and thus, make simplifying assumptions about the nature of relevance. However, they describe IIR as being concerned with the "communication processes that occur during retrieval of information by involving all major participants in information seeking and retrieval, i.e., the searcher, the socio-organizational context, the IT setting, interface and information space". Although IIR evaluations have characteristics of each of the two major perspectives, fundamentally IIR is about humans, cognition, interactions, information and retrieval, and most IIR researchers would probably align their research with the cognitive perspective. Although many IIR evaluations use standard test collections and make simplifying assumptions about things such as relevance assessments, many of these studies are still being conducted to further our understanding of how people interact with systems to do IR.

3.2 Text Retrieval Conference

Three Tracks have attempted to develop a TREC-style evaluation framework for studying interaction and users: the Interactive Track (TRECs 3–11), the HARD Track (TRECs 12–14), and ciQA

[3] It is more likely the case that the label "laboratory" is inappropriate since it is a place, not an approach.

[4] According to Järvelin [147] the user-centered approach was subsumed by the cognitive perspective in the 1990s.

(TRECs 15–16). Each of these Tracks experimented with different types of evaluation frameworks, but none were successful at establishing a generic evaluation framework that allowed for valid cross-site comparisons. Of the three Tracks, the Interactive Track was the most responsible for developing some of the first accepted protocols and measures for IIR evaluation.

3.2.1 TREC Interactive Track

The TREC Interactive Track lasted nine years and made some of the most important contributions to the development of a method for IIR system evaluation [80, 128, 175]. The work of this Track will be summarized here; interested readers are referred to the chapter by Dumais and Belkin [80] in the edited book describing TREC [288].

In the initial year of the Track (TREC-3), participants were required to recruit subjects who were tasked with creating an optimal routing query. Fifty standard routing topics were provided and participants were allowed to recruit any number of subjects so long as each participating site contributed at least one routing query for each topic. Subjects developed their queries using training data and there was no standard protocol by which participating sites were required to operate. Participating sites could investigate any aspect of interactive behavior, including the influence of different system features on behavior. Participants experimented with a variety of interface and system features to assist subjects in completing the routing tasks. While subjects searched the training database when constructing their queries, they did not search for and save documents that they believed were relevant to the topic. Their only job was to construct routing queries. Major findings were that automatic techniques for creating routing queries performed better than human techniques, that the routing task was difficult for subjects to do and that a lack of standard protocol made it difficult to compare results.

In the next iteration of the Track (TREC-4), an *ad-hoc* search task was used instead of a routing task. The *ad-hoc* task required subjects to find and save as many relevant documents as possible for 25 topics. Subjects were also asked to create a final, best query for

the topic. Again, there was no standard protocol for administering the experiments, but participants were required to have at least one search conducted for each topic. Participants were also required to log each search, provide specific kinds of descriptive data about search behavior, including a narrative account of at least one search. Although subjects' job was to find and save documents they considered relevant to the topics, their judgments and performances were evaluated using relevance judgments made by TREC *assessors*, who created topics and established the benchmark relevance assessments.[5] Major findings from this iteration of the Track were that subjects' relevance assessments differed from assessors and that standard TREC metrics were not suitable to interactive searching scenarios because they were in part measuring other things besides performance — in particular, the extent to which subjects' relevance assessments matched assessors' — and they were computed based on 1000 retrieved documents which does not make much sense in an interactive setting since users are unlikely to examine 1000 documents.

The next year of the Track (TREC-5) saw a decline in participation, but ultimately laid the foundation for the next several iterations of the Track. In the previous two years the Track had four and ten participating sites, respectively, but only two sites completed participation in the Track at TREC-5, although others were involved with planning the Track [17, 21]. Based on experiences in TRECs 3 and 4, it was decided that the best approach would be to develop a method for comparing the performance of different IIR systems at different participating sites, rather than comparing human and system performance. For this Track, a new task was created which did not correspond to tasks used in other established TREC Tracks (e.g., routing and *ad-hoc*). This task was dubbed the *aspectual recall task* and required subjects to find documents that discussed different aspects of a topic rather than all documents relevant to a topic. This task was used by the Interactive Track in TRECs 5–8, with only slight modifications. Twelve topics

[5] Usually TREC assessors are recruited by the National Institute of Standards and Technology (NIST) which manages TREC. In most cases, these people are retried intelligence analysts, but some Tracks have used other kinds of people for topic creation and assessment.

were created based on previous *ad-hoc* topics (as compared to 50 routing topics in TREC-3 and 25 *ad-hoc* topics in TREC-4). Participating sites were required to provide the list of documents saved by subjects, search logs and a narrative account of one search. TREC assessors used the saved documents to compile a list of unique aspects and a key that showed which documents discussed which aspects. Aspectual recall was then computed, as well as standard precision.

In addition to these guidelines, the Track also developed an experimental design that required participating sites to implement two systems: a baseline, which was provided to participants, and an experimental system, which participants created. The Track also created topic and system rotations to control for order effects that required participating sites to study at least four subjects. These additional requirements were likely the reason why so few participants managed to complete the Track; it is reported in Dumais and Belkin [80] that even the two sites that did participate were unable to complete the experiment as planned. Thus, one of the major findings from this iteration related the amount of resources and time required to craft and implement the experimental design. There were concerns about the small number of topics (keeping in mind that topics are sampled just as users are) and results showed strong subject effects, topic effects and subject–topic interaction effects.

The TREC 6–8 Interactive Tracks basically used the same evaluation model that was developed in TREC-5 with some minor changes. Nine sites were able to complete the study in TREC-6 and it is described by Dumais and Belkin [80, p. 136] as being "the first true cross-site comparison in the interactive track". Only six topics were used, which allowed each subject to search for all topics (unfortunately this meant that each subject's experimental session lasted approximately 3 hours). The required, shared baseline system remained. The amount of data provided by the nine participating sites allowed for a more rigorous statistical analysis which found significant effects for topic, subject, and system. It was also the case that participants desired to create their own baseline systems that would allow them to better focus on interactive techniques that interested them. This major change happened in TREC-7. Note that this meant that cross-site comparisons

were no longer possible since each participating site created their own experimental and baseline systems.

TREC-9 represented a departure from the aspectual recall task which was in part motivated by a desire to reduce the amount of time required of subjects and explore additional types of tasks and collections. In this iteration, eight fact-finding tasks were created which consisted of four *n-answer* tasks (these tasks required subjects to find some number of answers in response to a question) and four *specific-comparison* tasks (these tasks required subjects to identify which of two provided answers was correct). The design required 16 subjects — a large increase over previous years, but since subjects were only given 5 min to answer each question, this resulted in experimental sessions that only lasted about one hour. Subjects were required to identify the answer and to save documents that helped them determine the answer. These answer-document pairs were then analyzed to determine performance.

The TREC-9 design only ran once and the Interactive Track moved to a 2-year cycle with TRECs 10–11 and focused on Web search. The goal of the first year was to define important issues and tasks worthy and possible of study in the Web environment. Subjects were provided with eight topics and searched the open Web. No standard instruments, experimental protocols or systems were required. In the second year, participants used the .gov collection from the TREC-11 Web Track and the Panoptic search engine was made available to participants [125]. The framework returned to a more tightly controlled experiment with a required design, protocol, and instruments.

Finally, in its last year (TREC-12), the Interactive Track was a subset of the Web Track. Subjects were asked to complete a topic distillation task, which asked them to construct a resource page (list of useful URLs) for a particular topic. The .gov corpus was used and two versions of the Panoptic search engine were made available to participants. Participants followed a specified experimental design and protocol. Interestingly, in this task, lists generated by human subjects were compared with lists generated automatically by systems, which was one of the original points of comparison in the Track during TRECs 3 and 4 (i.e., automatic versus human).

The work of the TREC Interactive Track resulted in many discoveries about IIR evaluation. First, assessors' relevance judgments were not generalizable and using these judgments to evaluate the performance of others was fraught with difficulty. Second, many standard evaluation metrics developed to assess system performance were not particularly useful in interactive settings. Third, including large numbers of topics in a laboratory evaluation was not reasonable given human limitations (both physical and cognitive). Larger numbers of topics require larger numbers of subjects and resources.

Although it was not possible to create an IIR evaluation model that was similar to the standard TREC evaluation model, the Interactive Track played an important role in establishing standards for IIR evaluation. This included a standard design and protocol, as well as standard techniques for reporting search logs and other data. The evaluation model of the archetypical IIR evaluation study derives directly from the work of this Track.

3.2.2 TREC HARD Track

The High Accuracy Retrieval of Documents (HARD) Track followed the Interactive Track and ran for three years [4, 5]. The set-up of this Track varied from year-to-year, but the primary focus was on single-cycle user-system interactions. These interactions were embodied in *clarification forms*. In most cases, participating sites used these forms to elicit feedback from assessors. Thus, the interaction consisted of a single exchange between the system and assessor. One of the initial goals of this Track was to represent and incorporate aspects of assessors' context into retrieval, thus, in addition to a corpus, topics and relevance judgments, this collection also contained user metadata describing context. The types of interactions that could occur were defined by the Track and there was no interactive searching performed by assessors. Instead, assessors completed these single-cycle interactions remotely (i.e., the assessors did not visit each participating sites' laboratories). In the first version of the Track, TREC assessors were also subjects — this was different from most iterations of the Interactive Track where assessors and subjects differed. In one iteration of

this Track, assessors represented a different kind of person than the traditional TREC assessor; they were interns and personnel at the Linguistic Data Consortium.

While the types of interactions were limited to those that could occur during a single exchange, the Track provided participants with an opportunity to engage in interactions with the assessors and elicit feedback from them. Common rules of the Track governed that some aspects of the interactions were the same across participating sites. To a certain extent assessors were also held constant since they each completed all the interaction forms for their particular topics. However, because the same assessor completed a number of interaction forms for a given topic, it was impossible to control for learning effects in this evaluation model — with each interaction assessors learned more and more about their topics, so while the Track studied single-cycle interactions, assessors actually engaged in multiple interactions (even if these were with different clarification forms). Although the evaluation model is not optimal for cross-site experimentation and full-scale IIR system evaluation, it can be used in a single-site study where researchers are interested in isolating and studying individual aspects of the search process while holding other aspects constant (e.g., [164]).

3.2.3 TREC ciQA Task

Following the HARD Track, ciQA (complex interactive question–answering) was introduced as a sub-task of the QA Track in 2006 and 2007 [68, 167]. The first year of this task was modeled closely after the HARD design where assessors completed forms that had been created and submitted by task participants. The ciQA task differed from the HARD task in that the focus was on complex question answering, rather than the traditional *ad-hoc* document retrieval task. The task also did not attempt to incorporate context into retrieval. At the time, there was little research devoted to interactive QA of the type represented by the TREC QA Track and one goal of ciQA was to encourage exploration in this area.

The 2007 version of ciQA allowed for any type of interaction, including full-scale interactions with working systems. Assessors interacted

with experimental systems remotely via the Internet. However, this new setup did not eliminate the problem of learning effects since assessors still engaged in a number of interactions with a number of systems for the same topic. More than anything, ciQA represented a first attempt at studying interactive QA and deploying a large scale evaluation exercise remotely. One of the more interesting findings of ciQA was the extent to which assessors could be considered as regular system users.

4

Approaches

This section provides an overview of different research approaches used for evaluation. Several approaches are discussed, but the emphasis in this paper is on laboratory evaluations.

4.1 Exploratory, Descriptive and Explanatory Studies

One way to think about research approaches is to consider specific goals of the research: exploration, description or explanation. Such characterizations can be found in almost any research methods textbook (e.g., [13]), but are useful to consider here since they suggest how studies should be evaluated. If the study goal is description, then evaluation criteria for explanatory studies should not be applied during the review.

Exploratory studies are typically conducted when little is known about a particular phenomenon. Exploratory studies often employ a variety of research methods with the goal of learning more about a phenomenon, rather than making specific predictions. Exploratory studies often have less structured methods than descriptive or explanatory studies and it is often the case that results from exploratory studies lead to descriptive or explanatory studies. Research questions are typically broad and open-ended and hypotheses are uncommon.

Descriptive studies are focused on documenting and describing a particular phenomenon. Examples of descriptive studies are those whose results characterize query logs and query behaviors (e.g., [146]). The main purpose of such studies is to provide benchmark descriptions and classifications. Although results of descriptive studies can become dated over time, temporal comparisons of results can made. As with exploratory studies, results of descriptive studies can be used to inform other studies. For instance, an analysis of a query log could give a researcher a principled way of selecting tasks to use in a laboratory study or suggest a hypothesis that can be evaluated as part of an explanatory study. Descriptive studies can lead to a weaker form of prediction via correlation analysis. However, such studies are not able to explain why a relationship exists between two variables.

Explanatory studies examine the relationship between two or more variables with the goal of prediction and explanation. Explanatory studies are often concerned with establishing causality and because of this require variables of interest to be isolated and studied systematically. Explanatory studies are often conducted in the laboratory since this is the environment that affords the researcher the most control over the situation. Explanatory studies use more structured and focused methods than exploratory or descriptive studies and involve hypothesis testing. Despite the name, it is important to note that not all explanatory studies offer explanations — many just report observations and statistics without offering any explanation. It is also important to distinguish between prediction and explanation: it is possible to build predictive models of events without actually understanding anything about why such events occur. Very often researchers stop at prediction and do not pursue explanation, but it is actually explanation that is tied most closely to theoretical development.

4.2 Evaluations and Experiments

In classic IR, experiment and evaluation have been used interchangeably, but these two types of studies need to be separated when discussing IIR. One can conduct an evaluation without conducting an experiment and *vice versa*. Evaluations are conducted to assess the

goodness[1] of a system, interface or interaction technique and can take many forms (some of which are discussed later). Experiments have historically been the main method for interactive system evaluation, but experiments can also be conducted to understand behavior. IIR experiments look similar to those conducted in social science disciplines such as psychology and education. For instance, it is common to evaluate the relationship between two or more systems and some set of outcome measures, such as performance or usability. This is a standard experimental model where the goal is to examine the effects of an independent variable (e.g., system type) on one or more dependent variables (e.g., performance and usability). Two important characteristics of experiments are that there are at least two things being compared (e.g., system type) and that some manipulation takes place. For instance, one might manipulate which system a subject uses.

In some types of IIR studies only a single system is evaluated. This is a weaker form of evaluation since it is not possible to demonstrate how much better users perform or how different their behaviors and interactions are since there is no point of comparison. Traditional usability tests are examples of this type of evaluation. Traditional usability tests are usually conducted with a single version of a system, with the goal of identifying potential usability problems. These types of studies are particularly important for *formative evaluation*: evaluation that is conducted during system development [97]. Formative evaluations can be contrasted with *summative evaluations* which assess the value of a finished or mature system.

4.3 Laboratory and Naturalistic Studies

Studies can also be characterized according to where they take place. Studies can take place in the laboratory or in a naturalistic setting. Most, but not all, experiments take place in the laboratory. It is important to note that if you are conducting a study in a lab, this does not automatically make the study an experiment. It is traditional to conduct other types of studies, such as usability tests, in labs even

[1] The term *goodness* is used as an abstract construct and may, of course, represent a number of things such as performance, usability or effectiveness.

when there are no experimental conditions or manipulations. Laboratory studies are good with respect to the amount of control researchers have over the study situation. This is particularly useful when trying to isolate the impact of one or more variables. Of course, one perennial criticism of laboratory studies is that they are too artificial, do not represent real life and have limited generalizability.

Naturalistic studies examine IIR in the settings in which it occurs. Log-based studies are examples of naturalistic studies since they capture behavior as it is occurs in real life. The behavior that is captured is thought to be more representative of the user's true behavior since the chances that it contaminated or biased by the study design or the researcher is much less than that captured in a laboratory. One important drawback to conducting naturalistic studies is that the researcher has little control over the setting, which can make it hard to make cross-user comparisons. The amount and types of data one collects might be as variable as the number of users in the study. It can also be difficult to administer naturalistic studies since they are more intrusive and the user often has to be willing to give up some privacy.

It is also possible to conduct natural experiments. One example is the study by Anick [11] who conducted live trials of an interface for query expansion. As a researcher at a large search engine company, Anick was able to distribute an experimental interface to a number of users and compare its use to the standard interface. In another example, Dumais et al. [79] deployed two working versions of a desktop search tool to 234 people within an organization and gathered data on its use over the course of six weeks.

4.4 Longitudinal Studies

Naturalistic studies are often longitudinal in that they take place over an extended period of time and measurements are taken at fixed intervals. Longitudinal approaches can also be incorporated into laboratory studies — users might be required to attend multiple sessions over time. Longitudinal approaches are often used when one wants to study if and how something changes over time. Although studies employing longitudinal approaches are more time consuming, they represent a necessary

and important type of study since many kinds of information seeking and retrieval activities take place over extended periods of time during multiple search sessions [181, 247]. Only studying single search sessions, such as the kind typically studied in the lab, places limits on what is known about IIR.

One important consideration in designing longitudinal studies is determining the duration of the study, as well as the measurement interval. Having some understanding of the behavior and some expectation about how often it occurs can help one make this decision. For instance, if the behavior occurs everyday then one might want a different duration and measurement interval than if the behavior occurs weekly. It is also the case that in the social world, user behavior can be governed by a number of external factors. For instance, the occurrence of a holiday or a project deadline will likely change the kinds of behaviors users exhibit and these behaviors may not represent their typical behaviors.

4.5 Case Studies

Another approach is the case study. This type of study is not seen a lot in IIR, but as the area grows, researchers may begin to do more of these studies. Case studies typically consist of the *intensive* study of a small number of cases. A case may be a user, a system or an organization. Case studies usually take place in naturalistic settings and involve some longitudinal elements. Researchers conducting case studies are less interested in generalizing their research results and more interested in gaining an in-depth, holistic, and detailed view of a particular case. The ability to generalize is traded for a more complete and robust representation of how something occurs in a small number of cases. Case studies are particularly useful when little is known about an area and for understanding more about details that sometimes get lost when averaging over large numbers of users.

4.6 Wizard of Oz Studies and Simulations

Wizard of Oz studies get their name from the well-known film/book of the same title. In this work, the protagonist Dorothy, travels a

great distance to visit the Wizard of Oz who at first glance appears very grand and intimidating. However, as it turns out, the Wizard is really just a small man behind a curtain orchestrating a grand Wizard façade. Wizard of Oz studies are similar in that researchers often imitate 'grand' systems that they would like to study. While users believe they are interacting with a real system, in reality there are one or more researchers 'behind the curtain' making things work. For example, suppose a researcher wanted to study a speech user interface for querying and interacting with an IR system. Rather than building the entire system, the researcher might first want to learn something about the range of desired communications and interactions. Users might be instructed to speak to the system while a researcher sits in another room and controls the system. Wizard of Oz studies can be used for proof-of-concept and to provide an indication of what might happen in ideal circumstances [67].

Wizard of Oz studies are simulations. Heine [126] discusses simulation experiments in the context of IR research. While systems are simulated in Wizard of Oz studies, another entity that has been simulated in IR studies is users. Rather than recruit and study real users, simulated users are used to exercise systems. Users may be defined by one or more characteristics which can take on a number of values. Simulated users consist of various combinations of these characteristics and values. Simulated users can also represent different actions or steps a real user might take while interacting with an IR system. Some concerns about the use of simulated users include the realism and utility of the simulated users and more fundamental issues about what it means to really study users and user behavior. Some positive things about simulated users is that IIR evaluations can be conducted more rapidly and with less cost and larger numbers of users (albeit simulated), who have a broad range of carefully controlled characteristics. It is also the case that one can control learning through programming — the researcher is able to determine if and how learning will take place during the study. Example studies of simulated users include Lin and Smucker [180] and White et al. [293].

5

Research Basics

This section discusses some of the basics of IIR research. Most of these things are necessary parts of any empirical research, but they are reviewed here because they form the foundation of research endeavors.

5.1 Problems and Questions

All studies are motivated by some problem or gap that exists in the research. Thus, the first step in conducting any study is to identify and describe the problem. This helps focus one's attention and provides a roadmap for the presentation of research results. Describing the problem and its relevant pieces also helps ensure that proper attention has been paid to what is currently known about a particular issue (and that a proper literature review has been conducted). Within the context of some given problem, the research question essentially identifies the piece of the problem that will be addressed by the study.

The research question[1] should be narrow and specific enough that it can be addressed in a study, but the specificity of the question will

[1] Although research question is used in the singular, it is common for studies to have multiple related research questions. It is advised to have multiple, simple questions rather than a single, complex question.

depend in part on the purpose of the study. For instance, explanatory studies typically have much more concise and narrow research questions than exploratory studies. Research questions should also be value-free in a sense that the researcher's opinion is not embedded in the question. If the researcher has some belief about the outcome of the study, then this should be framed as part of the study hypothesis, not research question.

Figure 5.1 shows some example research questions from IIR studies. The first example identifies a broad question that was part of an exploratory study. Example 2 is a descriptive question, while the final two examples are of specific, focused explanatory research questions. Since description often leads to explanation, some studies might have both a descriptive and explanatory research question. What is known about a particular phenomenon and the extent to which one wants to study it determines the specificity with which questions can be asked.

Researchers who are very focused on their particular system might also be tempted to pose a question in the form, "Is System X better than System Y?" This is probably okay (even though it could technically be answered with a binary response), but in some ways detracts from more specific types of questions that can be asked regarding the differences between System X and System Y. A better approach would be to identify the expected differences (e.g., in performance or usability) and formulate specific questions about these differences, rather than to lump everything together in a single question. The general question might be fine for a strict evaluation, but more specific questions are

Example 1: How do people re-find information on the Web? [268]
Example 2: What Web browser functionalities are currently being used during web-based information-seeking tasks? [163]
Example 3: What are the differences between written and spoken queries in terms of their retrieval characteristics and performance outcomes? [62]
Example 4: What is the relationship between query box size and query length? What is the relationship between query length and performance? [22, 159]

Fig. 5.1 Some example research questions from IIR studies. Example 1 is exploratory, Example 2 is descriptive, and Examples 3 and 4 are explanatory.

better suited for situations where the researcher wants to understand IIR behaviors or phenomenon that transcend a specific system.

5.2 Theory

A theory is "a system of logical principles that attempts to explain relations among natural, observable phenomena. A theory appears in abstract, general terms and generates more specific hypotheses (testable propositions)" [95, p. 132]. Littlejohn [182, p. 21] describes theory as "any conceptual representation or explanation of a phenomenon." Thus, two of the most important qualities of theories are that they offer explanations of particular phenomena and allow researchers to generate hypotheses. The testing of hypotheses, in turn, allows researchers to further refine and extend theories. Theories can also be developed and refined through grounded theoretical approaches [108].

Historically, IR and IIR have been driven by innovation and technology; the emphasis has been on applied and practical aspects of science. As a result, there has been less in the way of theoretical development. Indeed, a lot of research does not even mention or consider theory. This is not to claim that there are no theories or theoretical constructs in IR and IIR — some examples include Robertson's [218] probability ranking principle, Ingwersen's [137] theory of polyrepresentation and Belkin's [19] anomalous states of knowledge. In the area of human information behavior there is even a book describing common theories and models [93]. What is claimed is that at present, theory innovation has received less attention than system innovation. Currently, research in IR and IIR emphasizes results over explanation, and many studies are not motivated theoretically.

5.3 Hypotheses

Hypotheses follow from research questions (or theory) and state expected relationships between the concepts identified in the questions (such concepts may be more or less definable, but they are eventually represented by variables). There are two types of hypotheses: *alternative* hypotheses and *null* hypotheses. An alternative hypothesis is

the researcher's statement about the expected relationship between the concepts under study. This is also known as the research hypothesis. Research hypotheses are called *alternative* because they present alternatives to the null hypothesis, which states that there is no relationship or difference. The null hypothesis is accepted by default; the scientific method places the burden on the researcher to demonstrate that a relationship exists. Although there are several published accounts of researchers proposing and testing null hypotheses, this is actually counter to the scientific method since null hypotheses do not need testing — they represent the default description of things. Scientists start with null hypotheses because logically it is easier to show that something is false instead of true.

In general, science is about accumulating evidence to demonstrate some relationship rather than providing definitive proof. The logic is such that we are not ever able to say that our alternative hypothesis represents the *truth* or that we have *proved* it. In fact, it can be argued that it is not useful to talk about truths, especially when studying the social world, but rather to talk about accumulation of evidence, which supports a particular hypothesis or points in a general direction. When we engage in hypothesis testing, strictly speaking we are able to make two statements about the relationship between the evidence we collect and our hypothesis: (1) our evidence allows us to *reject the null hypothesis*, in which case it is shown that our hypothesis provides a better (but not the only) description of what is going on, or (2) our evidence does not allow us to reject the null hypothesis, in which case *we fail to reject the null*. The burden of proof lies with the researcher, and even then, absolute proof is not a useful construct. Instead, researchers show that the evidence they collect demonstrates that the null hypothesis does not adequately describe what is going on; the alternative hypothesis offers an alternative explanation of what is going on, but it still may not adequately capture what is happening.

At the most basic level, a hypothesis should state a relationship between two or more things. One common mistake that many researchers make is not actually posing testable hypotheses or not fully articulating the comparison they would like to make. By nature, hypotheses are comparative and suggest the existence of at least two

things. For instance, it is not sufficient to say, "System A is usable," or that "System A is more usable," but "System A is more usable than System B." While one could provide descriptive statistics that show, for instance, that 70% of subjects rate System A as usable, without at least two groups, one cannot perform any statistical testing. There is no way to test the first statement, particularly since there are no benchmarks upon which to base rejection of the null. In the statistics section of this paper, it will be shown that it is possible to have a hypothesis that does not explicitly name two groups, but this is when one group is the population and the population parameter is known.

Hypotheses can also be directional or non-directional. The hypothesis that *System A is more usable than System B* is a directional hypothesis since the direction of the difference is given. Contrast this with a non-directional form of this hypothesis, *there is a difference in usability between System A and System B*. Finally, strictly speaking, hypotheses should be stated at or near the beginning of a study. Researchers often create hypotheses after they begin examining their data, but the scientific method calls for hypotheses to be identified clearly before any data are collected.

5.4 Variables and Measurement

Variables are present in almost all studies, although they play less of a role in qualitative studies. Variables represent concepts. Specifically they represent ways of defining, observing and measuring the concepts that researchers aim to study. Relevance, performance, and satisfaction are all concepts. To investigate concepts, researchers must engage in two basic processes: conceptualization and operationalization. These processes involve articulating definitions, but at two different levels.

5.4.1 Conceptualization and Operationalization

Conceptualization is the process by which researchers specify what they mean by particular terms. Some terms have very agreed-upon meanings. For instance, if we talk about someone's sex, most people would understand what we mean. However, other terms can have a variety of meanings. For instance, there is no universally agreed-upon

definition of relevance. We know from studies that have attempted to define relevance that there are many interpretations of this term, as well as many manifestations [235, 236]. Thus, the first step in attempting to measure a concept like relevance is to define it. Such a definition is considered a *working* or*nominal* definition — it represents a *temporary* commitment on the part of the researcher and helps frame the study and delineate findings. No claim is made about the universality of the definition.

Sometimes it is useful to subdivide a concept into dimensions to make conceptualization easier. For instance, in defining relevance, one might first identify specific dimensions such as those articulated by Saracevic [234] — algorithmic, topical, cognitive, situational, and affective. Next, one might provide specific definitions for these terms rather than trying to provide a single, all-encompassing definition for relevance.

After articulating conceptual definitions, the next step is to provide *operational* definitions, which state the precise way the concept will be measured. For instance, one might decide to measure topical relevance by asking subjects to indicate how useful they find documents and giving them a five-point scale to indicate this. One might define algorithmic relevance as the system's estimate of the likelihood that a user will find a document useful given a particular query and use the relevance scores produced by the retrieval system as an indicator. Usability, a concept that plays an important role in many IIR evaluations is often subdivided into dimensions such as effectiveness, efficiency, and satisfaction. Doing this is only part of the process since one must also define these concepts and state how they will be measured and observed. Very often researchers do not carefully articulate the conceptual and operational definitions that they employ in their reports; if this has not been done, it is difficult to evaluate the quality and appropriateness of the measures, and the validity of the work. This also makes it difficult to compare findings across studies.

5.4.2 Direct and Indirect Observables

A useful distinction to make between IIR measures relates to whether they are directly or indirectly observed. Direct observables are often

byproducts of a user's behaviors and interactions and are produced as the user searches. For instance, number of queries entered, number of documents opened, and the amount of time spent searching are examples of measures that are directly observable. Indirect observables are those things which cannot be observed and that essentially exist within the user's head. An example of an indirect observable is satisfaction.

For both types of observables — direct and indirect — it is important to ensure that the equipment used to make the observations is valid and reliable. Direct and indirect observables present different issues in this regard. With direct observables there is a ground-truth — we can physically count the number of queries a user enters and compare this value to what is recorded by some other instrument, such as a logger. However, for concepts that are only indirectly observable, instrumentation is more difficult. Not only must researchers be concerned with whether indirect measures are good representations of particular concepts, but they must also be concerned with how this information is captured (e.g., does a five-point Likert-type item adequately capture satisfaction?). Ground-truth exists in each individual user's head and there is no way to compare what we can observe through a self-report scale to this truth.

IIR is concerned with many more indirect observables than direct observables, which would suggest that measurement and instrumentation should be priority research issues. Unfortunately, there are not a lot of research programs focused on measurement, which makes it difficult to understand the extent of measurement problems in IIR evaluations. Many measures are developed in an *ad-hoc* fashion and there are few well-established measures and instruments, especially for indirect observables. Ultimately, any new measure should be both valid and reliable. These issues are discussed later.

5.4.3 Independent, Dependent, and Confounding Variables

Another distinction that can be made is between *independent*, *quasi-independent*, and *dependent* variables. Using the language of cause and effect, independent variables are the causes and dependent variables are the effects. In experiments, researchers typically manipulate the

independent variable — for instance, asking people to use particular systems or assigning them to particular experimental conditions. Quasi-independent variables are variables that can create differences in some outcome measure, but are not manipulated by the researcher. Sex is a good example of a quasi-independent variable. A researcher might be interested in examining differences in how males and females use an experimental and baseline IIR system, but a researcher cannot manipulate anyone's sex. The researcher might ask equal numbers of males and females to use each system, but this variable is not under the researcher's control in the same way as system type. Dependent variables are outcome variables, such as performance and satisfaction. In most IIR evaluations, researchers are generally interested in examining how differences in one or more independent variables impact one or more dependent variables.

Confounding variables (or confounds) are variables that affect the independent or dependent variable, but have not been controlled by the researcher. Often researchers are unaware that such variables exist until they begin a study or after a study ends. If a researcher realizes that such variables exist before the study starts, then the researcher can control the effects of the variables. For instance, a researcher might believe that search experience impacts how successful a person will be with an information retrieval system. If the researcher were testing two IIR systems, it would be important to ensure that equal numbers of subjects with high and low search experience were assigned to use each of the systems. If more subjects with high search experience were assigned to use one of the systems, then it might be found that subjects did better with this system, but the cause could not be attributed to the system since another potential explanation exists. In this case, search experience would be considered a confounding variable.

In addition to independent and dependent variables, moderating and intervening variables are also used to represent relationships among concepts. However, at present, these are used less frequently in IIR evaluations. Moderating variables affect the direction or strength of the relationship between an independent and dependent variable. For instance, consider the above example regarding the relationship between system, search experience and performance. Suppose the

researcher designed two interfaces, one which supported advanced search and another which supported simple search. The researcher might be interested in investigating how well subjects perform with each of the systems. The researcher might further believe that subjects who have high search experience will perform better with the advanced system, while those with low search experience will perform better with the simple system. In this situation, search experience is said to moderate the relationship between interface type and performance because different levels of search experience (high and low) result in different types of relationships between the independent and dependent variables.

Finally, an intervening variable provides a connection or link between an independent and dependent variable. For instance, a larger query box might lead subjects to enter longer queries. These longer queries, in turn, might lead to better performance. The connection between the independent and intervening variables and intervening and dependent variables basically represents two causal relationships. One would not say that a larger query box caused differences in performance, but rather that a larger query box caused differences in query length which in turn, caused differences in performance.

5.5 Measurement Considerations

When designing measures, there are a number of properties to consider. Most of these properties are related to measures that have a response set. Such response sets might contain numeric or textual choices or categories for responding.

The first property is related to the *range of variation* that is expected to occur in the concept being measured. The range of variation is the extent to which a measure presents an adequate number of categories with which to respond. Range of variation is closely related to the preciseness of the measure. For instance, when creating an instrument for eliciting relevance judgments from subjects, is a binary scale, a tertiary scale, or a five-point scale provided?

Exhaustiveness is closely related to range of variation and is the extent to which a response set can be used to characterize all elements

under study, where an element might be a system, document or user. For instance, with a binary relevance scale, a user might have a difficult time characterizing a document that is partially relevant. Another common example, although not used in most IIR evaluations, is race. Very often some races are missing from the list of choices (exhaustiveness) and others are not differentiated enough (range of variation). Understanding both the range of variation and how exhaustive measures need to be is very much related to the researcher's knowledge and expectations of the typical variation that exists within the elements under study.

Another property to consider is *exclusiveness*. This property is the extent to which items in the response set overlap. When this property has been violated, there might be more than one response that can be used to characterize a single object. For instance, a user might be provided with the following options for indicating relevance: not relevant, partially relevant, somewhat relevant and relevant. Most subjects would have a difficult time distinguishing between the middle two options (unless the researcher provided some very good definitions of each choice). Not all measures have to be exclusive, for instance, some questions allow a user to make more than one choice, but such questions can technically be viewed as a series of binary items.

Another property is *equivalence*. This property is the extent to which the items in a response set are of the same type and at the same level of specificity. Consider a scale that is meant to assess a person's familiarity with a search topic and has at one end of the scale the label *very unfamiliar* and at the other end, *I know details*. It would better to associate the first label with *very familiar*, and the second label with *I know nothing* since these are true opposites and at the same level (*knowing details* is slightly more specific than *being familiar*).

A final property is *appropriateness*. This is the extent to which the provided response set makes sense in relation to the question being asked. Consider the following question which might be asked of subjects, "How likely are you to recommend this system to others?" If the researcher provided subjects with a five-point scale with strongly agree and strongly disagree as anchors, then this response set would be inappropriate because the scale anchors do not match the question.

5.6 Levels of Measurement

The level of measurement used to represent a variable is a critical concept since it ultimately determines what types of statistical tests are possible. Researchers often wonder what type of statistical test is most appropriate for their data. The answer to this question, in part, lies with the levels of measurement. There are two basic levels of measurement: *discrete* and *continuous*. One of the biggest differences between discrete and continuous measures is the extent to which the values represent real numbers. This, in turn, impacts the extent to which data produced from these measures can be used in mathematical functions. As one moves from discrete to continuous levels of measurement, one is able to conduct more sophisticated types of statistical analyses.

5.6.1 Discrete Measures

Discrete measures provide and elicit categorical responses. These categorical responses can be textual or numeric. Discrete measures are divided into *nominal* and *ordinal* data types. Nominal data types provide response choices that represent different kinds of things but not different degrees. Ordinal measures provide response choices that are ordered, where choices represent different degrees. A classic example of a nominal measure is *sex*, which has two *levels* (choices or responses): male and female. These two levels are different from one another, but there is no order among them — one is not better or more than the other. Instead they represent the exhaustive set of choices that all subjects would need to be able to classify themselves. The most common type of nominal variable in IIR evaluations are independent variables such as interface type and task-type.

The two common ways that ordinal measures are used are as rank-order measures and as Likert-type[2] scale measures. An example: a rank-order measure is when a subject is given a set of documents and asked to order them from most relevant to least

[2] Likert-type is used to describe numeric scales (regardless of points) that are used to elicit data from users. The term *Likert-type* is used because the original Likert scale was a five-point scale that measured agreement and was scored and administered a bit differently than how such scales are often used today [178].

relevant. This type of measure allows the user to show the perceived relationship among the documents with respect to relevance, but this is a relative measure rather than absolute. For instance, we could identify which documents were more relevant than others, but we could not discuss the magnitude of these differences. That is, we could not say how much more relevant one document was than another. The difference between documents ranked 1 and 2 might be slight, while the difference between documents ranked 2 and 3 might be large. The differences between consecutive points are not equal.

The other common type of ordinal measure in IIR is Likert-type scales. While such scales give the impression of a true number line, the values do not represent real numbers (instead, one can think of the numeric values as labels). If we provided subjects with a five-point scale, where 1 = not relevant and 5 = relevant, and asked them to judge a set of documents, then we would still be in a position to describe which documents were more relevant than others, but we would be unable to describe the amount of these differences. We could say that a document rated 4 was more relevant than a document rated 2, but we could not say that a document rated 4 was *twice* as relevant as a document rated 2 since the scale contains no true zero.

To use a Likert-type scale subjects have to perform some calibration. This is unlikely to be consistent across subjects or even within a subject: one subject's 2 may not represent the same thing internally as another subject's 2. Because Likert-type scales are not represented by real numbers, the types of analyses that can be done with them are limited. This is unfortunate since these measures are the *sine qua non* of any science whose aim is to study human behavior and attitudes. Because of this, an accepted practice in the social sciences is to promote Likert-type measures to a continuous data type so that more sophisticated analyses can be done.

5.6.2 Continuous Measures

Continuous measures are divided into *interval* and *ratio* data types. For each of these types, differences between consecutive points are equal, but there is no true zero for interval scales. The most common examples

given for *interval* level data are the Fahrenheit temperature scale and intelligence quotient (IQ) test scores. For both measures, a score of zero does not indicate the complete absence of heat (indeed, the freezing point is 32 degrees) or intelligence. However, it is the case that the absolute differences between temperatures of 40 and 80 degrees, and 50 and 90 degrees are the same. However, it is not appropriate to say that a temperature of 80 degrees is twice as warm as a temperature of 40 degrees since there is no true zero on the scale.

The *ratio* level of measurement represents the highest level of measurement. A true number line underlies measures of this type. Common examples of ratio levels of measurement include time and almost any measure that can be verbally described as *the number of occurrences*. For instance, the number of queries issued, the number of pages viewed, and the number of documents saved. It is sometimes difficult to imagine these values being zero, but it is possible. For example, it is *possible* for someone not to enter a single manual query (imagine a system that supports browsing) or not open any documents during a search session. Another nice thing about ratio level data is that it can be transformed into ordinal and nominal level data. For instance, based on recall scores, an ordinal measure called *search expertise*, with the following levels: poor performers, medium performers and top performers could be created.

6

Experimental Design

The basic experimental design in IIR evaluation examines the relationship between two or more systems or interfaces (independent variable) on some set of outcome measures (dependent variables). IIR evaluations can include other independent variables as well such as task-type, and quasi-independent variables such as sex and search experience. One important part of experimental design which will be discussed in detail is rotation and counterbalancing. Tague-Sutcliffe [261] was one of the first to write formally about this in IIR. This allows one to control aspects of the study that might otherwise introduce experimental confounds. This section also presents other issues related to experimental design including study mode, protocols, tutorials, timing and fatigue, and pilot testing.

6.1 Traditional Designs and the IIR Design

Traditional designs can be discussed in terms of pre-experimental designs and experimental designs. These are standard designs that are discussed and presented in a number of research methods textbooks (e.g., [13]). They are not a creation of IIR and do not always fit perfectly with IIR study situations, but they do provide different

Group	Time_1		Time_2
1	O	E	O
2	O	C	O
3		E	O
4		C	O

Fig. 6.1 Solomon four-group experimental design.

ways of thinking about study design and measurement. The distinction between pre-experimental and experimental designs rests on the absence of a control group and baseline measurement. Figure 6.1 presents a well-known experimental design, the Solomon four-group design [47]. The different groups in this design can be used to illustrate other types of research design, including pre-experimental designs. A pre-experimental design with no control group or baseline measurement is represented by Group 3. In this group, an experimental stimulus (E) (e.g., a system) is introduced and then an observation or measurement (O) is taken of some outcome measure (e.g., performance). One of the most common types of studies in IIR that follow the design depicted by Group 3 is the single system usability test. There is no comparison or control system. Instead, subjects use one system and some initial feedback is collected regarding its goodness. Note that this type of study does not allow for the testing of hypotheses related to the system because there is only one system being studied. No comparison is possible, except with pre-determined population parameters, which are unlikely to exist. It is important that one looks closely at one system studies before deciding they are usability studies and not experiments. Many experiments only involve a single system, but some other variable of interest is manipulated and of interest. It is possible for IIR evaluations to have independent variables that are not tied directly to a system. The system may just be used as an instrument to facilitate information search (e.g., [168]).

The other attribute that makes this (Group 3 only) a pre-experimental design is that a baseline measure of the outcome variable has not been taken. In traditional experimental models, baseline measures of the outcome variables of interest are elicited before the stimulus is introduced. This is depicted by Group 1 in Figure 6.1. For instance, if one were evaluating a new drug designed to help people lose

weight and the outcome measure was a person's weight, one would need to obtain a baseline weight for each subject to know if the drug was associated with a decrease in weight — without this measure it would not be possible to determine this. In the context of IIR, one general goal of many evaluations is to determine if a particular system helps subjects find relevant documents. Attempting to elicit a baseline measure before the system (stimulus) is introduced does not make much sense and would probably not be possible. We can also imagine that the goal of an IIR system is to help subjects learn something about their information problems. To evaluate this, we would really need to measure subjects' knowledge of their information problems before and after they used the system.

Baselines are used in IIR evaluations, but in a way that differs slightly from the classic experimental model. In the context of IIR evaluations, baselines are often introduced as an alternative to the experimental system. Instead of taking a baseline measure before a user interacts with a stimulus, the baseline is more often represented by one level of the stimulus variable. For example, if the stimulus variable is an IIR system, it might have two levels: experimental and baseline. Thus, baselines in IIR evaluations are more similar to control groups (C) in Figure 6.1. In the *traditional experimental model*, the stimulus variable is usually either present or absent and a control group is used along with pre-treatment measurement (Figure 6.1, Groups 1 and 2). In Figure 6.1, the *classic IIR* design is represented by Groups 3 and 4. This model ostensibly functions as the archetypical IIR evaluation design.

A baseline (or control in the traditional model) is generally defined as the status quo, which raises some interesting questions with respect to IIR evaluations. Specifically, if IIR systems are under study and baselines represent subjects' normal experiences, then in most cases this would be a commercial search engine. However, it is not possible or valid to compare an experimental IIR system to a commercial search engine.[1] For instance, a researcher may be using a closed collection of newspaper articles; if a commercial search appliance were used to access

[1] This may be possible if you work for a commercial search engine company.

this collection as a baseline, it might not work optimally because of characteristics of the corpus and search algorithm. Thus, such an evaluation would not be comparing similar situations. Of course, whether a commercial search engine is a valid baseline depends greatly on the purposes of the study and the system. Even though it may not be possible or desirable to use a commercial search engine as a baseline, it is important to recognize subjects' previous search experiences and search norms will impact their interactions with, and expectations of, any experimental IIR system.

Developing a valid baseline in IIR evaluations often involves identifying and blending the status quo and the experimental system. For instance, if a researcher developed a new technique for displaying search results, then a baseline method of doing this could be modeled after methods used by commercial search engines. If the experiment was done using a proprietary system or well-established system, then the baseline could be the retrieval method currently used by that system (given that one was testing the workings of the system). Things get a bit more difficult when the experimental system or interface is something that subjects have never seen. Researchers often develop experimental IIR systems from scratch using languages such as Java. There is a good chance that the interface will look very different from a Web-based system to which subjects are accustomed. In this case, if one were comparing a new search interface feature, it would not be reasonable to compare this to a standard Web search engine since the number of differences between these two systems would be great. If differences were found, it would be difficult to relate them to the specific search interface feature of interest and to rule out the possibility that these differences were not caused by some other feature or aspect of the system.

As mentioned earlier, the design depicted in Figure 6.1 is called the *Solomon Four-Group Design* [47]. It was developed to address several major threats to the internal validity of experiments. These will not be discussed here, but suffice to say the four groups allow the researcher to control a number of threats to validity. The Solomon Four-Group Design is quite nice, but requires large numbers of subjects, since the groups are independent. Many researchers in other disciplines use the classic experimental design (Groups 1 and 2 only), while others

(IIR included) use a modified design based on Groups 3 and 4. This design is called a *Posttest-only Control Group Design* [47]. Campbell and Stanley [47] argue that these are the only two groups needed, if subjects have been randomly assigned to the groups.

All of these designs rest of the assumption that subjects comprising each of the groups are equal across a range of characteristics. Characteristics which might, if not equally distributed across groups, conspire to generate spurious results — results not caused by the stimulus, but by some other characteristic of the group. Random assignment can be used to increase the likelihood that these characteristics are distributed equally across groups. While it is usually not possible to conduct true random sampling in IIR evaluations, it is possible to randomly assign subjects to groups (or conditions).

6.2 Factorial Designs

Currently, the more common way for researchers to discuss experimental design in IIR is as *factorial designs*. This is particularly useful when studying the impact of more than one stimulus or variable. The models presented above assume a binary stimulus (experimental and control), but the researcher might also be interested in studying the impact of a number of factors[2] on one or more outcome variables. Factorial designs accommodate this. In the preceding example there was one factor, system type, which had two levels, experimental and baseline. If the researcher believed that there might also be differences in the outcome variable based on the sex of subjects, then sex would be an additional factor, with two levels, male and female. This is tightly coupled with the previous levels of measurement discussion; the factors in a factorial design should be discrete. The levels represent distinct categories rather than ratio level values.

There is a specific notation and language for describing factorial designs. If the relationship between the two factors mentioned above were examined (system type and sex) in relation to an outcome measure such as performance, then the experiment is described as a 2×2 factorial

[2] Used as a synonym for independent variable.

2×2 Factorial

		System		
		A	B	Total
Sex	Males			
	Females			
Total				TOTAL

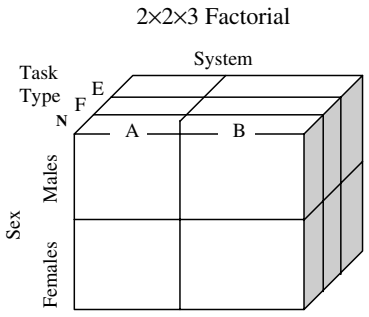


Fig. 6.2 Example factorial designs. The first example is a 2×2 design with one independent variable, system and one quasi-independent variable, sex. The second example is a 2×2×3 design with one additional independent variable, task-type, which has three levels: navigational (n), fact-finding (f), and exploratory (e).

design. Both the number of digits and their magnitudes are meaningful. The number of digits describe the number of factors (system type and sex) and the magnitude of each number describes the number of levels of each factor (experimental, baseline; male, female). If another factor called task-type were added which had three levels (exploratory, fact-finding, and navigational), the experiment would be described as a 2×2×3 factorial. These two designs are illustrated in Figure 6.2. Such illustrations aid in describing a study and allow researchers to understand and communicate the different types of comparisons that are available. The different combinations of levels generate different conditions (in the 2×2 there are four conditions and in the 2×2×3 there are 12 conditions). Each condition will have some value on one or more outcome variables. Comparisons can be made using cell, column and row values. Each factor adds another dimension to the representation and studies with four or more factors do not lend themselves as easily to this type of representation and are not conducted that often anyway because they require large numbers of subjects and it is difficult to interpret results.

6.3 Between- and Within-Subjects Designs

Studies can also be characterized with respect to whether the independent variables are *between* or *within* subjects. This is an important

distinction which should be made in all reports. Between-subjects means that subjects experience only one level of the variable, while within-subjects means that subjects experience all levels of the variable. Studies can be mixed along this characterization: some variables can be between-subjects, while others can be within-subjects. For instance, system type might be a between-subjects variable, while task-type might be a within-subjects variable. This means that each subject would only use one system, but would have to complete all three task-types. Some variables are necessarily between- or within-subjects. For instance, values on the sex variable are completely beyond the control of the experimenter and reside outside of the study. In the classic IIR evaluation study, system type (or interface type) is typically within-subjects to facilitate comparison of the two systems. Otherwise, it is not possible to ask subjects to compare the systems since they would have only used one of them. In other cases, it might be desirable to make a variable between-subjects to avoid exposing subjects to all conditions of the experiment, which might led to contamination.

6.4 Rotation and Counterbalancing

Rotation and counterbalancing are cornerstones of most experiments and evaluations and are often associated with systems and tasks in IIR evaluations [260]. The primary purpose of rotation and counterbalancing is to control for order effects and to increase the chance that results can be attributed to the experimental treatments and conditions. Although *treatment* typically refers to the things that a researcher tests or manipulates (e.g., interface), it also refers to the tasks and topics which subjects execute when engaging in IIR regardless of whether these items are variables of interest. In most IIR evaluations that involve searching, search tasks are necessary in order for subjects to exercise systems. Even though they may not be treated by the researcher as independent variables, they do function as variables and therefore must be controlled. This is typically achieved through rotation.

Two types of designs can be used to systematically rotate variables, the *Latin square design* and the *Graeco-Latin square design* (Graeco is

also spelled Greco). The Latin square design accommodates a single variable, while the Graeco-Latin square design can accommodate multiple variables — it is essentially a combination of two or more Latin squares. To illustrate the different designs, let us assume that we are testing three interfaces using six different search topics and that each user will complete two topics per interface. The task will be held constant and will be a document finding task.

6.4.1 A Basic Design

First, let us look at a basic design with no rotation (Figure 6.3), where rows represent subjects and columns represent interfaces. Topics are represented in the cells of the table. There are two major problems with this design — the first is related to topic order and the second is related to interface order. A Latin square can be used to control for one of these things, but not both. A Graeco-Latin square is needed to control both.

The major experimental confounds that are introduced by this design are caused by order effects. Specifically, learning and fatigue can produce results that are attributable to the experimental design rather than the treatments. As subjects complete each consecutive topic, they learn more about the experimental situation and the experimental system (assuming that the systems are similar). With each topic encountered, subjects potentially transfer what they learn by completing one topic to their interactions with the next topic, which might result in better performance on topics that are presented last, as opposed to first. Since the order of interfaces is fixed, it is also likely that the subject's

	Interface type		
Subjects	Interface 1	Interface 2	Interface 3
S1	1, 2	3, 4	5, 6
S2	1, 2	3, 4	5, 6
S3	1, 2	3, 4	5, 6
S4	1, 2	3, 4	5, 6
S5	1, 2	3, 4	5, 6
S6	1, 2	3, 4	5, 6

Fig. 6.3 A basic design with no rotation. Numbers in cells represent different topics.

earlier experiences will impact later experiences. If, on average, subjects perform best with the last interface, it may be a function of a learning effect, rather than the goodness of the last interface. Another problem with a fixed order for topics and interfaces is related to the potential interactions among the topics and the system. Some topics may be easier than others and some systems may do better with some topics than others. If, for example, it was found that Interface 2 was the best, it may be because Topics 3 and 4 were easier than the other topics. In this case, while the researcher may attribute differences to the interfaces, the differences are really caused by the topics.

Fatigue can also impact the results. As subjects engage in more and more searches, they are likely to become fatigued, especially in experiments that last over one hour. At the beginning of the study, subjects may be more motivated and attentive than at the end of the study. When subjects become fatigued they may move quickly through the experiment just to finish. They may become cognitively exhausted and just not be able to perform as well as they did at the start of the study. If, for example, it was found that Interface 1 was the best, it may be because subjects were more energized and worked harder in the beginning of the study than at the end.

6.4.2 A Latin Square Design

To improve the design in Figure 6.3, a *Latin square* can be used to control for the effects of one of the variables — either topic or interface. Latin square designs are used to rotate and control for a *single* variable and in Figure 6.4 this variable is topic. The items in the cells of a Latin square are distinct and should appear an equal number of times in each row and each column. This can be accomplished fairly easily: for each row, topics are shifted among the columns in a systematic way. Since there are six topics, we need six rows (or six subjects) to get through one topic rotation. Note that each user completes all topics. It is also important to note that these designs do not eliminate learning or fatigue, but distribute their impact equally across all treatments and conditions.

While the rotation in Figure 6.4 is balanced with respect to topics, it is problematic for other reasons. There are two important things that

	Interface type		
Subjects	Interface 1	Interface 2	Interface 3
S1	1, 2	3, 4	5, 6
S2	2, 3	4, 5	6, 1
S3	3, 4	5, 6	1, 2
S4	4, 5	6, 1	2, 3
S5	5, 6	1, 2	3, 4
S6	6, 1	2, 3	4, 5

Fig. 6.4 Basic design with Latin square rotation of topics.

	Interface type		
Subjects	Interface 1	Interface 2	Interface 3
S1	2, 4	1, 6	5, 3
S2	3, 5	2, 1	6, 4
S3	4, 6	3, 2	1, 5
S4	5, 1	4, 3	2, 6
S5	6, 2	5, 4	3, 1
S6	1, 3	6, 5	4, 2

Fig. 6.5 A basic design with Latin square rotation of topics and randomization of columns.

this type of rotation does not address. First, it is possible that there may be some interaction among the topics, such that encountering Topic 4 after Topic 3 makes completing Topic 4 easier. Note that for all rows of the table except one, Topic 4 always follows Topic 3. One can see this visually in the design via the diagonal — this indicates that there is still some order preserved in the table (it is easiest to spot this along the '6' diagonal). One way to address this problem is to randomize the order of the columns (excluding the column headings). One could assign numbers to each of the columns and then use a random number generator to determine the column orders in the rotation. Figure 6.5 illustrates the table once this has been done. The properties of the Latin square are still maintained and topics are no longer completed consecutively. Note that even after randomization of the columns (not topics) it is still the case that each topic will be completed first, second, third, etc. an equal number of times.

The second thing that a standard Latin square design does not address is the order effects introduced by the interfaces. One assumption behind a Latin square rotation is that there is no interaction

between the items represented by the rows and columns. Notice in Figures 6.4 and 6.5 that Interface 1 is always used first, Interface 2 second and Interface 3 third. The previous discussion of order effects as they relate to a fixed topic order also applies to a fixed interface order. Learning and fatigue may conspire to impact the results.

6.4.3 A Graeco-Latin Square Design

The solution to the problem described above is to rotate the order in which subjects experience the interfaces. This can be accomplished with a *Graeco-Latin square* which is a combination of two or more Latin squares. This is essentially equivalent to reproducing the Latin square in Figure 6.4 above three times, each with a different interface order. A single representation of this is displayed in Figure 6.6. In this Figure, the interfaces are now represented within the cells instead of as column headings. The column headings represent points in time (or order) and the rows represent subjects. For instance, the first user would use Interface 1 to complete Topics 1 and 2, and then Interface 2 to complete Topics 3 and 4, etc.

Subjects	Time 1	Time 2	Time 3
S1	I_1: 1, 2	I_2: 3, 4	I_3: 5, 6
S2	I_1: 2, 3	I_2: 4, 5	I_3: 6, 1
S3	I_1: 3, 4	I_2: 5, 6	I_3: 1, 2
S4	I_1: 4, 5	I_2: 6, 1	I_3: 2, 3
S5	I_1: 5, 6	I_2: 1, 2	I_3: 3, 4
S6	I_1: 6, 1	I_2: 2, 3	I_3: 4, 5
S7	I_2: 1, 2	I_3: 3, 4	I_1: 5, 6
S8	I_2: 2, 3	I_3: 4, 5	I_1: 6, 1
S9	I_2: 3, 4	I_3: 5, 6	I_1: 1, 2
S10	I_2: 4, 5	I_3: 6, 1	I_1: 2, 3
S11	I_2: 5, 6	I_3: 1, 2	I_1: 3, 4
S12	I_2: 6, 1	I_3: 2, 3	I_1: 4, 5
S13	I_3: 1, 2	I_1: 3, 4	I_2: 5, 6
S14	I_3: 2, 3	I_1: 4, 5	I_2: 6, 1
S15	I_3: 3, 4	I_1: 5, 6	I_2: 1, 2
S16	I_3: 4, 5	I_1: 6, 1	I_2: 2, 3
S17	I_3: 5, 6	I_1: 1, 2	I_2: 3, 4
S18	I_3: 6, 1	I_1: 2, 3	I_2: 4, 5

Fig. 6.6 A basic design with Graeco-Latin square rotation for topic and interface.

Note that this design has the same problem as the design in Figure 6.4: Interface 2 always follows Interface 1 except when Interface 2 is first. To address this problem, the same column randomization strategy described above can be applied. The column randomization should be applied after the Graeco-Latin square has been built, otherwise it cannot be ensured that each topic will be paired an equal number of times with each system.

Randomization should be used to assign subjects to the different rows in the table, even when the columns have been randomized. All experimental designs assume random assignment of subjects to conditions. To accomplish random assignment, numbers could be assigned to the rows in Figure 6.6 and a random number generator could be used to determine the order of the rows. Random assignment to condition controls for any potential differences that might be attributable to subjects. The assumption is that any individual differences in subjects (e.g., intelligence, search experience, and motivation) that might impact the results will be equally distributed across condition and therefore controlled as much as possible.

Notice that the rotation in Figure 6.6 provides insight into how many subjects are needed for the study. We know that we need at least 18 subjects to get through the rotation once and to keep the study completely balanced we would need to recruit subjects in batches of 18. However, this is not the only way to determine an appropriate sample size. Statistical power, representativeness and generalizability are also important factors.

6.4.4 Using the Mathematical Factorial to Construct a Design

Another method that can be used to construct an experimental design makes use of the mathematical factorial to enumerate all possible orders for topics and interfaces. However, it is important to note that this is *not* a Latin square rotation — it is a factorial rotation. It is also important to note that this type of rotation is infeasible and cannot be used to create a completely balanced designed in most cases. For instance, in our example with three interfaces and six topics, we would first need

to do a factorial for interface type ($3! = 3 * 2 * 1$), which results in six possible orders. Next, we would need to do this for the six topics ($6! = 6 * 5 * 4 * 3 * 2 * 1$), which results in 720 possible orders. To make the experiment completely balanced, we would need 4320 subjects ($6 * 720$). It is unlikely that anyone would have the resources to recruit and study 4,320 subjects for a single study and besides, studying this many subjects is not really necessary since at some point statistical power plateaus. One might select a portion of these orders, but this would not result in a completely balanced design. However, there are some types of situations, where a factorial rotation is feasible. For instance, two interfaces ($2! = 2 * 1$) and four topics ($4! = 4 * 3 * 2 * 1$) results in 48 possible orders.

6.5 Randomization and User Choice

Another method that can be used to create experimental rotations is to randomly create orders. This can be done by combining different orders of the interfaces and topics, or in conjunction with a Latin Square, where the main variable of interest, interface type, is rotated using a Latin Square and topics are randomly assigned to subjects. It is often the case that researchers want to include more topics in a study to increase generalizability and randomization is selected as a way to assign topics to subjects. However, topics are unlikely to be equally represented in the data set (unless very large numbers of subjects are studied). Thus, results may be attributable to topics and/or topic interactions with other independent variables. If one can use a Latin or Graeco-Latin Square design, then it is a better choice for ensuring a more balanced experimental designed.

Another approach is to give subjects a choice of topics. For instance, subjects might be presented with 10 topics and allowed to select four that they would like to research using the experimental systems. The justification for this approach is it helps increase subjects' motivation [292]. However, if one does this, one should be careful not to give subjects too many choices and have some control over how many topics are completed with particular systems. The danger in letting people choose is that the choices may naturally create a situation where topic

effects are present. There will likely be an unequal distribution in subjects' choices, resulting in some topics being overrepresented in a study and others being underrepresented. It may also be the case that some system–topic pairs occur more frequently unless extra effort is taken to prevent this.

6.6 Study Mode

The mode in which a study is administered can also vary. IIR evaluations can be administered in batch-mode, where multiple subjects complete the study at the same location and time or in single-mode, where subjects complete the study alone, with only the researcher present. The choice of mode is ultimately determined by the purpose of the study. In studies where subjects are deceived in different ways, completing different sequences of activities or will be interviewed, single-mode studies are more appropriate. If the experiment is relatively self-contained, subjects do similar things and can be directed via computer, then batch-mode is appropriate.

Batch-mode studies are very efficient — more subjects can be ran in a shorter period of time. However, it is important to note that subjects can influence one another even when they do not communicate verbally. For instance, in a batch-mode design, the first person to finish the study will likely signal to others that the end of the experiment is approaching. As a result, the remaining subjects might work faster and be less thoughtful, even if they are in different conditions that require more time. Thus, one should think carefully about non-verbal signals that are present in batch-mode studies, what these signals might communicate and how they might contaminate or change a subject's experiences and subsequent behaviors.

Studies can also be administered via the Web instead of in the laboratory. Toms et al. [272] adapted the traditional TREC Interactive Track IIR evaluation model so that it could be run on the Web. The WiIRE framework provided an infrastructure where researchers could plug-in different systems or interfaces for evaluation and tailor common instruments, such as questionnaires, to their needs. Researchers are increasingly experimenting with different ways of administering

evaluations online, although the impact of this on the quality of the evaluation data is unclear. The main concern is that allowing people to login to a system and complete a study in any environment potentially introduces confounding variables that will be unknown to the researcher. For instance, one subject might complete the study while sitting in a loud environment, another might multi-task between the study and other tasks, including text messaging and emailing, while another might solicit help from others or refer to alternative resources of information while searching. Of course, these all represent real use scenarios (and subjects can be instructed about what is expected of them) and this may be what interests the researcher. However, such studies should not be treated as controlled experiments, because they are not.

6.7 Protocols

A study protocol is a step-by-step account of what will happen in a study. It is useful to have a document describing in detail exactly what should happen to guide the researcher. Check lists and other such documents can be used to ensure consistency in the administration of the study. This consistency helps maintain the integrity of the study and ensure that subjects experience the study in similar ways. Creating a detailed protocol also helps ensure that the experiment will run smoothly and that the researcher knows what to expect. In cases where multiple researchers are conducting a study, a protocol helps ensure that the same steps are followed for each subject.

6.8 Tutorials

When subjects encounter new IIR systems it is often the case that they need some instruction on how to use them. Many of the systems that IIR researchers investigate are experimental and thus, differ from the standard systems to which subjects are accustomed. In the past, researchers have created print tutorials to introduce subjects to an experimental system, while others have verbally administered tutorials. Of the two, the print option is best because it ensures that the

presentation is consistent. These days, an easy method for creating a tutorial is to record a video tutorial using screen capture software. This video can be played for each user and it can be guaranteed that what is told and how it is told is consistent. It is best to first develop a script before creating a video tutorial.

There are objections to the use of tutorials and other instructional materials on the grounds that they potentially bias subjects and that in real life people do not read instruction manuals. The issue related to bias is arguably the more important objection; the tutorial may suggest to subjects how they should interact and behave. If one is using a measure such as uptake of a new feature and the feature is prominently discussed in the tutorial, the measure may just reflect how cooperative subjects are, rather than their real interests in the feature. However, if the purpose of the experiment is to evaluate a new feature, then asking people to use the feature seems reasonable since it must be used in order to be evaluated. When it is necessary to provide a tutorial researchers should ensure consistency and balance in the presentation and consider how this experience might influence subjects' behaviors and the study results.

6.9 Timing and Fatigue

Another issue that needs to be considered is the length of time the study will last. This is a critical issue because typically subjects are performing activities that take some length of time to complete. Unlike studies in psychology, where hundreds of trials can be conducted in a single hour, very often only four search tasks can be completed in one hour. Moreover, search activities can be exhausting both mentally and physically. There are no set rules on how long one should give subjects to complete tasks; this is usually contingent on the type of task and study purpose. For instance, in an evaluation of Web search result surrogates, Käki and Aula [158]) imposed short time limits in an attempt to simulate how people actually scan surrogates in real life. In many other evaluations, subjects are given 10–15 min to complete search tasks.

6.10 Pilot Testing

One way to get an estimate of how long a study will last is to conduct a pilot test. Pilot tests help researchers do a number of other things besides estimate time. They help researchers identify problems with instruments, instructions, and protocols; allow systems to be exercised in the same way they will be in the actual study; provide researchers with an opportunity to get detailed feedback from test subjects about the method; help researchers gain comfort with administering the study; and finally, they can be used to train inexperienced researchers. Ultimately, pilot tests help researchers identify and eliminate potential confounds and errors that might otherwise compromise the integrity of the study results.

7

Sampling

Many different items are sampled in IIR evaluations including users, tasks, topics and documents, although the biggest emphasis is on users. The term *element* will be used to refer to items that are sampled, although in most of the discussion that follows users are the focus.

First, it is important to note that it is generally not possible to include all elements from a population in a study, which is one of the main reasons for sampling. In most cases we do not know or have access to all elements in a population. Thus, populations are sometimes described as *theoretical*. It is to this theoretical collection of elements that researchers aim to generalize their results.

The population is usually not mentioned in most IIR reports. Instead, it must be assumed that the target population is all people who engage in online information search, or all literate people who engage in information search between the ages of 18–70 or just all people. There is an implied population behind all studies that involve samples (whether the elements are people or tasks or documents), even if it is not stated explicitly.

Any discussion of sampling must include a discussion of the specific numbers of elements that should be included. In other words, what

is a sufficient sample size? There is no principled, all-encompassing response to this question. There are some general rules of thumb that can be used to estimate how many subjects one needs based on one's design. For instance, counter-balancing a design will show how many subjects are needed to get through one rotation. There are also formulas that can be used to conduct power analysis in cases where a factorial design is used to determine how many subjects are needed in order to achieve a specific power[1] and formulas for survey research to determine appropriate sample sizes given specific margins of error and confidence levels.[2] In general, more is better, although there is a point at which one receives diminishing returns with respect to statistical power. Sample size is used in almost all statistical computations and if the sample is small, then it will be difficult to find statistically reliable results, unless the effect is extremely strong.

It is important to note that some types of methods, in particular qualitative methods, do not rest on the notion of probability sampling and are not focused on statistical testing. It is often the case that qualitative studies and some naturalistic studies have small numbers of subjects. There is a trade-off between the number of subjects that can be included and the intensiveness and depth of the interactions that can occur with those subjects.

Ultimately, the purpose of the research determines the sampling approach and the sample size. The important thing is for researchers to understand the limitations associated with their sampling strategies and exercise caution with interpreting results and generalizing findings. One assumption behind inferential statistical testing is that the sample is representative of the population from which it was drawn and that it was drawn using probability sampling techniques (although this is a theoretical assumption because in most cases this is violated). One should be mindful that a statistic is an estimate of some value in the population (known as a *parameter*); if this estimate is made on the basis of an unrepresentative sample, then statistical test results will not be reliable.

---

[1] See http://www.stat.uiowa.edu/~rlenth/Power/ for an example.
[2] See http://www.surveysystem.com/sscalc.htm for an example.

There are two major approaches to sampling: probability sampling and non-probability sampling. With few exceptions, most sampling in IIR is of the non-probability variety because it is nearly impossible to meet all of the criteria of probability sampling, especially when sampling people.

7.1 Probability Sampling

Probability sampling suggests ways of selecting a sample from a population that maintains the same variation and diversity that exists within the population. If there were no differences among people and all people were exactly alike, then there would not be a need for sampling; in fact, a researcher would only need a sample size of 1. However, since people vary along a number of characteristics (e.g., sex, level of education, and search experience), the goal of probability sampling is to create a sample that contains the same variation of these characteristics that exists within the population. Such a sample is termed *representative* because the distribution of these characteristics in the sample matches their distribution in the population. For example, if females comprised approximately 60% of the population and males 40%, then a representative sample would contain roughly 60% females and 40% males. The important thing about representative samples is that they increase the *generalizability* of the results. Generalizability is related to the extent to which the study results reflect what would happen in the entire population.

Although it is nearly impossible to have a perfectly representative sample, a probability sample is more representative than a non-probability sample. Probability sampling rests on the assumption that all elements in the population have an equal chance of being selected. There are several techniques that can be used to accomplish this, most notably simple random sampling. However, all of these techniques rest on the assumption that all elements in the population are known and that all elements will be included when selected. Stated another way, in order for each element to have an equal chance of being selected for the sample, it must be known *a priori*, otherwise it does not have a chance of being selected. If one's population is all students at a particular

university, then this information would likely be available. However, if one's population is all adults between the ages of 18 and 60, then it unlikely that a comprehensive list of such elements exists. In most sampling situations, especially those in IIR evaluations, knowledge of all of the elements is not the norm, especially when people are being sampled. This is also the case if one considers the world of possible tasks, topics and documents that exist — it would be impossible to enumerate all of these in order to execute probability sampling. However, if a small corpus of documents functioned as the population (e.g., a TREC corpus), then it would be possible to use probability sampling to randomly select documents.

The other problematic piece of this is that all selected elements would need to be included in the sample. With documents or other inanimate objects, this is not problematic so long as the researcher has access to these items. With human beings, this is more problematic since people cannot be forced to participate in a study. People who decline to participate might possess some characteristic that distinguishes them from those who agree to participate. This characteristic will then be absent from the sample which will in turn compromise the representativeness of the sample to randomly select documents.

As mentioned earlier, in order to conduct probability sampling, every element must have an equal chance of being selected. This assumes a constant probability and that one either draws the total sample simultaneously or that sampling is done with replacements. Otherwise, those elements selected later will have a greater chance of being selected than those selected earlier, since the probability of being selected changes as each element is removed from the population and added to the sample.

The other possible method for ensuring that all elements have an equal chance for inclusion is sampling with replacements, although this is really only a theoretical solution. Sampling with replacements means that each element that is selected is returned to the population and essentially exists in two places, the sample and the population. For instance, if one were drawing names from a hat, this means that after a name is drawn, one would note that this element was part of the sample and then return it to the population (i.e., the hat) and continue

drawing more elements. This ensures that each element always has the same chance of being selected, but this also implies that some elements can be selected more than once. Of course, it is not practical to include the same person in the same study more than once, so sampling with replacements cannot be used when each element in the sample needs to be unique.

Another factor affecting the representativeness of the sample is its size. Even when one is able to employ probability sampling, if enough elements are not selected, then the sample will not be representative of the population. For instance, if a population size is 1000 and a researcher randomly selects 10 elements for the sample, then it is unlikely that the variation that exists in the population will exist in the sample. In this example, the sampling ratio is 1%. However, the relationship between a sample and population size is not linear with respect to statistical power. At some point, statistical power begins to plateau and each increment in the sample size adds virtually no statistical power.

There are three major probability sampling techniques: simple random sampling, systematic sampling, and stratified sampling. *Simple random sampling* is the most basic type of probability sampling. First, one creates a list containing all elements that are to be sampled and associates numbers with each element. For instance, if items are listed in a spreadsheet, the row numbers could function as numeric identifiers. Next, one uses a random number generator or a random number table to identify which elements will be included in the sample. It is assumed that most readers are familiar with these techniques. The second technique, *systematic sampling*, is also likely to be familiar to readers. When using this technique, every kth element in the list is selected for the sample. K can be determined by dividing the population size by the desired sample size.

Stratified sampling is a technique that can be used in conjunction with simple random sampling or systematic sampling. The purpose of stratified sampling is to subdivide the population into more refined groups, which are defined according to specific strata, and then select a sample that is proportionate to the population with respect to the strata. For instance, if one were sampling documents that were

associated with specific topics such as science, technology, and literature, one might want to ensure that the proportion of documents associated with each topic in the sample was equal to the proportion of documents associated with each topic in the population. In this example, there would be one strata, topic. After a desired sample size has been determined, proportions can be used to determine target numbers of documents for each topic. Elements from each group would then be sampled using a random number generator, with the number of elements selected for each sample group proportionate to the size of these groups in the population. It is possible to include multiple strata (e.g., topic, publication source, date). This would require the researcher to identify target proportions at a finer level. Multi-dimensional matrices are helpful for visualizing multiple strata where each cell will have a specific proportion associated with it.

Systematic sampling can also be used in conjunction with stratified sampling. Elements of the population can be grouped and sorted according to the various strata in a list format, rather than a matrix. For instance, documents might first be sorted according to topic, then publication source and then date. Once the list has been assembled and sorted accordingly, systematic sampling can be used to select elements from the list. Specific proportions do not have to be associated with the strata because an equal proportion of elements should be selected since the list is ordered.

7.2 Non-Probability Sampling Techniques

In most IIR evaluations, non-probability sampling techniques are used. There are several reasons for this. Researchers often do not know all of the elements in a population and therefore cannot generate the lists required for probability sampling. Even if the elements are known, researchers may not have access to them. For instance, some people may refuse to participate or some documents may be impossible to obtain. Financial constraints and other resources also limit what is possible. Even if one were able to select a random sample of people in a geographic area for an IIR study, it is unlikely that the project budget

would be large enough to pay the travel and lodging costs for potential subjects.

The biggest weakness of non-probability sampling techniques is that they limit one's ability to generalize. This does not mean that research using non-probability sampling should be dismissed, but it does mean that researchers should exercise caution when generalizing from their data and be explicit about sampling limitations in their reports.

There are three major types of non-probability sampling: convenience, purposive, and quota. *Convenience sampling* is the most common type of sampling used by researchers in a number of areas, including IIR. This type of sampling is used in IIR and IR more broadly, to sample users, topics and documents. Convenience sampling is relying on available elements to which one has access. When researchers recruit undergraduate students from their universities or people that are geographically close to them, this is convenience sampling. When newspaper articles or congressional reports are used to create a corpus because one has access to these documents (including copyright permissions), this is convenience sampling. These represent smaller subsets of what is possible, but this is usually a result of practical constraints not researcher negligence. Even the massive log-based studies conducted at search engines companies rely to a certain extent on convenience sampling, since they only include users of a particular search engine. Of course, they have a lot of users, so this helps, but it still is not a probability sample since people choose to use one search engine over another.

Purposive or *judgmental sampling* happens when a researcher is interested in selecting subjects or other elements that have particular characteristics, expertise or perspectives. Inclusion and exclusion criteria are usually associated with purposive sampling; such criteria indicate who may be included in a study. For instance, during an initial evaluation of a new IIR interface, one may purposively select usability experts or students enrolled in a human–computer interaction course to gain very detailed and critical feedback about the interface. Subjects without such expertise may not be able to consider and articulate as wide a range of responses as those with such expertise.

The last type of non-probability sampling is *quota sampling*. It is identical to stratified sampling except that the technique for populating

the strata (or cells) is a non-probable technique (most likely convenience sampling). For instance, after one has created the strata and identified the target number of subjects for each stratum, during recruitment, cells of the table are populated on a first-come-first-served basis. The researcher might send a solicitation to all undergraduates and ask them to characterize themselves according to these strata when replying to the solicitation. The researcher would then be able to ensure that the appropriate numbers of subjects with each characteristic and combination of characteristics are included in the sample. The other major difference between quota sampling and stratified sampling is that the proportions associated with each stratum are usually not as accurate for quota sampling as they are for stratified sampling since information about such proportions in the popular may not be available.

7.3 Subject Recruitment

There are many different methods for recruiting subjects including posting signs, sending solicitations to mailing list, inviting those who work in one's organization, using subject pools and using referral services. Researchers also work in conjunction with others to identify and recruit subjects when they do not have immediate access to target subjects; for instance, a researcher might work in conjunction with the military to recruit intelligence analysts. Recently, many researchers have relied on crowdsourcing and mechanical Turk to recruit people to make relevance assessments via the Web [7]. As more people begin to experiment with conducting IIR evaluations on the Web, additional recruitment strategies may be observed such as Web advertising, mass mailings and virtual postings in online locations.

Because many researchers rely on convenience sampling and are located in academic institutions or technical–industrial settings, there has always been an overrepresentation of undergraduates, computer science students and researchers, and library and information science students in published studies. As long as one describes faithfully the sample and the recruitment techniques, then readers can make their own determination about the validity and generalizability of the results.

Having such groups as subjects is not inherently bad, but extrapolating wildly from results is. The more important question is about the extent to which these groups may be biased. In particular, many researchers have used their lab mates or those in their research group or even themselves as study subjects. This is very problematic because these people likely know something about the purpose of the study and the desired outcome. When study subjects also end up analyzing the data, there is even more room for concern.

7.4 Users, Subjects, Participants and Assessors

There are a variety of names given to humans who are studied in IIR research. The most common are users, subjects, research participants and assessors. A general rule of thumb for distinguishing among these different labels is as follows. The term *user* is often used in situations where those being studied are actual users of a system. For instance, in log-based Web studies the data are usually generated by real users. These users are not using the system for the sole purpose of generating research data (indeed, most probably do not even realize how their data are used), but instead the data is a byproduct of their normal use of a system.

Interestingly, the phrase *user study* was originally used in information science to describe studies that investigated people's information seeking needs. Siatri [242] traced the first user studies in information science to the 1948 works by Urquhart [278] and Bernal [26] who studied the distribution and use of scientific information, and the reading habits and needs of scientists, respectively. These days user study is used more generally to describe any study that involves human participants, which really dilutes its meaning.

The terms *subjects* and *research participants* are used to describe people who knowingly participate in a research study. Subjects and research participants are the subset of the user population that has been selected for inclusion in a study. The sole reason that these people are using a system is because they are part of a study. The term subject has typically been associated with laboratory studies, while research participant has been associated with naturalistic and qualitative studies.

It is also the case that some people dislike the term subject as it is said to be dehumanizing.

The term *assessor* has been used to describe people whose sole purpose is to make relevance assessments. There is a fine line between an assessor and a subject; one could justify the distinction by noting that the only data produced by assessors that are of interest are the relevance assessments. Although traditionally the behavior of assessors has been of little concern, several researchers have started to investigate assessors [14, 226], thus treating them more as research subjects.

8

Collections

8.1 Documents, Topics, and Tasks

Most IIR evaluations require subjects to search for information. Thus, one important consideration is the identification of a set of documents (or more generally, information objects) for subjects to search and a set of tasks or topics which directs this searching. Along with these things, a researcher must also make some decisions about how the relevance of the information objects to the topics will be determined. These items — corpus, topics, and relevance judgments — comprise what is commonly known as a *test collection* in IR. In IIR evaluations, these items can be thought of as instruments just like questionnaires and logging software.

Although *collection* is often used as a synonym for *corpus*, in this paper these words will be used to indicate two separate things in the way that they are used in the context of TREC [121]. A *collection* consists of topics, a corpus, and relevance judgments. A *corpus* is the set of documents, or information objects, that subjects access during a study. IIR evaluations vary on the extent to which they have each of these components. In some studies, subjects are provided with standard

topics or tasks, search the entire Web and make the final relevance judgments. In other studies, subjects are provided with topics, asked to search a specific corpus and although they are able to make relevance judgments, these are ultimately compared against a gold standard (e.g., TREC relevance assessments).

8.1.1 TREC Collections

TREC collections, and specifically those used by the Interactive and HARD Tracks, have been used in a number of IIR evaluations. If one is conducting a controlled, laboratory study or system evaluation, these collections are attractive for a number of reasons. There is a finite and theoretically knowable set of documents and some information is available about the number of documents that are relevant to different topics. It is also the case that relevance assessments exist and different kinds of performance metrics can be computed.

There are also a number of limitations associated with using TREC collections in IIR evaluations. It is generally known that users' queries retrieve different documents than the batch queries used in system-centered evaluations, so it is possible that subjects will find documents that were not included in the relevance pools [129]. If a document was not in the pool, then it would not have been judged by the original assessor. Strictly speaking, documents in this situation are considered not relevant. One might be tempted to just independently assess these documents, but this potentially perturbs any findings since the majority of assessments will be from a single assessor and a small number will be from people who did not experience the original assessment context. The stability of the TREC collection rests on the assumption that the relevant documents are relevant to a single user at a single point in time. Mixing assessments made by others violates this assumption.

A bigger problem is related to the extent to which relevance assessments generalize. Numerous studies have demonstrated that relevance assessments do not generalize across subjects [80, 129]. Indeed, it is understood that different people will make different relevance assessments given the same topics and documents. TREC acknowledges

this and makes no claims about the generalizability of the relevance assessments [288]. Their stance is that these assessments represent one user's judgments at one point in time. This is fine when one is comparing relative system performance, but problematic when trying to use these assessments as benchmarks in studies with new subjects. It has even been argued that the performance measures that are computed in IIR evaluations using TREC relevance judgments actually represent the extent to which subjects agree with the assessors, rather than performance [80].

An alternative is to completely ignore the relevance assessments made by assessors and instead generate new relevance assessments based on subjects' actions. This will impact one's abilities to compute certain performance metrics, especially those based on recall, but these may not represent performance concepts that are well-suited to the purpose of the study anyway. Creating new assessments can be done using a consensus approach [301] or by just accepting what individual subjects identify as relevant for the topic. One problem with consensus-based approaches is that one will likely end up with a lot of documents that were saved by a single user, but discarded by others. However, in cases where researchers are using graded relevance, a consensus-based approach provides a useful way to grade documents.

Another limitation of using TREC collections in IIR evaluations is that most of the corpora are newswire text. Although it is the case that a large number of internet users get news online[1] this is not the only type of searching task that is performed and more often than not, users browse the same news sources daily rather than actually search for news articles. More recent TREC Tracks have explored different kinds of corpora, such as blog and legal.

Finally, another major criticism of using TREC collections in IIR evaluations is that the topics are artificial. However, it is possible to make this criticism of any study that uses artificial topics and tasks, so this is not unique to TREC collections.

[1] According to the latest report from the Pew Internet & American Life Project Tracking surveys 72% of online users view news online http://www.pewinternet.org/trends/Internet_Activities_6.15.07.html.

8.1.2 Web Corpora

In many studies, the Web is used instead of a corpus. Subjects are allowed to search the Web and there are no constraints on what pages and resources they can use. The major drawback to using the Web is that it is impossible to replicate the study since the Web is constantly changing. Two subjects in the same study might issue the same query at two different points in time and get completely different results. Although network delays are less of an issue these days, it is also the case that different subjects might experience different reaction times depending on the time day. Factors such as authority and quality enter more into subjects' relevance assessments since these are not controlled for as they would be in a closed collection. If the subject knows that all documents are from *The New York Times*, then the subject does not have to worry as much about source credibility. This suggests that subjects should be required to provide more and different kinds of relevance assessments when searching the web, than when searching closed corpora [271]. It is also the case that subjects may base their relevance judgments on more things than just the text (e.g., design, style, and images). Finally, with open web searching, researchers are only able to use a small number of the well-established performance metrics.

In log-based Web studies, corpora are often built on-the-fly, as subjects visit the Web pages. For researchers working at search engine companies, these corpora have traditionally consisted of the search engine home page, the search results page and first level search result. Recent browser plug-ins allow these researchers to gather page visits that go beyond this traditional tripartite set, but some newer plug-ins have the potential to corrupt this data.[2] Regardless of who is conducting the study there is potential to create an offline corpus of documents viewed by study subjects and to perform future experiments with this collection (cf. [269]).

It is also possible to study IIR in the context of open or closed corpora that are accessible via the Web. For instance, one might study IIR in the context of a digital library, a proprietary database, an internal business corpus or intranet. Searching such circumscribed corpora

[2] http://mrl.nyu.edu/∼dhowe/trackmenot/

provide more control over the retrieval situation and a greater understanding of the corpus and range of document types. Assumptions can also be made about the quality and kind of documents within the corpora. However, the difficulty is ensuring that the corpus is of a sufficient size so that the retrieval task and results are meaningful.

Finally, there are some closed Web corpora that have been crawled and assembled by various research groups, including the TREC .gov corpus. These corpora share some of the positive things related to closed corpora, mainly they are stable, theoretically definable and self-contained, and allow for the computation of recall-based measures.

8.1.3 Natural Corpora

Natural corpora are corpora that have been assembled over time by study participants. Studies using natural corpora are most often in the context of personal information management [119]. The benefits of using such corpora are that they add to the realism of the study and allow subjects to interact with documents that are meaningful to them and with which they are familiar.

The use of natural corpora also has drawbacks. The biggest problems are lack of replicability and equivalence across participants. Each subject's corpus will differ in size and kind. One subject may have a small set of word processing documents and Web pages, while another subject may have gigabytes of documents representing a variety of file types. Thus, the researcher needs to know something about the number and kind of files available on the machine to interpret the subject's behavior. Cross-user comparisons are difficult since each subject's experience will be different and dictated in part by their own corpus. There may also be unknown document type–tool interactions. For example, the experimental tool may handle some document types better than others and there is no way to control the distributions of such document types across subjects' corpora. Furthermore, if the corpus resides on the subject's machine, the subject may prepare for the study by organizing, deleting and filing things, which changes the natural state of the corpus. Finally, researchers are unable to do follow-up experiments since the corpus resides on the subject's machine and changes constantly.

To overcome some of these problems, natural corpora can be transferred to a research machine where they can be controlled. This might be necessary if the IIR application being tested is not robust enough to be deployed in other environments. The downside to this is that subjects will have to spend time preparing and transferring these corpora to the researcher. Of course, there is also the possibility that some organizing and self-censoring will occur because subjects may be self-conscious about how their corpora exist in natural form. As users accumulate more and more electronic information and increased attention is given to personal information management, we will likely see more researchers developing methods for using and experimenting with natural corpora.

8.1.4 Corpora and Behavior

One important thing to point out about choice of corpora is that it will impact the way in which subjects behave and seek information. For instance, behavior with a Web-based corpus with lots of hypertext links will be different from behavior with a TREC newspaper corpus that has no hyperlinks. The search strategies and tactics that subjects employ when using their own personal corpora might differ from those they employ when searching the open Web. When designing an IIR study, it is important to recognize the potential impact of the corpus on behavior and interactions.

8.2 Information Needs: Tasks and Topics

A user's information need is perhaps one of the most critical aspects of information seeking and retrieval. This need forms the basis of the user's activities and relevance judgments. Much has been written about the nature of information needs and it is generally accepted that people often have a difficult time articulating their information needs and translating them into a vocabulary that is appropriate for a system [19, 20, 25, 104, 267, 297]. Research has also shown that information needs evolve during the search process; this evolution results in dynamic relevance assessments — that is, as people learn more about their information needs, their relevance behaviors change [266, 281].

Although it is difficult to create an all-encompassing definition of an information need, most information needs can be characterized in terms of task and topic. In the published literature, these three terms (information need, task, and topic) are often used interchangeably. However, it is important to distinguish among them so that one is clear about what is being studied. A task represents the goal or purpose of the search — this is what a user wants to accomplish by searching. For example, a traditional task is gathering information to write a research report. Other tasks include planning travel, monitoring sports scores, navigating to a homepage, or re-finding previously seen information. The topic represents the subject area that is the focus of the task. For example, one might be gathering information to write a research report about the malaria epidemic in Africa or one might be planning travel to Australia. One task might be associated with several topics and one topic might be the focus of many different tasks. It is the combination of the specific task and topic that forms the information need.

Historically, IR focused on the topical aspects of the information need. In the early years, systems were developed for trained searchers who were searching on behalf of a client, typically a researcher or scientist looking for exhaustive information about a particular topic. The task was often constant (or assumed to be constant) and was more recall-oriented. Even after the target user group changed to include non-expert searchers, the model search task was still somewhat stable since IR systems were only located within specific environments. With the development of Web IR, different kinds of task models began to develop as the types of users, tasks and use environments diversified. Web IR, in particular, has brought to the forefront precision-oriented information needs, where users are looking for one or a small number of documents rather than all of the documents about a particular topic. Currently, there is a swing back towards recall-oriented information needs, including exploratory tasks, where users are looking for a larger number of documents and have information needs that are unfocused and evolving [190].

Although research trends can impact which tasks are considered important, underlying all IR research is some *user model*, which is an abstract representation of target users, and one or more *task models,*

which represent the goals of users. One user model that has been used a lot in IR is that of a librarian or other search intermediary. Other examples include intelligence analysts and undergraduate students. Examples of task models include finding documents for a survey article, homepage navigation, and fact-checking. The user and task models help define the particular behaviors and activities the IR system is intended to support and help determine the appropriateness of particular evaluation measures and study participants. Although studies may have one or more task models, typically there is only one user model. User and task models are often implied and inherited from research tradition. For instance, underlying many IIR evaluations is the TREC *ad-hoc* task which is modeled after exhaustive searching.[3] Appropriate user models include students writing a survey paper for a class, or intelligence analysts preparing a briefing for a military official. Historically, user and task models have been held constant within a particular study and the only thing that changed was the topic (hence the reason for referring to *TREC topics*), but current research is starting to employ different task models within a single study and using task as an independent variable. Some common task-types that have been investigated include navigational, known-item, fact-finding, resource finding, homepage finding, exploratory and informational.

Task has been shown to affect users' information seeking behaviors and relevance judgments in a variety of ways and is a good candidate variable for understanding more about search systems and user behavior [163, 170, 171, 249, 266, 281]. Vakkari [279] provides an overview of task-based information searching. Byström [43, 44, 45] has conducted a large number of studies investigating task complexity and how tasks can be defined, measured and studied. In particular, Byström and Hansen [44] distinguish among work tasks, information seeking tasks, and information retrieval tasks. Li and Belkin [177] developed a faceted approach to the conceptualization of task. Ingwersen and Järvelin [139] have also provided task classifications. Bell and Ruthven [24] and Gwizdka [114] explore measures of task complexity and difficulty. Toms et al. [273] examined the effects of several task-related variables on interactive

[3] All TREC collections (and Track) have some user and task model associated with them.

search behavior. Kim and Soergel [172] identified various task characteristics and proposed how they can be used as independent variables in research studies.

Search behavior and relevance judgments can also vary according to topic, but usually this is a result of variations in user related variables, such as how much a user knows about a particular topic (e.g., [134, 291]), and corpus related variables, such as how many relevant documents are available about a particular topic (e.g., [135]). It is common in many IR and IIR evaluations to investigate performance and other dependent measures with respect to topic in a *post-hoc* fashion (i.e., at the end of the study), but it is uncommon to treat topic as an independent variable.

8.2.1 Generating Information Needs

Creating information needs for a study is difficult for a number of reasons. It is not always clear at what level of specificity a task or topic should be defined or how many facets should be used to describe the need. For instance, tasks such as gathering information to write a research report or planning travel can be broken down into a series of sub-tasks or, conversely, they can be grouped together into the broad task of seeking information. A topic such as elephant poaching can be further broken down into techniques, places, penalties, policies, etc. or it could be described at a higher level simply as elephants. Other considerations that must be made is whether there is a sufficient number of documents in the corpus and whether target users will have the basic abilities and knowledge to complete specific tasks. One of the most difficult aspects of creating information needs is ensuring the needs are appropriate to what is being studied, but are not over-engineered to guarantee success. Unfortunately, there is little formal guidance for creating tasks and information needs and it can be argued that it is impossible to create artificial information needs, since information needs are generally believed to reside within a person's head. Indeed, most research focuses on information tasks, rather than information needs. Elsweiler and Ruthven [86] outline some steps that one might execute to create search tasks in the domain of PIM.

In many IIR evaluations, where information needs are assigned to subjects, researchers are using TREC collections, which come with topic descriptions. This is one benefit of using TREC collections. However, there are times when researchers must create information needs, especially if they are studying the open Web. A common approach is to examine query logs and work backwards from the queries to develop information needs. It is important to note that queries are not synonymous with information needs. Often users will issue a number of queries during the resolution of an information need. This has been one criticism of log-based studies that only passively monitor what users do — such studies amass large amounts of queries, but it is unclear to what end these queries were written. The ability to isolate and study search sessions which might be comprised of a number of queries from a single user helps address this, but it still does not allow one to understand the nature of the information need. Despite the difficulties of going from a set of queries to an information need, such queries provide researchers with some insight into the kinds of things for which people are searching.

Another way to develop information needs is to work with experts who would either be assigning such information needs to target users as work tasks, or have the same kinds of information needs [138]. For example, if a system were designed to support intelligence analysts, then a group of intelligence analysts might be enlisted to help develop the search tasks. A final way to approach information need (or task) creation is to first identify different characteristics of an information need — for instance, information needs can be well-defined or ill-defined, fleeting or persistent — and then combine the facets in different ways to construct the needs (e.g., [177]).

8.2.2 Simulated Work Tasks

One important development in IIR evaluation and experimentation has been the simulated work task, which Borlund [34] describes as a short cover story that describes the situation leading to the information need. Simulated work tasks go beyond simple topic-based descriptions of needs by providing more contextual information that is tailored

towards target users. Borlund and Ingwersen [35] note that the simulated work task describes the following to the user: the source of the information need, the environment of the situation, and the problem which has to be solved. This problem serves to make the test person understand the objective of the search. Such descriptions are further proposed to provide a basis against which situational relevance can be judged.

Simulated work tasks are comprised of two major parts, the simulated work task situation and the indicative request (although the indicative request is identified as optional). Together they are called the simulated situation. An example is shown below in Figure 8.1. Notice in this description, the task situation is tailored to the target users — university students. Also notice the indicative request, which provides an example of what subjects might search.

One of the primary rationales for developing simulated work task situations is the criticism that assigned search tasks are artificial, that subjects may not have a context for executing the task and making relevance judgments and that subjects may simply be unmotivated to search for artificial tasks. Of course, the reason for assigning tasks to subjects is to control the search situation and produce conditions that allow for comparison. Borlund's work provides an empirically validated way to use assigned tasks while also personalizing them. This simulated work task method calls for members of the target user group to generate the tasks to ensure that they are relevant to study subjects. This aspect is often overlooked by those using this method to create tasks for their own studies.

Simulated Situation

Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Indicative request: Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.

Fig. 8.1 Example of a simulated situation/simulated work task situation [32, 33].

Other researchers have tried to address the motivation problem by presenting subjects with a choice of assigned tasks [153, 292]. When using TREC topics, motivation and requisite background knowledge for making relevance assessments become crucial issues since many of these topics were developed by retired US intelligence analysts.[4] When such topics are assigned to undergraduate research subjects, there may be a disparity in motivation and background knowledge. While it may not always be possible to create simulated work task situations, researchers should understand the limitations and constraints of whatever type of information needs are used.

8.2.3 Natural Tasks

Another type of information need that is used in IIR evaluations is natural information needs (or tasks). These are more commonly used in naturalistic studies where experimental control is less of an issue and the study focus is more exploratory in nature. Natural tasks are those tasks that subjects normally conduct in their everyday life. While the behaviors observed from subjects conducting natural tasks are more representative of those subjects' behaviors, it can be difficult to generalize and compare findings across subjects. Another difficulty is that tasks will likely be at varying levels of specificity, stability and completion, the amounts of information available to address different tasks will vary and subjects will know varying amounts about how to complete the tasks.

8.2.4 Multi-Tasking and Multiple Search Episodes

Most people acknowledge that users in the real world engage in multi-tasking and that information-seeking often takes place across multiple search episodes [181, 247, 251]. There have been a few studies of these behaviors, but there have not been a concentrated effort to develop tasks that would be appropriate for more controlled research settings. To study multi-tasking, one would need to create sets of tasks that can

[4] There are some collections where retired intelligence analysts or other US government personnel did not create the topics. For instance, topics in the TREC HARD 2004 collection were developed by summer interns at the Linguistic Data Consortium.

be done simultaneously. More importantly would be the development of a work task situation to provide background for why a user would want to conduct such tasks. The same can be said for tasks that need to be completed across multiple search episodes. Some research on re-finding has experimented with techniques for studying tasks that require at least two episodes (or search sessions) for resolution [48] and there is at least one example from IIR [181], but few researchers have attempted to study multi-tasking and multiple search episodes in IIR. There is not much guidance for creating tasks appropriate for these situations either, although Czerwinski et al.'s [66] diary study of multi-tasking behaviors in an office environment provides some groundwork.

9

Data Collection Techniques

Corpora, tasks, topics, and relevance assessments are some of the major instruments that allow researchers and subjects to exercise IIR systems. Other types of instruments, such as loggers, questionnaires and screen capture software allow researchers to collect data. This section provides an overview of some of the data collection techniques and instruments researchers have used to understand what happens during IIR evaluation.

9.1 Think-Aloud

The think-aloud method asks subjects to articulate their thinking and decision-making as they engage in IIR [87]. The researcher will need to somehow capture this data and this can be accomplished with an inexpensive computer microphone. Most computers come with recording software, so think-aloud data is relatively inexpensive to collect. However, there are problems associated with using the think-aloud method. Most subjects have a difficult time simultaneously articulating their thoughts and completing whatever IIR task has been assigned to them. In many cases, the IIR system is novel and subjects do not have the additional cognitive resources to engage in think-aloud. Think-aloud is

also awkward and unnatural — most people do not go around articulating their thoughts as they complete tasks. If subjects get quiet, the researcher can prompt them to continue thinking-aloud, but in general, people have a difficult time doing this and the researcher will likely do a lot of prompting, which might eventually annoy and distract subjects. Some researchers have proposed that subjects complete a short training task before they start searching to get accustomed to think-aloud. For instance, subjects might be asked to solve a puzzle while thinking-aloud before they start using an experimental system. Although this practice may help subjects become acquainted with think-aloud, it does not change its awkwardness. Finally, there is the question of whether the IIR task is too complex for think-aloud, since it was originally designed to be used with more basic tasks [87].

9.2 Stimulated Recall

An alternative to think-aloud protocol is stimulated recall (see [112] for an evaluation of this method and Rieh [217] for an application example). Stimulated recall (also called *talk after*) is used to collect the same type of data as think-aloud protocol, but differs in that data is collected during and after the search. The researcher records the screen of the computer as the subject completes a searching task. After the task is complete, the recording is played back to the subject who is asked to articulate thinking and decision-making as the recording is played. General instructions can be provided or the subject can be asked specific questions or to focus on specific features or processes. Although there is a delay between the subject executing the task and discussing what was going on, the researcher is generally able to get better data since the subject's attention is not divided. Discussing a completed event with another person is also a more natural activity than thinking-aloud during the event.

Instrumenting stimulated recall is more expensive than think-aloud since the researcher needs to purchase screen recording software. However, there are several relatively inexpensive pieces of software that can be used to record both the screen and audio. Such software can be used to record the screen while the subject searches and then to record the

screen and audio as the subject engages in stimulated recall. Screen recording software can also be used for other experimental tasks, such as creating tutorials.

9.3 Spontaneous and Prompted Self-Report

Another technique for collecting data from subjects while they engage in search is spontaneous or prompted self-report. This technique is not used a lot, in part because it is difficult to orchestrate and can be intrusive, but the basic idea is simple: elicit feedback from subjects periodically while they search. Subjects are not required to continuously verbalize their thoughts (as with think-aloud), but are instead asked to provide feedback at fixed intervals or when they think it is appropriate. The purpose of this technique is to get more refined feedback about the search that can be associated with particular events, rather than summative feedback at the end of the search. This is particularly useful in search situations that span long periods of time. Both qualitative and quantitative feedback can be gathered with spontaneous self-report.

Spontaneous self-report can be hard to integrate into a study because it is almost always intrusive, but one might imagine a persistent window that is visible on the screen at all times where subjects can communicate their feedback. Another method of implementation is to have a window pop-up at fixed intervals or have the researcher ask the subject to provide feedback at specific intervals. Either way, the subject is interrupted and might, after some time, become annoyed.

9.4 Observation

Observation can take two forms in IIR evaluation. In one form, a researcher is seated near subjects and observes them as they search or complete IIR activities. The researcher is usually trained to focus on particular events and behaviors and takes notes that describe their observations. Observation can also be conducted with a video camera or screen capture software. In these situations, the researcher usually conducts the observations at a later date, and the act of observing is actually combined with data analysis. Thus, observation can occur

in real-time or at play-back time (for recordings). Subjects might be uncomfortable because their actions are being monitored, but subjects are not required to do anything extra as with the previously described methods.

During real-time observation by a researcher, the subject is not interrupted, but can be asked follow-up questions about particular events later during post-search interviews. Typically researchers focus on recording things not available in a log, such as the subject's verbal or non-verbal reactions. Researchers might also note unusual interactions or instances when the subject seemed confused or frustrated. It is important to note that when observations are made at play-back time and the researcher has used screen capture software, it is not possible to record non-verbal communication. The only thing one can observe is what is happening on the screen. Non-verbal communication can be captured with a video camera, but recording people's faces puts them at more risk so one must be certain to take extra precautions with this type of data.

Observation is extremely time-consuming and labor-intensive, both in the collection and analysis of data. It is also prone to selective attention and researcher bias. It is extremely important that the researcher has some practice before commencing to observe and has some idea of what particular data need to be recorded and how it will be used. A protocol can be developed to guide the process, as well as a form to structure the observations. Training of observers is also important, especially if multiple researchers act as observers. Effort needs to be made to ensure that researchers record the same types of things and at the same level of detail.

9.5 Logging

One of the oldest and most common methods for collecting data in IIR evaluations is transaction logging (see [144] for a review). Papers were published at the *SIGIR Conference* in 1979 and 1983 describing the use of computer monitoring and its application to understanding user behavior [30, 77]. In 1993, there was a special issue of *Library Hi Tech* devoted to transaction log analysis [210], including an article on

its history and development [209]. The use of transaction logs has been around for quite some time, although the recent explosion of studies using Web transaction log data has re-popularized this approach. Four types of logging will be discussed: system, proxy, server, and client logging. Researchers face three major challenges in using log data: ensuring the validity and reliability of the logger, extracting and preparing data generated by the logger, and interpreting the data.

System logs refer to those that are written for an experimental IIR system. In this scenario, the logging application is built as part of the IIR system and records all interactions specified by researchers. The current practice is to write logs using XML, which greatly facilitates data extraction and analysis. System logs are typically used to characterize the interaction and record both what the system does as well as how the subject reacts. For instance, typical logs will record the subject's queries, the results shown to the subject and the results selected by the subject. This type of logging used to be referred to as transaction logging when applied to operational library systems. Because most current library systems are Web-based, there is a blurring between traditional system logs and Web logs, since in many cases these are the same things.

Common types of logging for studies conducted via the internet are server, proxy and client logging. The primary differences between these types of logging are where the logging takes place and what types of information are available to be logged. Server-side logging has traditionally been used in large-scale log studies conducted by search engine companies. Server-side logging takes place on a server and is thus limited to communications between the user and the server. Most servers come with some type of logging application to record basic data about which resources and services are requested and when they are requested. Such requests are typically associated with the IP address of the machine making the request as well as the client application that is making the request. One limitation of server-side logging is that only resources and services requested from a single server are recorded.[1]

[1] Currently, many search engine companies track users' activities after they leave the search engine server through a browser extension. While initial search logs from these companies

Thus, server-side logging does not track what subjects do once they leave a particular server. Server-side logging also does not record users' interactions with their local machines. This is a particularly important limitation since pages that are requested multiple times within a single period of time (for instance, when the user uses the back button on the browser) are not always routed through the server.

A final problem with server-side logging is that unless users are associated with particular usernames or IP addresses it can be difficult to determine unique identities and associate log records with these identities. In the past, the problem was primarily a result of different users using the same machine; before so many people owned computers, many shared computers that were found in a centralized location (e.g., computer laboratory and public library). Although more people own computers nowadays, many local networks are configured to dynamically assign IP addresses, which prevent one-to-one correspondences between users and IP addresses. Additionally, at least one browser extension has been created which corrupts the information that is sent to the server.[2]

Proxy logging also takes place at a server, but not the server from which the user requests resources. Rather, all communications between a user and a server are routed through an additional server, known as a proxy server. The proxy server may provide traditional internet services, but its main function is to log the communications that occur between a user and all servers to which they make requests. Proxy applications can also be used to modify or change what the user experiences. For instance, Joachims et al. [152] used a proxy application to change the order in which search results were displayed to subjects. Web browsers can be easily configured to point to a proxy server. While proxy logging allows the researcher to track the user's interactions with a number of servers, it still misses many actions that occur on the client machine.

The most comprehensive type of logging happens on a user's local machine via client-side application. Client-side logging is used to refer to logging applications that reside on the user's local machine. The

only contained information about communications with the server, more information is now available in these logs.

[2] http://mrl.nyu.edu/∼dhowe/trackmenot/

term *client* originally described particular applications that made requests to servers, but it is used as a synonym in this context to refer to the user's local machine (e.g., the user's desktop or laptop computer). Client-side logging applications can be used to record the user's interactions with all applications that run on the local machine, including Web browsers and word processing applications. Some of these loggers even record the user's interactions with the operating system. Client-side logging provides a more robust and comprehensive log of the user's interactions and solves most of the problems of server-side logging. Namely, unique identifiers can be associated with records if there is a one-to-one correspondence between the computer and the user and all requests, even those that operate on the cache, are recorded. Furthermore, additional activities that only occur on the client, such as scrolling and printing, are recorded.

Client-side logging is perhaps the most comprehensive type of logging, but it is also the most expensive and difficult to implement [89]. Building a client-side logger from scratch requires a great deal of knowledge and time. Glass box [61] is an example of a very robust client-side logger, but unfortunately it is only available to those working on certain US government funded projects. Jansen et al. [145] has made available Wrapper, an open source application designed to record subjects' Web interactions.[3] There are some commercial applications that can be used too, but researchers must be careful to evaluate these programs to make sure that the logs produce an accurate and reliable record of what occurred. Indeed, researchers bear this responsibility no matter what type of logging is used.

A final approach to logging involves instrumented Web browsers, such as those used by Kellar et al. [163]. This approach consists of creating a specialized browser or plug-in that works in conjunction with an existing browser. Instrumented browsers allow the researcher to focus on people's interactions with one particular application (i.e., the Web browser), to have more control over what is logged, and to log interactions with experimental features. Instrumented browsers also allow researchers to incorporate other data collection tools into

[3] http://ist.psu.edu/faculty_pages/jjansen/academic/wrapper.htm

the logger. For instance, the browser might contain widgets that allow subjects to classify the pages they view according to come criteria.

The primary benefit of any type of logging is that a record of the user's activities and interactions is created. The completeness of records varies depending on the type of logging (or more specifically where this logging takes place). Logging is also a useful method for capturing users' natural search behaviors in studies that occur outside laboratory settings. Most logging applications can run in the background while the user works without causing any disruption or delays. Perhaps the biggest limitation to logging is that only electronic observables can be captured — activities that occur beyond the computer or application are not captured and the purpose and intent of the observed actions is often unclear. Grimes et al. [111] conducted a comparison of data collected via query log, field study and using an instrumented browser and found that the query log provided the least useful data for individual events, but the most useful for understanding the scope of user's activities. Their paper title sums up their conclusion: query logs alone are not enough.

9.6 Questionnaires

The questionnaire is one of the most popular methods of collecting data from subjects in IIR evaluations. Almost all IIR evaluations have some type of questionnaire. Questionnaires can consist of closed questions where a specific response set is provided (e.g., a five-point scale) or open questions where subjects are able to respond in any way they see fit (e.g., what did you like most about this system?). Closed questions typically produce quantitative data, while open produce qualitative data. Thus, closed questions are useful for providing numeric representations of subjects' attitudes and feelings and allow researchers to make statistical comparisons. Open questions are useful for gaining more unique and varied insight into subjects' experiences and for understanding the reasons behind particular attitudes and behaviors. Responses to open questions also allow researchers to better interpret and contextualize subjects' responses to closed questions.

Closed questions typically take one of two forms: Likert-type scales or semantic differentials. In the first format, a series of statements are provided such as, *the system was easy to learn to use*, along with five- to seven-point Likert-type scales for responding, where one scale end-point (or anchor) represents strong agreement and the other represents strong disagreement. Traditional Likert scales measure agreement on a five-point scale, where the scale labels are: strongly agree, agree, neutral, disagree and strongly disagree [178]. The classic Likert measurement also produces a summative measure across a set of items. In IIR, a range of scales points are provided (although it is most common to have five- or seven-point scales), a range of scale labels are provided (e.g., not at all, somewhat, very much) and a summative measure is usually not produced, which is why this type of scaling is often referred to as Likert-type. There is little guidance about the appropriate number of scale points, although some researchers have suggested that seven-points are optimal for eliciting relevance assessments [265]. What is clear is that an odd number of scale points allow subjects to select a mid-point, while an even number does not. It is also the case that scales with more points can be converted into scales with fewer points.

The use of semantic differentials is another common way to format closed questions. Semantic differentials present pairs of antonyms at opposite ends of a scale without numeric labels. Instead of requiring subjects to identify discrete numbers, this scale allows subjects to place a mark along a line, indicating a more continuous-type of measure. While the semantic differential presents individual lines indicating where subjects should mark, other scales have been formatted to present a straight line without any demarcations. Of course, both of these formats only represent a difference in user interface: different points along the continuum are eventually coded into discrete numbers by the researcher for analysis.

Questionnaires can be administered via different modes: electronic, pen-and-paper and interview. Research in psychology and public opinion polling has found that questionnaire mode affects how people respond to questions [42, 274]. Kelly et al. [165] investigated questionnaire mode effects in the context of an IIR experiment and found that subjects' responses to closed-questions were significantly

more positive when elicited electronically, than via pen-and-paper or interview. Although this was only a single study with 51 subjects, these results suggest that questionnaire mode can impact subjects' response behaviors in IIR studies.

Questionnaires are used at various points during a study to collect data. The five most frequently used questionnaires are demographic, pre-task, post-task, post-system, and exit. The determination of which among this set is appropriate and what they contain depends on the purposes of the study. Table 9.1 lists some of the data that are typically elicited via each type of questionnaire.

In addition to these types of questionnaires, researchers sometimes give subjects specialized instruments (usually in the form of a questionnaire) that are used to characterize subjects along some standard measure (e.g., cognitive style, personality type, and spatial ability). These types of instruments are usually administered near the beginning of the study. It is also possible to use questionnaires to pre-screen potential subjects during recruitment if a particular characteristic is of interest. For instance, a researcher might only be interested in subjects who are novice searchers. Pre-screening can also helps to ensure a balanced design if a researcher is aiming for equal numbers of subjects across some characteristic (e.g., sex).

Instructions about how to design questions will not be presented here, but it is critical that appropriate attention is paid to this process and that questionnaires are piloted before they are used (see [207] for more information about question design). Important considerations to make when creating questions are related to the wording and ordering of the questions and choice of scale points, labels and anchors. With respect to wording, it is particularly important that the questions are not biased, loaded or double-barreled. For closed questions, it is also important that scale labels are appropriate to the question and that range of variation, exhaustivity, exclusivity and equivalence are considered. For open questions, the wording and placement of questions within the questionnaire are critical issues.

There are additional considerations to make about closed and open questions. Because a response set is provided for closed questions, the researcher has to ensure that the response set is appropriate. There is

Table 9.1 Commonly used questionnaires in IIR evaluations.

Questionnaire	Purposes	Administration
Demographic	This questionnaire is used to elicit background information about subjects. This information is typically used to characterize and describe subjects, but it can also be used to explore and test specific hypotheses. For instance, a researcher might be interested in investigating the difference between male and female behavior, or among people with different amounts of search experience.	This questionnaire is usually given at the start of the study, but it can be given at the end. The rationale for waiting until the end is that subjects are likely to be fatigued and it is better to get this "easy" information then.
Pre-task	This questionnaire can be used to assess subjects' knowledge of the search task and/or topic. Questionnaire items are usually directly related to the search task in which the subject is about to engage.	Subjects complete this questionnaire before searching occurs so that the search experience does not bias responses.
Post-task	This questionnaire is most often used to gather feedback about the subject's experiences using a particular system to complete a particular task. Thus, the primary goal of this questionnaire is to assess the system–task interaction.	This questionnaire is administered following each task.
Post-system	This questionnaire elicits feedback from subjects about their experiences using a particular experimental system. It is typical to administer this type of questionnaire during within-subjects studies where subjects use more than one system. The assessment is usually focused on the subjects' experiences using the system to complete a number of tasks and represents an overall assessment of a particular system.	The questionnaire is administered after subjects finish using a system. Subjects complete one questionnaire for each system.
Exit	If the study is a between-subjects study, then this questionnaire functions similarly to the post-system questionnaire. However, for within-subjects studies, this questionnaire can be used to elicit cross-system comparisons and ratings.	As its name implies, it is typically administered at the end of the study.

a danger that the questions and response sets will bias subjects since appropriate topics and responses are suggested. Data elicited by closed questions is homogenous and easier to analyze, but it is important to remember that the numeric scales are not based on a true number line, but instead represent a set of labels (albeit sometimes ordered). They are subject to individual interpretation and can be difficult to compare; there is no true zero and one subject's six may be another subject's five. Open questions take longer to administer than closed questions and responses are more difficult to interpret and analyze. People typically use different words to describe the same things and some subjects are better at clarifying and explaining their responses than others.

9.7 Interviews

Few IIR evaluations consist solely of interviews, but interviews are a common component of many study protocols. Most often, the interview is used as a delivery mode for a set of open-ended questions which might just as easily be delivered via print or electronic questionnaire. Although an interview mode is not necessarily required to ask such questions, it presumably allows one to get more individualized responses and allows some flexibility with respect to probing and follow-up. Kelly et al. [165] compared subjects' responses to a set of open-ended questions across three modes: interview, pen-and-paper, and electronic and found that while subjects' responses were longer in the interview mode than in the other two modes, the number of unique content-bearing statements they made in each mode were about equal. These results do not mean that the interview mode is not useful, but rather that it may not be useful for certain types of questions. The researchers asked questions that were similar to those asked in traditional IIR evaluations, which really are not designed with depth-interviewing in mind. If one were interested in asking more complex, abstract questions then it is likely that the interview mode would be more appropriate.

Another place where interview techniques can be used in IIR evaluations is during stimulated recall. During stimulated recall, subjects verbalize their decision-making processes and thoughts while watching

a video recording of a search they recently completed. During this process, the researcher might interrupt with specific pre-planned questions. These questions might be used to find out about something specific, or to probe remarks or actions made by subjects.

When planning to conduct interviews, researchers must first decide what type of interview they would like to administer: structured, semi-structured or open. This will be dependent on the purpose of the research. Researchers typically create an interview schedule which is a list of all of the questions they would like to discuss with the subject. In some cases, the researcher might go through this list of questions one-by-one in the same order for all subjects. This is typically the case with the short interviews that are conducted at the end of an IIR study. When the interview is the primary method used in a study, it is common to skip around and diverge from the list of questions. In these types of interview situations, the researcher has more flexibility and subjects, in many ways, direct the interviews since they can determine what topics will be discussed and in what order. The interview schedule functions to remind researchers of all of the topics that they would like to cover, but it usually does not determine the sequence of questioning. Along with an interview schedule, a researcher might also use other types of stimuli to facilitate and structure the interview. For instance, screen shots of an interface can be used to provide a focal point for discussion or even a video recording of a confederate using the experimental system.

9.8 Evaluation of End Products

A final method of collecting data during IIR evaluations focuses on the outcome or product of the search. Very often the focus is on the resolution of the work task, as opposed to the search task [in Byström and Hansen's [44] language]. This approach is used less frequently because it is more time consuming for subjects since it requires them to complete an additional and more complex task beyond finding relevant documents. The additional task is to actually use the information in a way that matches the user model behind the search task. For instance, the user model behind the traditional IR search task is a

person collecting materials that can be used to generate a report about a particular topic. The methods discussed in this section ask subjects to use the documents they find to create papers, reports, or other end products. These end products, in turn, are studied along with the subject's interactions with the system. Example studies that have used such methods include Egan et al. [83], Halttunen and Järvelin [116], Kelly et al. [166], Marchionini and Crane [192] and Vakkari [280].

From a philosophical point-of-view, this approach differs from others discussed in this section in that what is being evaluated and studied extends beyond the system. The IR system is viewed as a tool that helps people accomplish some other goal; thus, IR is not viewed as an end unto itself, but as an activity that supports a larger goal. The approaches discussed here focus on examining that larger goal. The underlying notion is that a better IR system will help people do a better job of achieving this goal (e.g., write better quality reports).[4] These approaches also assume a different perspective of relevance. Specifically, a distinction is made between the differences in relevance behavior exhibited during information finding and information use. The notion is that subjects engage in a different kind of relevance behavior when they are selecting relevant documents than when they are making choices about what to actually use and include in the final product.

The methods described in this sub-section are all executed within the context of a traditional IR task model where people engage with an IR system to find documents that support the writing of a report. Some researchers have studied students who are working on a writing assignment for a course, while others have included report writing as part of the study protocol. The former approach requires coordination with an instructor and it may not be possible to use an experimental IR system for such critical tasks. The latter requires the researcher to create a writing task, which might be artificial since the subjects are not performing it for any purpose other than the study. The difficultly in both cases is that many things contribute to the end product, including a person's writing and organizational skills. However, if two or more

[4] It may become increasingly difficult to analyze the system's contributions as subjects are able to compensate for poorly performing systems [243].

groups are being studied (e.g., if there are two systems being studied), then randomly assigning subjects to groups should distribute subjects with differing skills equally.

IIR researchers have used several approaches to study end products: examination of references, expert assessments and cross-evaluation. Examination of references is the most basic way to evaluate the extent to which the documents found by the user during searching were used to complete the larger task. The reference list, of course, does not tell the entire story. It is likely that many documents used during other information-seeking stages do not make it into the reference list. These documents play an important role in the creation of the final product, but they are not visible from the reference list.

Researchers have also asked experts to assess the quality of the final products [280, 282]. If the study is conducted in conjunction with a class, then this expert might be the course instructor who assigns grades to the final products. Course instructors will be experienced evaluators of the end products and will likely have established grading rubrics (even if they are internalized). If the evaluators are not very experienced, then the researcher will need to develop a rubric and spend time training evaluators to use it consistently. While these final products are a more realistic representation of how the information is used, it is difficult to coordinate and administer this type of assessment.

Cross-evaluation was developed as both a method and tool to facilitate comparison and rating of reports generated by subjects of information systems [257]. It requires subjects to perform three activities: use an information system to find relevant information, create a short report summarizing their findings, and evaluate other subjects' reports. Reports are evaluated according to seven quality criteria related to aspects of the information contained within the reports as well as aspects of the report itself. Cross-evaluation was developed in the context of interactive question–answering systems with intelligence analysts as subjects, but it could be extended to other types of situations. Interestingly, cross-evaluation can be a motivator for subjects since their work will be reviewed by others.

10

Measures

Measurement is fundamental to IIR research, but there are few research programs dedicated exclusively to the development and evaluation of measures for IIR. A large number of measures have been used but, at least until 1992, most could be categorized as relevance measures, efficiency measures, utility measures, user satisfaction, and success measures [254]. Su (2003) conducted a second review of evaluation measures in 2003 in preparation for an evaluation of search engines. Similar classes of measures were found except that success was replaced with connectivity. In her review of IIR research, Su (2003) further identified the following classes of measures for characterizing subjects and their information needs and behaviors: background (e.g., professional field, age, and sex); experience (e.g., use of IR systems and use of the internet); and information needs/search requirements (e.g., search topic, purpose of search, and time period of documents). Yuan and Meadow [299] also reviewed the research and created a classification of variables used in IR user studies. Classes included variables related to the study participants, searches and outcomes. Variables related to searches included an extensive list of items that ranged from specific

search tactics to performance measures. Boyce et al. [39] also compiled a list of measures in information science research.

Over time, four basic classes of measures have emerged as the standard: contextual, interaction, performance, and usability. The first set of measures includes those used to characterize subjects: such as age, sex, search experience, personality-type, and those used to characterize the information-seeking situation: such as task-type and subjects' familiarities with topics. Also included in these measures are geographic location and time. These measures basically describe the context in which information search occurs. It is beyond the scope of this paper to list every possible measure that fits this description. Ingwersen and Järvelin [139] provide an extensive discussion of context in information seeking and retrieval, while Dourish [78] provides a theoretical examination of the concept. Contextual measures in IIR evaluations can be elicited via questionnaires (e.g., age, sex, and topic familiarity) or controlled by the researcher (e.g, task-type). Many of these measures can be characterized as socio-cognitive measures or individual difference measures. While researchers have discussed context for many years and the difficulties with defining and measuring it (e.g., [6, 58, 75]), recently there have been large efforts in IR and IIR to systematically incorporate context into retrieval and evaluation (see e.g., [38, 227]).

The second set of measures includes those used to characterize the interaction between the user and the system and the user's search behaviors, such as number of queries issued, number of documents viewed and query length. These types of measures are typically extracted from log data.

The third set of measures are performance-based measures related to the outcome of the interaction, such as number of relevant documents saved, mean average precision, and discounted cumulated gain. These measures are also typically computed from log data.

The final set of measures includes those based on evaluative feedback elicited from subjects. Such measures often probe subjects about their attitudes and feelings about the system and their interactions with it. Although this class of measures is referred to as *usability* for simplicity sake, this class of measures includes a variety of self-report measures.

In IIR, performance measures have typically been separated from usability measures, even though as we will see shortly, *effectiveness* and *efficiency* are standard dimensions of usability and are often measured in HCI and ergonomics research with measures such as recall, task completion, error rate and time. Thus, in HCI performance is most often subsumed under usability, while in IIR we treat performance as a separate entity from usability and usually use the term *usability* as a synonym for self-report measures. The likely reason for our unusual use of this term is that performance measures have always been part of IR evaluation even before it was common to include subjects in evaluations. As it became more common to study subjects, various self-report-based usability measures made their way into the evaluation literature, but performance was still a first-class measure. Many early IIR evaluations discussed measures such as satisfaction and user-friendliness without mentioning the term usability. Common dimensions of usability are *effectiveness*, *efficiency*, and *satisfaction*, and what typically happens in IIR is that self-report measures are used to elicit evaluative data from subjects about these qualities. Again, this is contrary to how usability is measured in HCI — effectiveness and efficiency measures often consist of objective measures (e.g., some of our standard performance measures), while satisfaction measures are elicited via self-report.

Regardless of how one labels such measures, the most important thing is to be clear about the definitions of the measures. Devising appropriate measures involves the provision to two basic types of definitions: *nominal* and *operational*, which were discussed in a previous section. Nominal definitions state the meaning of concepts; for example, one might define performance as the ability to find relevant documents. Operational definitions specify precisely how a concept (and its dimensions) will be measured. For instance, an operational definition of *learnability* might include three questions and a five-point Likert-type scale for responding. Alternatively, one might operationalize learnability as the length of time it takes a user to learn to use an interface. Without both nominal and operational definitions it is impossible to know exactly what concepts researchers hope to capture with their measures, and it impossible to evaluate the credibility of the results.

There are also many validity and reliability considerations that are related to measurement, which are discussed in Section 12. Historically, IIR researchers have not been as concerned about measurement validity and reliability as other researchers who rely on self-report measures and questionnaires to elicit data from human subjects. Thus, studies that use self-report metrics which have not undergone serious scrutiny with respect to validity and reliability are susceptible to measurement error caused by the items themselves. Measurement bias is well-documented in many literatures including psychology, public opinion polling and public health. There is an entire field called psychometrics, which is devoted to understanding how to better measure psychological phenomena. This research has shown that people exhibit a number response biases when completing self-report measures, including inflation where the tendency is to rate things more positively than they are, and acquiescence where the tendency is to agree with everything.

Studies have also shown that slight variations in study procedures can impact study results and that method variance, in general, is almost always a potential problem [215]. Method variance refers to variance that is attributable to the measurement method rather than to experimental stimuli [94]. Systematic error variance, of the kind that might be caused by invalid or unreliable measurement techniques, is particularly problematic since it can produce results that might appear meaningful, but are only a function of the measurement technique. Method variance is a threat to any study regardless of whether self-report data are being elicited, but it is a more acute problem for studies that rely on self-report data. It is also problematic when no serious attempt is made to generate valid and reliable measurement techniques and the accepted standard is to just generate measures, especially self-reported ones, in an *ad-hoc* fashion.

Finally, the selection and interpretation of any measure, and particularly performance measures, should be grounded by the purposes of the system and the task the user is trying to accomplish. If a user is asked to complete a high-precision task such as finding one or a small number of documents that answer a particular question, then assessing recall makes little sense, since it is not appropriate to the retrieval

situation. If a system is designed to support exploratory search, then more interaction might be better than less.

10.1 Contextual

The measures presented in this section describe the context in which information search and interaction occurs. These include measures used to characterize subjects, such as age and sex, and those used to characterize the information need, such as task-type and domain expertise. It is common in most IIR evaluations to elicit some basic measures to describe study subjects, tasks and information search situations, but these measures are not always used as independent variables.

10.1.1 Individual Differences

Boyce et al. [39, p. 202] state, "the purpose of measuring user characteristics separately from the search process is to be able to use them to predict performance or to explain differences in performance." Such differences are often referred to as individual differences. Borgman [31] and Dillon [76] provide overviews of individual difference research. Borgman's article is focused on individual differences in information retrieval. Dillon's article is targeted to HCI researchers and strongly grounds individual differences in psychology research. Individual differences research was very popular in IIR during the period 1980s to 1990s, but has not received as much attention lately.

Fenichel [88] provides an overview of some of the more common measures of individual differences in the context of online searching. These include variables such as sex of subject, age, college major, profession, level of computer experience, and level of search experience. The latter two variables, in particular, do not figure as prominently in current research because there is not as much variability in these factors as there used to be, especially considering the typical subject in most IIR evaluations. Today, if one wanted to study many of these variables, one would need to purposively sample for them. However, there are still expected differences between many groups of people — for instance, one would expect subjects with advanced degrees in library science, computer science or human–computer interaction to be different from

subjects from the general population [90, 237, 238]. Therefore, it is important to report such characteristics, even if they are not independent variables, and carefully consider how they might impact the study results. Morahan-Martin [198] reviews research related to sex differences and internet usage, while Ford et al. [100] and Lorigo et al. [185] investigate sex differences related to information search. Ford et al. [99] investigate internet perceptions and cognitive complexity as additional ways to measure individual differences.

Another set of individual difference measures are those related to intelligence, creativity, personality, memory, and cognitive style. One nice thing about studying these types of measures is that there are a large number of standardized instruments found in the education and psychology literatures. Cognitive style, in particular, continues to receive a great deal of attention. Cognitive style is related to how people think about and approach problems. Ford et al. [100, 99], citing Riding and Cheema [216], state that there are two basic aspects of cognitive style: the wholist-analytic style characterizes users according to whether they tend to process information in wholes or parts and the verbal imagery style characterizes users according to whether they are verbal or spatial learners. Example instruments for assessing cognitive style include the learning style inventory, the remote associates test, the symbolic reasoning test and the Myers Briggs type indicator (Borgman, 1987). Ford et al. [99] use the cognitive styles analysis and approaches to studying inventory (which categorizes people as engaged in deep learning, surface learning or strategic approach) to measure cognitive style. The effects of cognitive style on information-seeking behavior in mediated search situations have also been investigated by Ford et al. [101]. Finally, computer and search self-efficacy have been studied as more refined measures of computer and search competency [57, 71].

10.1.2 Information Needs

Another important set of contextual variables are those that characterize the information need. Example measures include those related to the task such as task-type, task familiarity, task difficulty and

complexity, and those related to the topic such as a topic familiarity and domain expertise.

One problem with studying many of these items is that it is difficult to devise instruments for measuring them. For instance, topic familiarity is often measured with a seven-point scale, which does not really provide much information about how much a person knows about a topic. Since such scales are not calibrated it is even more difficult to make comparisons across familiarity levels. Domain expertise is often measured using credentials — for instance, a person with an advanced degree in molecular biology might be said to have high domain expertise in this subject area. Again, such coarse classifications often make it difficult to interpret study results and reach conclusions.

Other attributes of the information need that are often measured include persistence of information need, immediacy of information need, information-seeking stage, and purpose, goals and expected use of the results. Ingwersen and Järvelin [139] provide an extensive review of many of these types of variables.

10.2 Interaction

Interaction measures describe the activities and processes that subjects engage in during IIR. These measures are basically low-level behavioral data — such behaviors might originate from the subject or the system. Some types of interaction measures are common to almost all IIR evaluations. For example, number of queries, number of search results viewed, number of documents viewed, number of documents saved, and query length. Other types of interaction measures are specific to the individual system being studied. Many interaction measures are frequency counts of the activities that occurred and can be related directly to interface functionality. This includes basic *uptake* measures that show with what frequency subjects are using a feature or application.

Since most interaction measures are counts, they are continuous data types and can be combined to form other measures. For instance, time can be divided by the number of documents saved, or the number of documents saved can be divided by the number of documents

viewed. Usage patterns can also be derived from interaction measures. For instance, Markov modeling can be used to determine the most probable sequences of actions or a researcher can attempt to assemble low-level interactions into search tactics and strategies [15, 90].

One of the most challenging aspects of using interaction measures is developing a framework for interpreting them. Relating these signals to concepts requires one to consider the purpose and nature of what is being studied. If a subject enters a large number of queries, is this good or bad? The answer to this question is likely related to the purpose of the system — if the purpose of the system is to help a subject learn more about a topic, then more queries might be a positive indicator. If the purpose of the system is to help a subject find a single answer, then more queries might be a negative indicator.

Thinking more broadly about interaction, another important question to ask is what is interaction? Interaction with computers has been studied in a number of disciplines and there is not necessarily any agreement over what it means. There has been few serious theoretical discussion of the concept of interaction in IIR (e.g., [16]). In IIR, an implicit definition of interaction is accepted, which is very closely tied to feedback. Spink [248] and Spink and Losee [250] provide extensive discussions of the nature of feedback and identify interactive feedback units, the smallest of which consists of the user responding to the system and the system using the user's response to produce new content.

10.3 Performance

While there are many well-established measures for classic IR performance evaluation, there are not too many for IIR. As a result, many IR measures are used in IIR evaluations. The fit is not always perfect and it is probably safe to say that most researchers are not completely satisfied with these measures. Nested within most of these performance measures is another measure — relevance, and this is where things often break in an IIR evaluation scenario. The major problem is that most classic IR measures were developed and evaluated under different retrieval circumstances where certain assumptions could be made about relevance judgments and behaviors. These assumptions

underlie many of the evaluation measures developed as part of TREC, so the applicability and usefulness of these measures to IIR evaluations can be questioned. When using a TREC collection in an IIR study, researchers must assume for the purposes of the study that relevance is binary (usually), static, uni-dimensional and generalizable. Although these assumptions are incongruent with a lot research and with what most researchers believe, it is a *suspension of disbelief* that is required to use the TREC collection. Although there have been some attempts to create test collections with graded relevance assessments (e.g., [245]), this has been the exception rather than the rule. As mentioned previously, most standard IR performance measures assume binary relevance and do not easily accommodate situations where there are graded relevance assessments.

There have been a series of studies that have compared results of batch-mode and user studies and found that systems which perform better in batch-mode studies do not always do so in user studies [129, 243, 275, 276]. There are a number of explanations of these findings and many relate to the nature of relevance. Specifically, users often discard documents that TREC assessors have found relevant and find and save documents that TREC assessors never evaluate. Since it is the TREC assessor's judgments that are used to evaluate system performance, conflicting results are possible depending on how performance is evaluated. Another issue to consider when using TREC-based performance metrics in IIR evaluations is whether the metric is actually meaningful to real users. A measure that evaluates systems based on the retrieval of 1000 documents is unlikely to be meaningful to users since most users will not look through 1000 documents. Furthermore, it is important to note the limitations of performance measures: these measures can show that a system is functional (if used in a systems-centered evaluation), but not necessarily usable.

Finally, the actual technique used to measure relevance can vary considerably. There have been a number of studies that have examined and studied (1) the concept of relevance (e.g., [33, 197, 224, 232, 235, 236]); (2) the criteria users employ when making relevance assessments (e.g., [271]); and (3) techniques for measuring relevance (e.g., [84, 162, 272, 283]). Suffice to say, the published research about how users make

relevance assessments and the actual measures that researchers employ to collect relevance assessments are not very aligned.

10.3.1 Traditional IR Performance Measures

Table 10.1 presents some of the classic IR performance measures. For more information about these measures and descriptions of other IR measures, see Voorhees and Harman [288].

10.3.2 Interactive Recall and Precision

The measures in Table 10.1 are based on an assessor's relevance judgments and batch-mode retrieval runs that can consist of up to 1,000 documents. In IIR evaluations, subjects usually are unable to search through 1,000 documents. It is also the case that subjects are typically instructed to save documents that they find relevant and as described earlier subjects' relevance judgments often do not agree with the assessor's relevance judgments, so using the benchmark relevance judgments to assess performance may not be meaningful. Some TREC topics have hundreds of documents that have been marked as relevant by assessors and it is unlikely in most situations that a subject will search long enough to find all of these documents.

To partially account for the mismatch between TREC relevance judgments and subjects' relevance judgments, Veerasamy and Belkin [284] and Veerasamy and Heikes [285] proposed the use of interactive recall and precision, and interactive TREC precision which compares TREC relevant documents with those saved by subjects (Table 10.2). *TREC relevant* means the document was marked relevant by an assessor. These measures basically try to account for the two-stage relevance process that happens in IIR evaluations that use collections with relevance judgments: first, an assessor makes a relevance judgment and then a subject makes a relevance judgment. Documents that the assessor marks as relevant may or may not be retrieved, viewed or saved by subjects.

Relative relevance (RR) is a measure for comparing the degree of agreement between two relevance assessments [33, 36]. This might be between the system's relevance score and a subject's or between an

Table 10.1 Some classic IR evaluation measures.

Measure	Description
Recall	The number of retrieved relevant documents divided by the number of relevant documents in corpus.
Precision	The number of relevant retrieved documents divided by the number of retrieved documents.
F-measure	The F-measure is a way of combining precision and recall and is equal to their weighted harmonic mean $[F = 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})]$. The F-measure also accommodates weighting of precision or recall, to indicate importance.
Average precision (AP)	Individual precision scores are computed for each relevant retrieved document (with 0 assigned to relevant documents that are not retrieved). These values are then summed and divided by the total number of relevant documents in the collection. Thus, AP has a recall component to it and is typically described as the area underneath the precision/recall curve. AP also takes into account the position of relevant documents in the result list.
Mean average precision (MAP)	This is a run level measure and consists of taking the average of the average precision values for each topic.
Geometric average precision (GMAP)	The geometric mean of n values is the nth root of the product of the n values. Robertson [219] recommends taking the logs of the values and then averaging. GMAP was developed for the TREC Robust Track, which explored retrieval for difficult topics and does a better job than MAP of distinguishing performance scores at the low end of the AP scale.
Precision at n	The number of relevant documents in the top n results divided by n. Typical values for n are 10 and 20, which is thought to better represent the user's experience since research has shown that this is the extent to which users look through Web search results [146].
Mean reciprocal rank (MRR)	This measure was developed for high-precision tasks where only one or a small number of relevant documents are needed. For a single task with one relevant document, reciprocal rank is the inverse of its ranked position. MRR is the average of two or more reciprocal rank scores (used when there is more than one task).

assessor's relevance scores and a subject's. Thus, this measure attempts to accommodate the subjective nature of relevance and provide a measure of the extent of this overlap. RR might also be used to evaluate the extent to which assigned information need descriptions are

Table 10.2 Modified versions of recall and precision for interactive IR [284, 285] and relative relevance [33, 36].

Measure	Description
Interactive recall	Number of TREC relevant saved by user/number of TREC relevant documents in the corpus.
Interactive TREC precision	Number of TREC relevant documents viewed by the user/total number viewed.
Interactive user precision	Number of TREC relevant documents saved by the user/total number saved by the user.
Relative relevance (RR)	Cosine similarity measure between two lists of relevance assessments for the same documents.

well- or ill-defined — presumably there will be less overlap in relevance judgments for ill-defined needs.

10.3.3 Measures that Accommodate Multi-Level Relevance and Rank

Two other problems with traditional performance measures is that they only accommodate binary relevance assessments and they do not take into account that relevant documents that are retrieved further down on the results list are less useful since subjects are less likely to view them. Not only must a subject expend some effort to get to these documents, by the time the subject arrives at the document its content may be less valuable because of what the subject has learned on the way to the document. Traditional measures such as precision and recall do not accommodate this situation. Two results lists — A and B — may have the same recall and precision scores, despite having very different orderings of the documents. MAP was created to address the ordering problem in systems-centered research, but it still maintained some of the problematic assumptions of the traditional TREC measures.

Järvelin and Kekäläinen's [148, 149] suite of cumulated gain measures and Borlund and Ingwersen's [36] ranked half-life measures are measures that have been created for use in interactive search situations where human searchers make relevance judgments (Table 10.3). In addition to these measures, Borlund [33] identifies several others that account for position of relevant documents including Cooper's

Table 10.3 Cumulated gain measures [148, 149] and ranked half-life [33, 36].

Measure	Description
Cumulated gain (CG)	Cumulated gain can be computed at different cut-off values for search result of lists of varying sizes. At the cut-off point, CG is the sum of the relevance values of all documents up to and including the document at the cut-off point.
Discounted cumulated gain (DCG)	Discounted cumulated gain discounts the value of relevant documents according to their ranked position. New relevance values are computed by dividing the relevance score of a document by the logarithm of its rank. The discounted relevance scores are then summed to a particular cut-off point.
Normalized discounted cumulated gain (nDCG)	The DCG measure is normalized according to the best DCG available for a given results list. This normalization transforms DCG scores, which can take on a large range of numbers, to a 0–1 scale, which is easier to interpret and compare.
Ranked half-life (RHL)	The point in the results list at which half of the total relevance value for the entire list of documents has been achieved. If binary assessments are used, this is the point at which half of the relevant documents in the list have been observed. If multi-level assessments are used, this point is when half of the sum total of all of the relevance values are observed.

[59] *expected search length*, Dunlop's [81] *expected search duration* and Losee's [186] *average search length*. Käki and Aula [158] also propose *immediate accuracy* which is the proportion of cases where subjects have found at least one relevant result by a particular cut-off position in a ranked list of results. Kekäläinen and Järvelin [162] also extend traditional precision and recall metrics to accommodate graded assessments.

Discounted cumulated gain is based on the notion that the lower a document's rank in a results list, the less likely the subject is to view it. For instance, the chances of a subject viewing a document ranked in position one in a search results list is greater than the chances of the subject viewing a document ranked 88th. The measure also assumes that the number of topically relevant documents in a corpus is likely to exceed the number of documents a subject is willing to examine [162]. This measure also allows for multi-level (or graded) relevance assessments, which makes it more versatile and reflective of how most subjects make relevance assessments. To compute cumulated gain the

search results are first viewed as a vector where each document is represented by its relevance value. The cumulated gain of each document is a function of its relevance value plus all of the relevance values of the documents ranked above it.

To compute discounted cumulated gain, the relevance value of a particular document is treated as a function of a document's relevance and its rank. The *discounted* part of the measure reduces the contributions of relevant documents that are ranked lower in the list. This is accomplished by dividing the relevance of the document by the log (base 2) of its rank.[1] For instance, assuming four levels of relevance, where four represents a highly relevant document, a highly relevant document at rank 16 would contribute a score of 1 [$4/\log_2(16)$, or $4/4$]. The base of the log can be adjusted to match varying types of users. For instance, patient or impatient users. DCG scores can also be normalized (nDCG) based on an ideal result list which can be created by ordering all documents judged from most relevant to least relevant. Cumulated gain-based measures are typically reported at particular cut-off values. In Kekäläinen and Järvelin [162] the use of graded relevance assessments (non-binary based) are extended to a number of other evaluation measures as well. Järvelin et al. [150] extended DCG for use in multi-query sessions.

Rank half-life (RHL) measures the extent to which relevant documents are located at the top of a ranked results list [33, 36]. This measure is similar to MAP and DCG in that not only is the relevance values of particular documents included in its calculation, but also their position in a ranked list. There are two measures associated with RHL — RHL and RHL-index. The RHL is the position at which half of the relevant documents are retrieved. If multi-level relevance values are used, this is the point at which half of the total relevance value for the entire list of results has been observed. The formula used to calculate RHL is the basic formula for the median of continuous data. Thus, lower RHL are associated with better retrieval performance since lower numbers indicate that more of the relevant documents were found

[1] The discount is not applied to the document in position one of the search results list since the logarithm of this would result in a denominator of zero.

near the top of the results list. The RHL-index allows one to compare two result lists at a particular cut-off value, given a precision value. The RHL-index is the RHL of the list divided by the precision of the list.

10.3.4 Time-Based Measures

Time has been used quite a lot in IIR evaluations, both at a gross level (e.g., the length of time it takes a subject to complete a search task) and a more specific level (e.g., the length of time a subject spends viewing a search result or engaging in a specific action). As mentioned previously, time-based measures can be difficult to interpret since this is dependent on the task, the objective of the system and the researcher's beliefs about IIR. Time-based measures are often used as indicators of efficiency, although as stated earlier, effectiveness (performance), efficiency and satisfaction can be separated from one another. Efficiency and time will be discussed again in Section 10.4 when traditional notions of usability are discussed along with evaluative self-report measures.

Researchers have used a variety of time-based measures, including the length of time subjects spend in different states or modes, the amount of time it takes a subject to save the first relevant article, and the number of relevant documents saved during a fixed period of time. The number of actions or steps taken to complete a task is another way to look at time and efficiency. Käki and Aula [158] formalize two time-based measures that have been used in IIR research, search speed and qualified search speed (Table 10.4). These measures are based on answers not relevant documents, but could be extended to cover this retrieval unit.

Table 10.4 Time-based measures from Käki and Aula [158].

Measure	Description
Search speed	The proportion of answers that are found per minute. This measure consists of dividing the total number of answers found by the length of time it took to find the answers. All answers are included in this computation regardless of whether they are correct.
Qualified search speed	This measure accommodates multi-level relevance and consists of computing search speed for each relevance category, including non-relevant.

Although it is more common to consider how long it takes subjects to perform particular actions, the length of time it takes the machine to perform particular actions is also a common time-based efficiency measure in IR. Most would agree that this impacts a subjects' experience with a system and likely contributes to their evaluation of the system. Cleverdon et al. [55] discuss the response time of the system, which can be measured with a simple time-based figure, or could be an analysis of computational complexity. Common measures of computational complexity are number of steps, iterations or computing cycles that are needed by the computer to perform a task and the amount of computing resources needed.

10.3.5 Informativeness

Informativeness is a measure of the output of a system proposed by Tague [258, 261, 263]. This proposed method for evaluating search results focuses on relative evaluations of relevance rather than absolute measures. The assumption behind this is that asking subjects to rank a set of search results from most informative to least informative results in more accurate data than asking them to associate absolute judgments with each result using a scale. While Tague [263] wrote quite a bit about the informativeness measures and explored this measure in the context of browsing, a large-scale validation of this measure was never achieved due to her death [103]. Freund and Toms [103] recently re-introduced this measure and explored it in the context of Web search. Interestingly, there are many current proposals to use relative relevance judgments to evaluate search results lists (e.g., it is generally accepted that clicks equal relevance (whether right or wrong)). Perhaps with this renewed interest in relative relevance judgments, Tague's informativeness measure will finally be validated and adopted as a standard method of evaluation.

10.3.6 Cost and Utility Measures

In the early days of IIR research, cost and utility measures figured prominently in the IR evaluation framework. Some researchers treated these measures as separate constructs from relevance, while

others attempted to use these measures as substitutes. Cooper [60] proposed the use of subjective utility as the benchmark with which systems should be evaluated. In his proposal, users associated dollar amounts with search results. Salton [230, p. 442] summarizes the utility-theoretic paradigm, "retrieval effectiveness is measured by determining the utility to the users of the documents retrieved in answer to a user query". Salton [229] identifies a host of cost-based measures including those associated with the operational environment, response and processor time.

Belkin and Vickery [23] identified utility as one of the major approaches underlying performance measures alongside relevance and user satisfaction. In a study of evaluation measures for IIR, Su [254] compared 20 measures, including actual cost of search and several utility measures such as worth of search results in dollars, worth of search results versus time expended and value of search results as a whole. Su [254] found the value of the search results as a whole was the best single measure of IIR performance. The 40 subjects involved in this study were responsible for the costs of their own searches which likely changed the importance of this variable to them.

The popularity of utility measures in the early years is not surprising since users were charged to use many operational IR services. Utility and cost functions have always been an important part of the evaluation of library and information services (e.g., [239]). Even though people are still charged to access databases and view the full-text of articles, this cost is usually incurred by the user's institution, thus the price of information services are often out of users' awareness, despite continuing to be an important issue for institutions. Furthermore, since so much information is freely available online these types of measures are arguably less relevant to the individual user. It is likely the case that when using IIR systems, most users are not thinking about the costs associated with the service or information, at least not in terms of monetary values.

Recently, Lopatovska and Mokros [183] investigated willingness to pay and experienced utility as potential measures of the affective value of a set of nine Web documents. While the results are limited given the small number of documents evaluated, subjects' responses seemed to

suggest that willingness to pay reflected the rational value of the documents for completing the task, while experienced utility reflected an emotional, task-neutral reaction to the documents. This work suggests that these measures may continue to have value in IIR evaluations.

10.4 Evaluative Feedback from Subjects

Much of the data elicited during an IIR study is self-report, evaluative feedback from subjects. Very often researchers refer to these as *usability* measures, but this is not entirely appropriate. In many cases, researchers do not provide any conceptual or operational definitions of usability and instead lump all self-report data together and call it usability data. Referring to all self-report data as usability data overly-restricts the types of questions that can be asked and does not encourage much thought about the nature of the data that is collected. Also, as discussed earlier, traditional usability measures typically include objective performance measures, but in IIR, performance has been treated as a separate category because of its importance in classic IR.

HCI research has shown that there is only a slight correlation between objective and subjective (i.e., self-reported) performance metrics, and that people tend to use the high-end of the scale when evaluating systems (i.e., inflation) [133, 203]. Anecdotal evidence from IIR research also suggests this, and recently researchers in IIR are starting to look at this empirically [168]. Whether or not one believes that objective and subjective measures should be correlated is a theoretical issue that has not been explored in IIR. While it can be argued that response biases such as inflation are not so problematic as long as relative differences can be detected, this merely sidesteps the bigger problems of the validity and reliability of the measurement instrument. Furthermore, in cases where a single system is being evaluated such an argument does not hold.

10.4.1 Usability

To start, let us examine one of the most used conceptualizations of usability. The International Organization for Standards (ISO) [141, p. 2] defines usability as the extent "to which a product can

be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction". Thus, effectiveness, efficiency and satisfaction are identified as the key dimensions of usability. This definition also highlights the importance of carefully defining user and task models since it emphasizes the articulation of *specified* users and *specified* tasks. Nielsen [202] defines *usability* as "a quality attribute that assesses how easy user interfaces are to use,"[2] and divides the concept into five dimensions: learnability, efficiency, memorability, errors, and satisfaction. The ISO definition is arguably the most commonly used definition of usability and will be used in this paper. The ISO standard divides usability into effectiveness, efficiency and satisfaction. Definitions of these concepts are presented below.

- *Effectiveness* is the "accuracy and completeness with which users achieve specified goals." In other words, a tool is effective if it helps users accomplish particular tasks.
- *Efficiency* is the "resources expended in relation to the accuracy and completeness with which users achieve goals." A tool is efficient if it helps users complete their tasks with minimum waste, expense or effort.
- *Satisfaction* is the "freedom from discomfort, and positive attitudes of the user to the product". [141, p. 2]. Satisfaction can be understood as the fulfillment of a specified desire or goal. It is often the case that when people discuss satisfaction they speak of the contentment or gratification that users experience when they accomplish particular goals.

To understand the range of measurement techniques that researchers have used to study usability, Hornbæk [132] conducted a content analysis of 180 studies published in the core HCI journals and proceedings and classified the measures using the ISO definition of usability. Hornbæk [132] found that the way in which each of these concepts is operationalized varies according to study purpose. One important finding was that often multiple items are used to get at

[2] This definition is arguably not very good since *use* is the root of usability and it is used in the definition.

any one of these dimensions (this is common for all self-report data). For example, instead of a single item to measure satisfaction, several items are used to measure the construct in different ways. In general, measurement theory suggests the use of multiple items to measure a concept; single items are not considered reliable enough [105, 286]. The use of multiple items allows researchers to check for response reliability, as well as some types of validity.

10.4.1.1 Effectiveness

Hornbæk [132] found that the most common way that *effectiveness* has been measured in HCI studies is according to error rate and binary task completion. In IIR, the most common way to measure effectiveness is using traditional performance measures, such as precision and recall and to elicit self-report data from subjects about their perceptions of performance. In addition to error rate and binary completion, Hornbæk [132] documented several other measures of effectiveness including completeness, precision (ratio between correct information and total information retrieved), and recall (subjects' ability to recall information from the interface). While precision is operationalized the way it is in IR, recall is not. Also note that error rate, one of the most common measures, is not used that often in IIR except perhaps for fact-finding tasks, so there are some differences in notions of usability in HCI studies and usability in IIR evaluations. However, error rates with respect to interaction and interface use are worth capturing in IIR evaluations. Although many interfaces are quite simple, with more complex interfaces subjects may be observed making a number of errors, such as false clicks. Quality of outcome and expert assessment were also identified by Hornbæk; these are discussed in more detail later.

10.4.1.2 Efficiency

One of the most common ways to measure *efficiency* is to record the time it takes a subject to complete a task [132]. This includes measures of overall time, as well as more precise measures which document the amount of time subjects spend doing different things or in different modes. These types of measures were discussed previously

in Section 10.3 because like effectiveness, they are typically used in IIR evaluations as measures of performance. Usage patterns are also included in this set of measures, although in this discussion they have been separated out as interaction measures. In addition to these measures, questionnaire items regarding efficiency and time can be created to elicit subjects' perceptions.

Hornbæk [132] includes task difficulty as an efficiency measure where difficulty is typically determined by experts. While task difficulty is related to time-based measures of efficiency (i.e., presumably it takes longer to complete a difficult task), task difficulty was included in this paper as a contextual measure since IIR researchers are typically interested in seeing how difficulty impacts behaviors and performance. Moreover, task is a central focus in many IIR evaluations, so it is useful to separate it from efficiency since it is studied in a variety of ways in IIR. Learning measures are also identified by Hornbæk [132]. These measures use changes in efficiency as indicators of learning — e.g., subjects becoming faster at text input over time or subjects taking less time to complete subsequent tasks. These types of learning measures have not been used a lot in IIR evaluation.

10.4.1.3 Satisfaction

In the traditional conceptualization of usability, effectiveness, and efficiency measures are typically objective measures. The third dimension — satisfaction — attempts to gauge subjects' feelings about their interactions with the system. Hornbæk [132] identifies system preference as a measure of satisfaction, although in IIR evaluations system preference is typically treated as a separate construct and reported alongside satisfaction rather than as an indicator of satisfaction. Although multiple items are often used to assess satisfaction, a general question about satisfaction (e.g., how satisfied are you with your performance?) is also usually included as a questionnaire item. Specific satisfaction items might be asked for each different experimental feature of the system. Subjects' perceptions of outcomes and interactions are also commonly elicited. Examples include questions that ask about satisfaction with retrieval results and/or with the system's response time.

10.4.1.4 Ease of Use, Ease of Learning, and Usefulness

There are many other types of constructs that researchers try to measure in IIR evaluations that are related to usability. These include ease of use, ease of learning, and usefulness. Ease of use is considered by some as an indicator of satisfaction, but it can also be defined as the amount of effort which subjects expend executing and/or accomplishing particular tasks. Ease of use is closely related to efficiency: if a tool is not easy to use, then it is likely to result in inefficient use.

Ease of learning is related to the amount of effort subjects expend learning to use a system. Ease of learning items typically attempt to answer questions about how hard a system is to learn to use. A system may be effective and efficient, but if it is intended to have a human user and that user cannot learn to use it because it is too complex, then the system may not be successful.

Finally, usefulness is related to whether a tool is appropriate to the tasks and needs of the target users. The tool may be effective and efficient, but if users have no use for it, then it has little impact.

10.4.1.5 Available Instruments for Measuring Usability

There are several instruments that have been developed to measure usability. Each of these instruments has undergone different amounts of testing with regard to validity and reliability. Some of these instruments cost money while others do not. Some of these instruments are appropriate to IIR, but most contain a lot of items that do not make sense in the IIR evaluation context. Many of these questionnaires were developed in industry (and many are industry standards), but most questions are too general to provide enough detailed information about the IIR situation.

One of the most well-known instruments for measuring satisfaction is the Questionnaire for User Interface Satisfaction (QUIS) [53], which elicits evaluations of several aspects of the interface using a 10-point scale, including the subject's overall reactions to the software, the screen, the terminology and system information, and learning and system capabilities. The USE questionnaire [188] evaluates four dimensions of usability: usefulness, ease of use, ease of learning,

and satisfaction. Each dimension is assessed with a number of items which subjects respond to with a seven-point scale. Another commonly noted usability questionnaire is SUMI (Software Usability Measurement Inventory) [256]. SUMI consists of 50 items and provides subjects with three coarse responses: agree, do not know and disagree. Wildemuth et al. [295] used several validated usability measures in a TREC VID interactive project including Davis' [69] measures of perceived usefulness, ease of use and acceptance. Davis' work is from the management information systems (MIS) research, which has more examples of validated usability measures for information system evaluation.

10.4.2 Preference

In studies of two or more systems with a within-subjects design, it is common to collect preference information from subjects. This is perhaps one of the most basic types of data that can be collected, but often provides the clearest indication of subjects' attitudes about systems. Typically at least one open-ended follow-up question is asked to obtain more insight into subjects' preferences. Although not as common, preference data can also be elicited about individual aspects of the systems as well, such as method of display. Thomas and Hawking [270] propose an evaluation method that is based on preference. In this method, subjects are presented with a split screen each displaying search results from two different search engines. Subjects are asked to make holistic evaluations, basing their preferences on entire lists rather than individual documents.

10.4.3 Mental Effort and Cognitive Load

Mental effort is a construct that has been used as an indicator of efficiency, although the construct itself is quite complex [117]. This construct has been extensively investigated by those who study human computer interaction in the areas of control systems and airplane cockpits. Jex [151, p. 11] proposes the following definition of mental workload, "mental workload is the operator's evaluation of the attentional load margin (between their motivated capacity and the current task demands) while achieving adequate task performance

in a mission-relevant context". Hart and Staveland [122, p. 140], who developed the NASA-Task Load Index (NASA-TLX), define workload as "the cost incurred by a human operator to achieve a particular level of performance". Hart and Staveland [122, p. 140] go on to state that workload is "not an inherent property, but rather emerges from the interaction between the requirements of a task, the circumstances under which it is performed and the skills, behaviors and perceptions of the operator".

The NASA-TLX consists of six component scales, which are weighted to reflect their contribution to the workload according to the subject. These six scales are then averaged to produce an overall measure. The six factors that comprise the NASA-TLX are: mental demand, physical demand, temporal demand, performance, frustration and effort. A separate instrument is used for the weighting. This instrument elicits rankings of the *relative importance* of the six factors. This includes all possible pair-wise comparisons of the six factors. The number of times that a factor is rated as more important during the pair-wise comparisons indicates the relative importance of the factor. The NASA-TLX has been used in a few IIR evaluations. Most recently Kelly et al. [168] used this instrument to evaluate interactive question–answering systems and found that while it provided a good indication of mental effort for individual systems, it was hard to compare systems since the TLX is designed to distinguish among tasks, not systems.

In many traditional studies of mental workload a common method is to ask subjects to complete auxiliary tasks in addition to the primary task [151]. By manipulating the amount of cognitive load a person experiences with the auxiliary task and observing user performance, the belief is that it is possible to get an idea of what parts of the primary task are the most demanding, or alternatively, the most engaging. As the difficulty of the primary task increases or as the subject's engagement with the task increases, their performance on the auxiliary task should decrease because there are fewer cognitive resources available. Dennis et al. [72] employed the dual task technique in their study of interactive Web search. Dennis et al. [72] identified a number of possible auxiliary tasks and decided on a digit-monitoring task. The researchers found mixed results with respect to the usefulness of the

dual monitoring task, but the work is interesting in that it was one of the first uses of this technique in IIR. However, it is important to note that it is not always clear what differences in user behavior mean in these types of studies — increased effort on a task might mean the task is more difficult or that the user is more engaged. What can be said is that some tasks consume more attention that others and this can be good or bad.

10.4.4 Flow and Engagement

Although flow and engagement have not been used a lot in IIR research, they suggest additional ways that IIR systems and users' experiences can be evaluated. The notion of flow was proposed by Csikszentmihalyi [63] and is defined as a "mental state of operation in which a person is fully immersed in what he she is doing, characterized by a feeling of energized focus, full involvement, and success in the process of the activity." Csikszentmihalyi [63] identified five characteristics of the experience of flow and Csikszentmihalyi and Larson [64] presented an instrument and method for measuring flow in everyday life. Bederson [18] related flow to human–computer interaction and system evaluation. Pilke [213] conducted interviews to see if flow experiences occurred during IT use, while Chen et al. [52] looked at flow in the context of Web interactions. Ghani et al. [107] developed four seven-point semantic differential items to measure flow in human–computer interaction scenarios and further explored its relationship to task characteristics [106].

Engagement is a relatively new concept that has not yet been used to evaluate users' experiences with IIR systems. O'Brien and Toms [205, p. 949] define engagement as, "a quality of user experiences with technology that is characterized by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, and interest and affect". Thus, engagement is a multi-faceted construct that encompasses many characteristics of a user's experience. O'Brien and Toms describe how the theory of engagement shares some attributes with flow, aesthetic, play and information interaction theories, but state that it is fundamentally different from these theories. O'Brien and Toms [205] provide a conceptual framework

for defining user engagement and have also developed a procedure for measuring user engagement [206].

10.4.5 Subjective Duration Assessment

As described earlier, subjects often exhibit a number of response biases when responding to self-report measures which can make it difficult to obtain valid and reliable data. Motivated by the finding that subjective and objective performance measures are often uncorrelated and a belief that this indicates a potential problem with the validity of self-report measures, Czerwinski et al. [65] explore the use of time estimation as a way to obtain an indirect measure of task difficulty from subjects. Czerwinski et al. [65] call this method of estimation *subjective duration assessment* and propose a measure called *relative subjective duration*. After completing tasks in Czerwinski et al.'s [65] study, subjects were asked to estimate the length of time it took them to complete tasks. This estimation was then compared to the actual length of time it took them. Czerwinski et al. [65] found that subjects underestimated times associated with tasks with high success rates and overestimated times associated with tasks with low success rates. Relative subjective duration provides an alternative method to elicit estimations of subjects' perceptions of task difficulty. While this only represents one such measure, it suggests a possible useful approach to discovering other indirect ways of assessing subjects' experiences with IIR systems.

10.4.6 Learning and Cognitive Transformation

As described earlier, IIR is typically not a goal unto itself, but is often done in support of some larger goal or task. One goal of most IIR systems that is implied rather than explicitly stated is that the system will help users learn about a particular topic. This, of course, is accomplished through retrieving, reading and evaluating documents. However, measuring the extent to which learning has taken place is difficult because it would require the establishment of some baseline measure of how much a subject knows about a topic and a post-test measure to determine how much they have learned. One example study that attempted to do this is Hersh et al. [130]. Developing some standardized

instrument to evaluate how much a person knows is not easy and each different topic might require a different assessment technique.

Collecting assessments of interaction outcomes (e.g., a paper) is another approach to gauging the extent to which a system helped a person learn about a topic or accomplish a goal. This is not technically a self-report measure, but such assessments are usually generated by experts or other subjects who use standardized instruments to make assessments. As noted earlier, example studies that have attempted to assess final products include Egan et al. [83], Kelly et al. [166], Marchionini and Crane [192] and Vakkari [280].

11

Data Analysis

This section focuses primarily on quantitative data analysis, since much of the data collected in IIR evaluations is quantitative in nature. Some of the data collection techniques described above yield qualitative data, so qualitative data analysis will be presented briefly. This brevity is not meant to indicate that qualitative data analysis is easier or less important, but rather that there just is not enough space to provide a detailed exploration of this topic.

11.1 Qualitative Data Analysis

While there are numerous approaches that one might take in doing qualitative research — many of which differ epistemologically and philosophically (see, for instance, [142]) — this article focuses on two of the more common approaches to qualitative data analysis, content analysis and open coding. Since interviews only form a small part of traditional IIR evaluations, the two techniques described below should provide adequate background for analyzing this type of data. Those interested in reading more about different qualitative research traditions and analysis techniques are referred to Miles and Huberman [195],

Denzin and Lincoln [73], Charmaz [50] and Glaser and Strauss [108]. References from the information science literature include Bradley [40], Dervin [74] and Fidel [91].

The goal of most qualitative data analyses that are conducted in IIR is to reduce the qualitative responses into a set of categories or themes that can be used to characterize and summarize responses. Perhaps the most important message that can be communicated about qualitative data analysis is that it is not as easy as it seems. In general, reports of qualitative data analyses in IIR are weak and usually inadequate. One reason for this is that those unfamiliar with qualitative data analysis often do not bother to report important details about how the data were collected and analyzed. For instance, consider the following three scenarios:

(1) A researcher records and transcribes an interview. Analysis is based on the transcriptions.
(2) A researcher records an interview, but does not transcribe it. Instead the researcher listens to the recordings once and takes notes. Analysis is based on these notes.
(3) A researcher does not record the interview, but takes notes during the interview. Analysis is based on these notes.

All things being equal, the quality of the data the researcher captures as well as the researcher's interpretations of this data in each of these scenarios is likely to vary. In Scenario 1, the recording and transcription processes will result in the most faithful record of what occurred during the interview. Thus, analysis based on this data will likely be better than if these steps were not taken. In Scenario 2, an accurate recording of what occurred during the interview exists, but there is no guarantee that the researcher has done a good job noting what occurred, which of course, has implications for the validity of the subsequent analysis. In Scenario 3, the representation of what occurred will be limited to what the researcher can physically record and also to what the researcher thinks is important to record at the time of the interview. The selectiveness of this process will be reflected in the analysis. It is also the case that one's ability to conduct a good interview will be compromised

since one is engaged in both interviewing and note taking. The point of this example is not to say that all interviews should be recorded and transcribed, but rather to point out the potential differences of each method and illustrate the importance of reporting the method in its entirety even if it *seems* trivial.

Another reason that reports of qualitative data analysis are typically weak in IIR is because researchers often do not make appropriate distinctions among different analysis techniques. For instance, it is common for researchers to use the term *content analysis* as a synonym for qualitative data analysis. Researchers use this word in a generic sense — there is content and it needs to be analyzed — but content analysis actually represents a very specific data analysis technique. While content analysis and qualitative analysis have some things in common, they represent two very different approaches to analyzing textual data.

Content analysis is most often used to analyze recorded communication — books, films, email messages, Web pages, and advertisements. At its inception, it was intended as a quantitative method, although there are now a number of variations, interpretations and uses of content analysis [201]. Content analysis was originally executed in much the same way that IR is executed — by counting the occurrences of words and other features. Traditional content analysis starts with a somewhat well-defined and structured classification scheme, including categories and classification rules. The categories are usually mapped to variables. For example, if one were analyzing a set of transcripts for mentions of the concept of relevance, then one might use a pre-defined vocabulary as indicators of this concept. Before the analysis can start, the researcher creates a codebook that links together the concepts of interest, the categories that represent them and the classification rules. The coding process is more structured and deductive than what it is in qualitative data analysis.

Most researchers in IIR engage in a less structured form of data analysis when analyzing qualitative data. The goal is still data reduction, but the process differs from content analysis in several keys ways. The codes and categories are usually developed inductively during the analysis process as the researcher analyzes the data. This process is

referred to as *open-coding* [193]. Strauss and Corbin [253, p. 62] characterize open coding as, "the part of analysis that pertains specifically to the naming and categorizing of phenomena through close examination of data ... during open-coding the data are broken down into discrete parts, closely examined, and compared for similarities and differences". Codes are suggested by the researcher's examination and questioning of the data. This process is iterative; when new codes are added previously categorized data are reviewed to see whether they need to be reclassified. Coding ceases when saturation has been reached and all relevant utterances have been classified. While researchers typically develop rough heuristics for classifying data into different categories, these are not as well-formed as those in content analysis, which has implications for reliability. With content analysis, some type of inter-coder reliability should be performed to ensure that items have been coded consistently. With open-coding, this step is not always required, expected or possible, although it is assumed that the researcher is analyzing the data consistently and faithfully.

The two approaches discussed in this section are not the only approaches to analyzing qualitative data. In some ways they represent two ends of a continuum. On one end is content analysis, which is highly structured and emphasizes reliability, and on the other end is open-coding, which is more fluid and emphasizes flexibility. Both of these approaches, as well as all the ones in between, can be used to analyze qualitative data in IIR evaluations.

11.2 Quantitative Data Analysis

Quantitative data analysis is a large and complex topic. In this section, basic statistical tests are presented which are used in the common IIR evaluation model where a researcher is comparing two or more systems or interfaces (independent variable) using a set of outcome measures (dependent variables) that are categorical or continuous in nature. Reading this section will not make anyone an expert, but it will help readers distinguish among different types of statistics, select appropriate statistical tests and understand how some statistics are computed. The following books were consulted during the writing of

this material: Cohen [56], Gravetter and Wallnau [110], Myers and Well [200] and Williams [296].

A statistic is an estimate of an unknown value in the population; these unknown values are known as parameters. Statistics are derived from samples and provide *estimates* of the values of unknown parameters in the population. Descriptive statistics characterize variables; most notably these statistics describe central tendency and variation. Descriptive statistics are the basic inputs of inferential statistics. Inferential statistics are used to compare the relationship among two or more variables and to test hypotheses. Inferential statistics allow one to make *inferences* about population parameters based on sample statistics.

Inferential statistical tests are often performed in order to determine whether null hypotheses can be rejected and *significant* (or *statistically reliable*) relationships exist among variables. The word *significant* in the context of statistics has a specific meaning; *significant* is used to describe situations where particular probability values are observed. Thus, *significant* should not be used as a synonym for *large* or *important* when presenting and discussing results. If it is claimed in a research report that a *significant* relationship was observed, then one should be prepared to present the statistical tests supporting this claim.

The inferential statistics reported in this paper are commonly used in IIR evaluations. In particular, the focus is on tests that are used in evaluations to compare two or more systems among a set of outcome measures. The statistics reported in this section also focus on parametric statistics rather than non-parametric statistics. Parametric statistical tests assume that the variables that are being examined are normally distributed in the population (this will be discussed in more detail later). When it is assumed that the variable is not distributed normally in the population, then non-parametric tests can be used. Non-parametric tests use different descriptive statistics in their computation than their parametric counterparts. For instance, two medians might be compared instead of two means. When appropriate, an equivalent non-parametric test is suggested for each parametric test.

The research scenario in Figure 11.1 is provided to facilitate the presentation of material in this section. This scenario is modeled after

A researcher has developed two experimental IR systems and would like to test them against one another and a baseline. These three systems will be called System A (the baseline), System B and System C (note that *system type* functions as one variable, with three levels). Subjects are given six search tasks to complete which ask them to find documents that are relevant to pre-determined topics. Each subject completes two searches on each system (a within subjects design).

The researcher is interested in comparing the three systems using the measures listed below. These measures are organized according to the instrument used to collect the data.

Demographic Questionnaire

Sex of Subject [Male or Female]

Pre-Task Questionnaire

How *familiar* are you with this topic? [5-point scale, where 1= know nothing about the topic and 5=know details]

Post-System Questionnaire

Usability [5-point scale, where 1=strongly disagree and 5=strongly agree]: It was easy to find relevant documents with the system.

Exit questionnaire

Preference: Which system did you prefer? [System A, B or C]

System Logs

Performance

- Average session-based nDCG

Interaction

- Number of queries issued
- Query length

Fig. 11.1 IIR Research Scenario.

the archetypical IIR evaluation and compares three systems.[1] However, it should not be used as a self-contained model since it only represents sample variables and is necessarily incomplete. The different variables have been purposely selected to illustrate the statistics that are discussed.

11.2.1 Descriptive Statistics

The first step in analyzing data is to examine the frequency distributions of each variable. This is useful for identifying outliers and anomalies, and human errors that may have been made during the process of building the data files. It is also useful because it helps one understand the appropriateness of different types of statistics since this depends in

[1] *System* is used in a generic sense. A researcher may also be comparing two interfaces or interaction techniques.

part on the distributions. Frequency distributions present the number of observations of each possible value for a variable. For example, the frequency distribution for the familiarity variable will show how many times each of the five-points was used by subjects.

Six of the most common types of distributions are shown in Figure 11.2. These are the (a) normal distribution (also known as the bell-shaped curve and Gaussian distribution); (b) peaked distribution; (c) flat distribution; (d) negative skew; (e) positive skew; and (f) bimodal distribution. These curves represent general, theoretical shapes. The normal, peaked, flat and bimodal distributions are symmetrical and the negative and positive skews asymmetrical. While the distributions of some real data will match these shapes nearly perfectly, most distributions only approximate them.

There are two main measures for describing a curve's shape: skew and kurtosis. For each measure, a value of zero represents a normal curve. The skew measures can be best illustrated with the negative and positive skewed distributions. The skew measure will move in a negative direction when the distribution approximates a negative skew and positive direction when it approximates a positive skew. Kurtosis can be illustrated with the peaked and flat distributions. The kurtosis measure will move in a negative direction when the distribution approximates

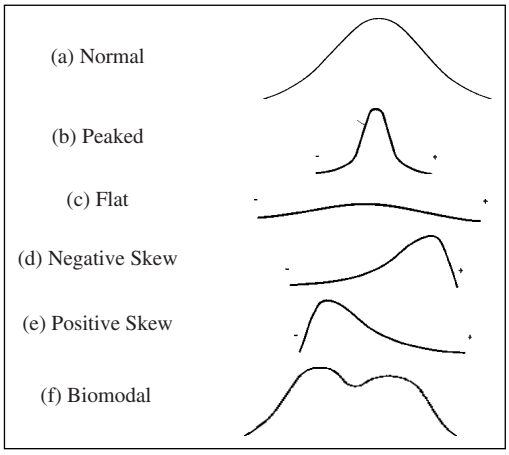


Fig. 11.2 Example distributions.

the flat curve (platykurtic) and a positive direction when it approaches the peaked curve (leptokurtic). A normal curve, which has a kurtosis of zero is said to be mesokurtic. These measures can be particularly useful when one is transforming or re-classifying data. For instance, if one decides to transform data collected using a seven-point measure into a three-point measure and is interested in having roughly equal numbers of observations for each point, then the kurtosis measure can be used to identify which of the candidate divisions of the seven-point measure results in the flattest distribution. Variables that are skewed can also be subjected to mathematical transformation in order to decrease the differences between points. For instance, a logarithm transformation can be performed (if appropriate) to minimize the differences between scores. We will discuss these distributions in more detail below, but first let us review some other descriptive statistics.

11.2.1.1 Measures of Central Tendency

There are two major classes of descriptive statistics: measures of central tendency and measures of dispersion. These statistics are useful for describing distributions so that the entire frequency curve does not have to be presented. Measures of central tendency describe how observations (or scores) cluster around the center of the distribution. There are three basic measures: mean, median, and mode. The definitions of each are provided below (Table 11.1). To illustrate these measures, let us assume that we have data from five subjects, each of whom submitted the following number of queries: 2, 1, 3, 5, and 1. The central tendency measures for this sample are: Mean = 2.4, Median = 2 and Mode = 1.

For the normal curve these measures are equal, but in distributions of different shapes the measures are not equal and can give misleading

Table 11.1 Measures of central tendency: mean, median and mode.

Central Tendency	
Mean	The sum of the scores in a distribution divided by the total number of scores. $\overline{X} = \frac{\sum X}{n}$
Median	The score that falls in the middle of the distribution.
Mode	The score that occurs the most frequently.

representations if used to describe the data. For a distribution that is positively skewed, the mean will represent a value that is higher than most other scores, especially the most common ones. The mean will also be greater than the median. For a distribution that is negatively skewed, the mean will represent a value that is lower than most other scores and will be slightly less than the median. In a bimodal distribution, there are two modes. The mean and median will be near the ebb where the two modes (or humps) come together. The mean, median, and mode are approximately equal in peaked and flat distributions.

Some measures are more or less appropriate for describing different variables given the level of measurement. For instance, it does not make sense to report the mean sex of subjects since sex is a nominal variable. Even though the researcher will probably assign numeric values to the two levels of sex to facilitate analysis, these serve no other function than assisting with computing frequencies. Thus, for nominal variables, the most appropriate measure of central tendency is the mode. One should also exercise caution when reporting the mean of an ordinal level variable when it represents categories.

11.2.1.2 Measures of Dispersion

While measures of central tendency describe where scores cluster in a distribution, measures of dispersion describe how scattered scores are about the center or how scores deviate from the mean. There are three basic measures of dispersion: range, variance, and standard deviation. Definitions of each are provided below (Table 11.2). Variance and the standard deviation are very similar. The major difference is that the standard deviation is smaller (since it is the square root of the variance). The dispersion measures for our sample data are: range = 4, variance = 2.80, and standard deviation = 1.67.

An earlier distinction was made between formulas for samples and for populations. In the formula for computing variance, if a population was studied instead of a sample, the denominator would be N instead of $n - 1$ (note that the capital N is used to indicate the size of the population, while the lower case n is used to indicate the size of the sample). There are different ways to note values that describe

Table 11.2 Measures of dispersion: range, variance, and standard deviation.

Measures of Dispersion		
Range	The difference between the maximum and minimum scores in a distribution.	$\text{Max}_x - \text{Min}_x$
Variance	The mean of the squared deviation scores. To compute the variance, first compute the mean for a set of numbers and then subtract each individual score from this mean. Square these values and then sum. This value is called *sum of squares*. Next, divide the sum of squares by the total number of scores minus 1 $(n - 1)$.	$\frac{\sum (\text{X} - \overline{\text{X}})^2}{n-1}$
Standard Deviation	The square root of the variance.	$\sqrt{\frac{\sum (\text{X} - \overline{\text{X}})^2}{n-1}}$

Table 11.3 Notations for descriptive measures of populations and samples.

	Population	Sample
Mean	μ	$\overline{X}$
Variance	σ^2	s^2
Standard deviation	σ	s
Number of observations	N	n

populations (parameters) and values that describe samples (statistics). This notation is provided in Table 11.3.

Subtracting 1 from the sample size is common in many formulas. It can be thought of as a penalty to ensure that the sample values are unbiased estimators of the population values. This is related to *degrees of freedom. Degrees of freedom* is the extent to which scores in a sample are free to vary. In our example set of query scores, the mean number of queries is 2.4 and the sum of all the values is 12. Degrees of freedom do not change this summative value (whatever it might be), but say something about how much freedom individual scores have to vary and still yield the same sum. That is, in order to arrive at the same sum of scores, each of the individual scores can vary so long as the value of one score is fixed. Any number of scores can be added and yield the value of 12, but one score will have to be reserved to maintain this. For instance, the 1st–4th query scores could be 2, 4, 4, and 1, which would add to 11. This is would mean that in order to maintain a sum of 12, the 5th score would have to be 1. Thus, given a particular sum of values, the values of all of the scores except one can vary. Degrees

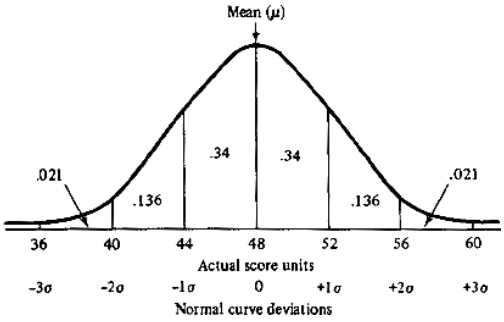


Fig. 11.3 The normal curve and its properties with respect to distribution of data.

of freedom indicate how many scores can vary and how many have to stay fixed.

The curve shown in Figure 11.3 is useful for illustrating dispersion and particularly, the standard deviation. The normal curve has special properties that allow one to make statements about how scores fall around the mean with respect to different units of the standard deviation. These properties are displayed on the normal curve in Figure 11.3. In normal distributions, approximately 68% of all scores fall within $+/-1$ standard deviation of the mean, 95% fall within $+/-2$ standard deviations and 99% fall within $+/-3$ standard deviations of the mean. For any item that is normally distributed in the population, these characteristics will hold. Thus, knowledge that the variable of interest is distributed normally within the population and knowledge of the mean and standard deviation of the sample allows one to understand the distribution of all scores.

Let us consider an example. Imagine if a researcher found that the mean score for the usability measure listed above was 3.5. This value provides an indication of the average value subjects marked for this question. Now imagine that the standard deviation was 0.25. This would indicate that the scores were fairly similar to one another; that is, that about 68% of the scores were between 3.25 and 3.75. Now imagine that the standard deviation was 2.0; this value would indicate that scores were fairly different from one another and that subjects used a range of values. The same number of scores (68%) would fall between the values of 1.5 and 5.5. For variables that are normally distributed

in the population, the mean and standard deviation allows one to form a basic understanding of the data. Thus, when reporting measures of central tendency some measure of dispersion should be reported. Each statistic provides an important and different characterization of the data. Reporting a measure of central tendency such as the mean without reporting a measure of dispersion can give misleading results.

Measures of dispersion describe the distribution of scores about the central tendency. If the standard deviation is large relative to the mean, then scores will be more varied. If the standard deviation is small relative to the mean, then scores will be more alike. This can be easily illustrated with the peaked and flat distribution in Figure 11.3. In a peaked distribution, scores are stacked and tightly clustered around the mean and as a result the standard deviation is small in relation to the mean. In a flat distribution, scores do not cluster and are spread evenly across a range of values; as a result, the standard deviation is large in relation to the mean.

One of the most common concerns that researchers have when preparing to conduct data analysis is that the distributions of *their* data are not normal. This is not unusual, since sample sizes are typically quite small and probability sampling techniques are not used. Since researchers are usually working from small samples, it is unlikely that the variable *in their sample* will be normally distributed. Consider the following illustration of this point. Two-hundred researchers across the world decide to collect data about the same variable in the same way. Each researcher can create an individual distribution curve from the data they have collected (a sample distribution). Each of these distributions is likely to vary, but when they are put together they should form a normal distribution (if the variable is in fact normally distributed in the population). By themselves, each of the 200 sample distributions may look strange, but many will be very similar — these distributions will form the hump of the normal curve. Many will be unique and differ from most others — these distributions will form the tails of the normal curve. Thus, depending on the sample distribution, one's sample data may or may not reflect a normal distribution. However, the requirement is that the data is distributed normally in the population.

11.2.1.3 *z*-scores

A common way to standardize and compare scores across a range of distributions is to create *z-scores*. Raw scores are transformed relative to the mean of the scores, so that the mean is set to the value zero on the normal curve in Figure 11.3. A z-score describes the exact location of a score within a distribution: the sign of the score ($-$ or $+$) indicates if the score is above or below the mean and the magnitude of the score indicates how much the score is above or below the mean. The formula for transforming raw scores into z-scores is

$$z = \frac{X - \mu}{\sigma}$$

where X is the raw score, μ is the mean of the all of the scores, and σ is the standard deviation. It is important to point out that in this formula population parameters are used for the mean and standard deviation (notice the use of μ and σ). The value produced by the formula situates the individual score relative to the mean and states how many standard deviations the raw score is from the mean and in which direction. The numerator in the equation is known as the *deviation score*; these values are used in a number of inferential statistical tests and will be re-visited.

Because z-scores rest on an assumption of normality, information about the likelihood of any particular score occurring in a population is available. Specifically, it is possible to indicate the probability of obtaining a score that is higher or lower than a particular target score. As described earlier, the normal distribution is symmetrical and has certain properties with respect to how scores are distributed about the mean. The most basic statement that can be made is that the chance of observing a score above the mean is 50% and the chance of observing a score below is mean is also 50%. Note that the z-score for a mean is equal to zero. The chance of observing a raw score that is greater than a z-score of $+1$ is about 16/100 since only about 16% of all scores are in this part of the curve. The chance of observing a raw score below this z-score is about 84/100. Thus, z-scores can be used to reason about the likelihood of observing particular scores by turning percentages into proportions. In this example, the z-score is an integer which can be

used easily with the normal distribution to find probability values for observing particular scores. In cases where the z-score is not an integer, the *unit normal table* can be used [27]. Portions of this table are shown below in Table 11.4.

Reasoning about the likelihood or probability of some event occurring is the basis of statistical inference. In the table above, the probability of observing a z-score of 1.65 is less than 0.05. The reader may be familiar with seeing probability values of 0.05 and 0.01 used in conjunction with statistical tests. These values refer to particular areas of the normal curve and the likelihood that some observed sample mean will fall in this area. The z-score formula in the preceding paragraph assumes that the researcher has knowledge of two population parameters — the mean and standard deviation. Of course, this is rarely the case in any study, especially IIR evaluations. In the next section, we will examine several examples that allow us to move from knowing some information about the population to not knowing any. This will allow us to look closely at some important statistical concepts, in particular those that try to account for discrepancies between sample statistics and population parameters.

Sampling error is the amount of error between a sample statistic and its corresponding population parameter. Thus, when computing inferential statistics one must somehow account for the error introduced by studying a sample instead of the entire population. As described previously, it is not always the case that a researcher's sample data will be distributed normally. However, an assumption is made that if many researchers collected many different samples from the same population, then the distribution of means for all the samples together will be normally distributed. That is, most researchers will find similar means; these will pile-up and form the hump on the normal curve. Some researchers will find means that vary different distances from the common mean; these values form the tails of the curve.

These ideas form the basis of the *central limit theorem*, which is at the core of most inferential statistical tests. This theorem states that the distribution of sample means will approach a normal distribution displaying the real mean and standard deviation of the population as

Table 11.4 Unit normal table for interpreting z-scores.

z-score	Proportion in body	Proportion in tail		z-score	Proportion in body	Proportion in tail		z-score	Proportion in body	Proportion in tail
0.00	0.5000	0.5000		1.00	0.8413	0.1587		1.64	0.9495	0.0505
0.01	0.5040	0.4960		1.01	0.8438	0.1562		1.65	0.9505	0.0495
0.02	0.5080	0.4920	...	1.02	0.8461	0.1539	...	1.66	0.9515	0.0485
0.03	0.5120	0.4880		1.03	0.8485	0.1515		1.67	0.9525	0.0475
0.04	0.5260	0.4840		1.04	0.8508	0.1492		1.68	0.9535	0.0465
0.05	0.5199	0.4801		1.05	0.8531	0.1469		1.69	0.9545	0.0455

n (the number of samples) approaches infinity. The basic idea behind this theorem is that as n gets larger, the distribution of sample means more closely approximates the normal curve — as a result the error between the sample and population means decreases. In the section on sampling, the relationship between sample size and power was discussed. This is embodied in the *law of large numbers*: larger samples will more representative of the population from which they are selected than smaller samples.

11.2.2 Inferential Statistics

As mentioned above, inferential statistics allow one to make inferences about population parameters based on sample statistics. Inferential statistics are most often used to test relationships between two or more variables and evaluate hypotheses. While it is possible to test the difference between a known population parameter and a single variable, it is rarely the case that the population parameter is known. Although uncommon, researchers who perform inferential statistics are meant to select appropriate tests during the design phase of a study. Thus, inferential statistics can be viewed as research tools about which researchers make choices, just as other instruments. Furthermore, the choice of which test to use is determined in part by variable data types (levels of measurement), so considering these things simultaneously will likely lead to better design choices.

11.2.2.1 *z*-statistic

Similarly to how one uses z-scores to reason about individual scores, one can also use z-scores to reason about the likelihood of observing particular sample means. The z-score formula changes slightly since we are dealing with sample means instead of individual raw scores. The z-score formula for reasoning about the likelihood of sample means is

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \quad \text{where } \sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}}.$$

The numerator in this formula is virtually the same: it is the difference between the real and observed means. The denominator is a bit different and is known as the *standard error*. Instead of being equal to the standard deviation of the population, it is defined as the square-root of the population standard deviation squared divided by the sample size. Basically, the standard error is a measure of the difference that would be expected to occur by chance between the sample mean and the population mean. This value is clearly a function of the size of sample size: when the sample size increases the standard error gets smaller and the sample mean tends to more closely approximate the population mean.

Although the z-score formula of sample means still assumes that the researcher has knowledge of the population mean and standard deviation, it can be used in special cases for hypotheses testing when this information is available. There are not any apparent situations in IIR where this might be possible, so an example from education will be provided. The point of providing such an example is to segue from z-scores to test statistics. In fact, in this situation it is more appropriate to refer to the z-score as a *z-statistic* (similar to the t- and F-statistics for the t-test and ANOVA, respectively), since it is used for hypotheses testing using sample data.

Suppose a researcher has created a two-week program that is meant to educate a person about information literacy (IL). The researcher hypothesizes that this program will have an effect on IL. The null hypothesis is that the program has no effect, or stated another way that there is no relationship between the program and IL. Information literacy (IL) will be measured using a nationally standardized test. It is known from national results that scores on this test are normally distributed and that the average IL score is 70 ($\mu = 70$) with a standard deviation of nine ($\sigma = 9$). Suppose the researcher enrolls 16 people in the study ($n = 16$). After the program these people complete the IL test and score a mean of 74.

To evaluate whether the difference between the average IL score of study subjects is statistically different from the average IL score of the population, a z-statistic can be computed to determine the probability of the sample IL score occurring in the population. Before this statistic is computed, the researcher must determine the boundary that

separates the high-probability samples from the low-probability samples. This boundary is known as the *alpha level* and by convention is typically set to 0.05, although in medical research this value is often smaller (0.01 or even 0.001). This value should be recognizable to many readers as the probability value (p-value) that is typically reported alongside statistical results. This value will separate the most unlikely of the sample means (from our imaginary distribution of means) from the more common ones (95% of them).

The extremely unlikely values defined by the alpha level make up what is called the *critical region* of the curve — basically this area is at the extreme ends of the normal curve. Because the researcher did not state a directional hypothesis, but just stated that there would be an effect, the 0.05 must be split evenly so that half of the critical region (0.025) is at the low-end of the curve and the other half is at the high-end of the curve. This is known as a *two-tailed test* — the experimental educational program can have either a positive or negative effect on IL scores; the sample mean may be found in the extreme left (low) or right (high) tails of the curve. As stated earlier, the researcher's hypothesis was non-directional — that is, it did not state whether this program would have a positive or negative effect, so a two-tailed test is most appropriate. The alternative is a *one-tailed* or *directional hypothesis test* where the researcher states that the impact of the experimental treatment (i.e., the education program) will have a positive (or negative) effect on IL scores. For one-tailed tests, the critical region is contained within one area of the curve, either the extreme right (for positive effects) or the extreme left (for negative effects), and the 0.05 is concentrated in one end or the other (it is not split) thus making the critical region larger.

Essentially, the value that is being tested in a statistical test is the *difference* between the population mean and the sample mean (in other statistical tests it is the difference between two or more sample means). Assuming a constant alpha level, it is easier to achieve statistical significance with a one-tailed test, since the larger, concentrated critical region accommodates a larger number of values. The necessary minimum difference between means to make it into the critical region will be smaller, and thus, more easily achievable, with a one-tailed test

than with a two-tailed test (although one could adjust the alpha level for one-tailed tests). In many IIR evaluations, researchers are able to make directional hypotheses. However, the standard practice is to use two-tailed tests[2] with an alpha level of 0.05, which is basically equivalent to a one-tailed directional test with an alpha level of 0.025. This is not particular to IIR; in most behavioral sciences two-tailed tests are normally used even when directional hypotheses are stated because the risk of rejecting a null hypothesis that is actually true[3] is too great with one-tailed tests. Since a two-tailed test requires stronger evidence to reject a null hypothesis it provides more convincing results that the null should be rejected. All tests discussed in this paper will assume two-tailed tests.

Let us get back to the example of the educational researcher. First, we would consult the unit normal table to determine the critical region of the curve for $p = 0.05$. The critical value separating the uncritical region from the critical region of the curve is displayed in the portion of the unit normal table in Table 11.4. We see that z-scores that are greater than or equal to 1.65 are in the portion of the tail that corresponds to an alpha value of 0.05. The z-statistic for our example data is 1.77, so we can reject the null hypothesis since the probability that a z-score of 1.77 came from the population distribution is less than 5/100.

When conducting hypotheses tests, there are two important types of errors that can occur, *Type I* and *Type II*. A Type I error occurs when a researcher erroneously rejects the null hypothesis. Type I errors can occur because of the researcher's actions — e.g., blatantly ignoring test results or setting the alpha value too low — but the more common source of Type I errors are anomalous results. In these cases, results are found to be statistically significant with a standard alpha value of 0.05, but there is something peculiar about the sample or testing method that caused the results. In the IL example above, it may have been that all of the members of the sample were just smarter than average, which would explain why they scored higher. The alpha level

[2] This is the default in most statistics packages.
[3] This is known as a Type I error.

actually determines the probability that a statistical test will lead to a Type I error. For instance, an alpha level of 0.05 means that there is a 5% chance that the data obtained in the study was a result of some anomalous condition. In other words, if the study were done 100 times, the same results are expected 95 of those times.

While the risk of a Type I error is actually quite small for single hypothesis tests, when researchers conduct multiple independent hypotheses tests on the same data set, the critical alpha value is often adjusted to further safeguard against obtaining statistical significance by chance. One common type of alpha adjustment is the Bonferroni correction, which reduces the critical alpha value by dividing some standard, such as 0.05, by the number of independent hypotheses that will be evaluated. For example, if a researcher examines five independent hypotheses, then a Bonferroni correction would change the critical alpha value to 0.01 (0.05/5). A less restrictive correction is the Holm–Bonferroni method.

A Type II error is failing to reject the null hypothesis when it should be rejected. In these situations, the test statistic does not fall into the critical region of the curve, even though the treatment may have a small effect. Strictly speaking, there is no way to determine the probability of making such an error. Probability values can provide a hint that a Type II error has occurred (for instance, if the test statistic falls into a region defined by 0.07 instead of 0.05), but this is not a universal explanation for why significant results were not found. In some cases, researchers often claim that results are *almost* statistically significant, but this inappropriate. If, for instance, two more subjects were included in the sample, the test statistic might move *further away* from the critical region instead of closer to it. Repetition of an experiment will allow one to *explore* the possibility that a Type II error occurred, but it is not a guarantee that results will change.

11.2.2.2 *t*-statistic and t-tests

The z-statistic in the preceding example is useful in cases where the population mean and standard deviation are known *a priori*, but this

is rarely the case in IIR evaluations. It is usually the case in most IIR evaluations, that two or more sample means are being compared with no knowledge of the population parameters. For instance, in the IIR research scenario, a researcher might be interested in comparing the differences in performance according to sex — is there a difference between the performance of males and females? The formula used to compute the z-statistic can be modified to account for the differences in what is known (or unknown) about the population parameters. First, the denominator in the formula for the z-statistic changes to one based on the variance of the sample, instead of one based on the standard deviation of the sample. The new denominator is called the *estimated standard error*:

$$z = \frac{X - \mu}{s_{\overline{X}}} \quad \text{where } s_{\overline{X}} = \sqrt{\frac{s^2}{n}}.$$

The second change accommodates the comparison of two sample means, which basically doubles all of the elements of the formula above, so that the above formula becomes:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}}.$$

In this formula, the two sample means are represented by $\left(\overline{X}_1\right)$ and $\left(\overline{X}_2\right)$. The null hypotheses, which assumes that these two samples were drawn from the same population and that there would be no difference between the means, is represented by the difference between μ_1 and μ_2 (which is eventually replaced by zero and excluded from the formula).

The magnitude of the standard error (the denominator in the formula) is determined by the variance of the observed scores and the sample size. With two sample means, there are two measures of variance (one for each sample) and therefore, two standard errors. In the formula above, the denominator is defined as:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}.$$

Research Question: What is the relationship between sex and perceptions of system usability?
Hypothesis: Males will rate the system as more usable than females.

	Sex		Deviations from samples means		Deviations squared	
	M	F	d_M	d_F	$d_M{}^2$	$d_F{}^2$
Usability scores	4	3	–0.2	1	0.04	1
	5	3	0.8	1	0.64	1
	4	2	–0.2	0	0.04	0
	5	1	0.8	–1	0.64	1
	3	1	–1.2	–1	1.44	1
Mean (M)	4.2	2				
n	5	5		Σd^2	2.8	4

$$t = \frac{\overline{X}_M - \overline{X}_F}{\sqrt{\left(\frac{\Sigma d_M{}^2 + \Sigma d_F{}^2}{n_M - n_F - 2}\right)\left(\frac{n_M + n_F}{n_M \cdot n_F}\right)}}$$

$$= \frac{2.2}{\sqrt{\left(\frac{2.8+4}{5+5-2}\right)\left(\frac{5+5}{5\cdot 5}\right)}}$$

$$t = \frac{2.2}{.58309} = 3.77$$

Fig. 11.4 Calculation of *t*-statistic using sample sex and usability data.

This formula is based on the pooled variance of the two samples,[4] where

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}.$$

SS represents the *sum of squared deviations*, which is best exemplified in Figure 11.4. Recall that the sum of squares is used in the computation of the variance and standard deviation. Within each sample group, each observed score is subtracted from the mean and squared. These values are then summed. The other figures in the formula for pooled variance as well as the parent formula for the standard error are based on the sample sizes of the two groups being compared. Sample size figures into the calculation directly (i.e., using the number itself) and indirectly through degrees of freedom (*df*). As a reminder, degrees of freedom is the extent to which scores in a sample are free to vary. For the *t*-statistic, this is equal to, $n_1 - 1 + n_2 - 1$, where n_1 is equal to the number of observations in the first group and n_2 is equal to the number of observations in the second group. This is often abbreviated as $n - 2$, where n is equal to the total number of observations in the study (i.e, total sample size).

[4] This is basically an average of the variances for the two samples.

To consolidate the formulas above, we can replace the $s_{\overline{X}_1-\overline{X}_2}$ in the t-statistic formula with the actual formula:

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}.$$

Figure 11.4 walks the reader through a complete example based on the IIR Research Scenario. In this example, the researcher is interested in investigating the effect of sex on usability. The t-statistic formula given in the example is a variation of the one above. Note that it does not include $(\mu_1 - \mu_2)$ since this value is assumed to be zero.

To determine whether a t-statistic is significant, the t-distribution table has to be consulted. This is similar to consulting the unit normal table that was consulted to determine if a particular z-statistic was significant. The objective is the same: to determine what t-value needs to be observed in order for the statistic to fall in the critical region of the curve.[5] To determine this value, we need to use the total degrees of freedom $(n - 2)$ and the alpha level. The example in Figure 11.4, $df = 8$ and the alpha level is 0.05. A portion of the t-distribution table [27] is presented in Table 11.5. To use this table, we first find the column that corresponds to an alpha level of 0.05 for a two-tailed test and then the row that corresponds to our degrees of freedom. The figure at this intersection tells us that we need a value that is greater than or equal to 2.306. In Figure 11.4, we see that the t-statistic is 3.77, which is greater than the critical t-value, so we can reject the null hypothesis.

Table 11.5 illustrates several important things about the t-statistic. First, note that all of the values in the table are positive, even though t-statistics can be negative. To use the table, one has to take the absolute value of the t-statistic. The only information that the sign adds is that it tells us which mean is greater. If the mean for Group 1 (male) is greater than the mean for Group 2 (female) the t-statistic will be positive. If the mean for Group 1 is less than the mean for Group 2 the t-statistic will be negative. Another thing to note from the table is

[5] Strictly speaking, target t-values associated with the critical region should be determined before the test statistic is calculated. Since computers do most this work for us all in one step, the order is not as important practically. However, the researcher should declare the acceptable alpha levels before conducting statistical analysis.

Table 11.5 Portion of the *t*-distribution. Degrees of freedom are associated with each row, while alpha levels (probability values) are associated with columns.

	Alpha Levels							
	One-tailed:	0.4	0.25	0.1	0.05	0.025	0.01	0.005
	Two-tailed:	0.8	0.5	0.2	0.1	0.05	0.02	0.01
Degrees of Freedom	1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
	2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
	3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
	4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
	5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
	6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
	7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
	8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
	9	0.261	0.703	1.383	1.833	2.262	2.821	3.250

that as sample size increases (as evident from the *df*), the critical *t*-value decreases, and so bigger sample sizes require smaller differences (also note the rate at which the critical *t*-values shrink as a result of increases in the *df*). Finally, keeping the *df* constant, notice how the critical *t*-value changes as a function of the alpha levels. To really understand how much easier it is to achieve significance with a one-tailed test instead of a two-tailed test (and why they carry a great risk of a Type I error), note that if we conducted a one-tailed test our critical *t*-value would have been 1.860 instead of 2.306.

The description above is for an independent samples *t*-test which arguably is the most common type of *t*-test conducted in IIR research. Another type of *t*-test which is also used in IIR research is the paired-samples *t*-test. To illustrate the difference between these two tests, we will re-visit the example study describing IL. In the original design of this study, the treatment group was compared to the population, but this study could have been designed in at least two other ways. In the first alternative design, the researcher could have two sample groups of subjects,[6] with one group receiving the education program and the other group not receiving the program. In the second alternative design, the researcher could have used a single group of subjects and given them a pre-test to elicit a baseline measure of IL, administered the program, and then given a post-test to measure IL. Neither of

[6] This design assumes that subjects are randomly assigned to conditions.

these designs requires knowledge of the population mean or standard deviation. Instead, to examine the null hypothesis in the first design alternative we would compare the mean IL scores of the two groups using the standard, independent samples t-test, while in the second design alternative we would compare the pre- and post-test IL scores of individual subjects. These two design alternatives illustrate the difference between when would use an *independent samples t-test* and when one would use a *paired-samples t-test*. The independent samples t-test examines differences in the means of two separate groups of subjects, while the paired-samples t-test examines differences within-subjects — subjects' pre- and post-test scores are compared with one another.

The formula for the paired-samples t-test is nearly identical to that used for the independent samples t-test except that the sample data are difference scores and are represented by D instead of X. This formula is

$$t = \frac{\overline{D} - \mu_D}{s_{\overline{D}}} \quad \text{where } s_{\overline{D}} = \sqrt{\frac{s^2}{n}}.$$

Notice that the *population mean difference* (instead of the population mean) is subtracted from *sample mean difference*. Also notice that the estimated standard error is given for the mean difference instead of the mean score. As was the case with the independent samples t-test, the *population mean difference* represents the null hypothesis and is set to 0 ($\mu_D = 0$).

11.2.2.3 *F*-statistic and Analysis of Variance (ANOVA)

Very often in IIR evaluations, researchers study a single independent variable with more than two groups or levels. Using the IIR research scenario, an example variable with three levels is system type. When one wants to compare the differences between three or more means, then *analysis of variance* (ANOVA) is used. There are several types of ANOVAs. In this paper, the single and multi-factor ANOVAs are discussed. The multivariate ANOVA (MANOVA) is introduced, but not discussed in detail.

The t-statistic measured the difference between sample means divided by the difference expected by chance (estimated standard error). The statistic produced by an ANOVA is called the F-ratio or F-statistic. It is similar to the t-statistic, expect it uses variance among groups, rather than means in its computation. Although variance is compared directly instead of means, the purpose of the test is to evaluate differences in means between conditions. Because there are more than two means, it is easier to compare variances.

The basic ANOVA formula is the ratio of differences in variances between the sample means to differences in variances expected by chance (called the *error variance*). For an ANOVA to be significant two major things need to happen: there needs to be a difference between at least one pair of means (for instance, between the mean performance for System A and System B, or System A and System C or System B and System C)[7] and the variance within each group must be small. When the within group variance increases, the error variance (the denominator in the F-ratio) also increases. The basic formula is given below.

$$F = \frac{variance\_between\_treatments}{variance\_within\_treatments}.$$

Although it is seemingly simple, it involves a number of calculations and uses the means and variances of the each group, as well as the total mean and variance.[8] When presenting the details of this formula, the example from the IIR research scenario stated above will be used, where system type is the independent variable and performance is the dependent variable. Thus, the researcher is interested in determining whether there is a statistically significant difference in performance according to system type.

To compute the ANOVA, the first step is to compute the overall variability using all of the data. After we have done this, we need to partition this variability into two components: variability between each condition and variability within each condition. This is known as *between-treatments variance* and *within-treatments variance*. The

[7] This is called *pair-wise comparisons*, which will be discussed in more detail later.

[8] In this discussion, we will use the term *grand* to refer to the total mean and variance of all of the subjects, and *sample* to refer to the means and variances for subjects in each condition (also called group or level).

between-treatments variance helps us understand how much variance occurs as a result of the different treatments, while the within-treatments variance helps us understand how much variance occurs by chance. Recall that the logic of hypothesis testing is to determine whether the probability of observing our results by chance is less than 5% (or 1%, depending on the alpha level). ANOVA attempts to measure this chance, or error variance.

Differences due to chance are typically attributed to two sources: individual differences and experimental error. These differences exist, but are independent of the treatment and should not be attributed it — the ANOVA attempts to partition these differences from differences caused by the treatment. Although steps can be taken to minimize the error introduced by these sources, such steps will never eliminate this error. For instance, random assignment to condition should distribute subjects with varying individual differences equally across groups, but there are lots of individual differences, so some error will still likely exist. Experimental error is related to how the experiment was conducted; a particular concern is the error introduced by poorly designed instruments, measures and protocols. A researcher can work to minimize these differences as well, but it is unlikely that they will be completely eradicated. To account for such chance differences, the ANOVA formula can be recast as

$$F = \frac{treatment\_effect + difference\_due\_to\_chance}{differences\_due\_to\_chance}.$$

The formula above helps us to not only understand the logic of the ANOVA, but also interpret the quotient it produces. In the above formula, if there was no treatment effect (treatment effect = 0) then the numerator and denominator would be equal since they measure the same thing, chance difference. This will result in an F-ratio of 1.00. When this happens, we fail to reject the null hypothesis since the only differences that were observed were due to chance. An F-ratio of 1 will never be statistically significant. It is also the case that F-ratios will always be positive numbers (unlike t-statistics). Since a value of 1 indicates no statistical significance, the values of the F-distribution will accumulate around this point on the distribution graph, similar to

how they accumulate around the mean in a normal distribution. Recall that to determine the significance of a z- or t-statistic we examine locations on a normal distribution. To determine whether an F-statistic is significant, the F-distribution is consulted. Instead of being normally distributed, the F-distribution is positively skewed. This is because F-values are always positive (essentially we are looking at the positive half of the normal distribution).

The exact shape of the F-distribution and the critical values needed to obtain statistical significance are determined by the alpha level and the degrees of freedom (*df*). There are three different *df* associated with an ANOVA: within-treatments (df_{within}), between-treatments (df_{between}) and total (df_{total}). The df_{total} is the sum of df_{within} and df_{between}, and is also equal to $n - 1$, where n is equal to the total sample size. The df_{within} is the difference between the number of levels of the independent variable and the total sample size. The df_{between} is equal to the number of levels of the independent variable (k) minus 1 ($k - 1$).

To demonstrate the relationship among sample size, levels of a variable, alpha levels, and critical F-values, a portion of the F-distribution is shown in Table 11.6 [27]. The df_{between} and df_{within} are used to locate F-values in this table. For instance, if we were examining the relationship between our three experimental IIR systems and performance, and we recruited nine subjects and randomly assigned three of these subjects to each system, then $df(\text{between}) = 2$, or $3 - 1$ and $df(\text{within}) = 6$, or $9 - 3$. Our critical F-values are equal to 5.14 for significance at the 0.05 alpha level or 10.92 for significance at the 0.01 alpha level.

In Table 11.6, notice the relationship between sample size (as evidenced by df_{within}) and critical F-values: as one increase the other decreases, but the rate at which this happens diminishes at some point. Also notice the extremely large F-values in the first row of the table. These represent cases where there is only one more subject than there are levels of the independent variable. In fact, to use this table, n must always be at least one point greater than the number of levels (k). For example, if a variable with eight levels was being tested on eight subjects (one subject per level) ($df_{\text{between}} = 7$ and $df_{\text{within}} = 0$), the critical F-value would be indeterminable. However, if nine subjects were used

Table 11.6 Portion of the F-distribution for $p < 0.05$ (top number in cell) and $p < 0.01$ (bottom number in cell). Degrees of freedom within groups is associated with rows, while degrees of freedom between groups is associated with columns.

	Degrees of Freedom$_{\text{between}}$						
Degrees of Freedom$_{\text{within}}$	1 ($p<0.05$) ($p<0.01$)	2 ($p<0.05$) ($p<0.01$)	3 ($p<0.05$) ($p<0.01$)	4 ($p<0.05$) ($p<0.01$)	5 ($p<0.05$) ($p<0.01$)	6 ($p<0.05$) ($p<0.01$)	7 ($p<0.05$) ($p<0.01$)
1	0161 4052	0200 4999	0216 5403	0225 5625	0230 5764	0234 5859	0237 5928
2	18.51 98.49	19.00 99.00	19.16 99.17	19.25 99.25	19.30 99.30	19.33 99.33	19.36 99.34
3	10.13 34.12	09.55 30.92	09.28 29.46	09.12 28.71	09.01 28.24	08.94 27.91	08.88 27.67
4	07.71 21.20	06.94 18.00	06.59 16.69	06.39 15.98	06.26 15.52	06.16 15.21	06.09 14.98
5	06.61 16.26	05.79 13.27	05.41 12.06	05.19 11.39	05.05 10.97	04.95 10.67	04.88 10.45
6	05.99 13.74	05.14 10.92	04.76 09.78	04.53 09.15	04.39 08.75	04.28 08.47	04.21 08.26

($df_{\text{between}} = 7$ and $df_{\text{within}} = 1$), the critical F-value is determinable, although impossibly large ($F = 237$). Thus, ANOVA is more accurate when there is a reasonable relationship between k and n. ANOVA is most accurate when there are equal sample sizes across condition. ANOVA is robust enough to handle unequal sample sizes, but the overall samples size should be relatively large and there should not be a huge discrepancy between the sample sizes for each condition.

ANOVA was originally developed for experimental situations where researchers have control over the assignment of subjects to conditions. However, in some cases ANOVA may be used to examine the effects of a variable that was not originally controlled in a study. For instance, in the IIR research scenario above, a researcher might be interested in examining the relationship between familiarity and performance. In this example, familiarity would be a *quasi-independent* variable since it was not manipulated by the researcher. It is unlikely that the distribution of familiarity scores will be equal across the five levels of the scale. If the distribution is too skewed, then the researcher might, for instance, consider reducing the five levels into three.

Conducting an ANOVA requires a number of calculations. First, recall that sample variance (which makes up both the numerator and denominator of the F-ratio), is equal to the sums of squared deviations (SS) divided by the *df*. We need three types of variances to compute the F-statistic: between- and within-treatment variance, as well as total variance. Thus, three different sums of squared deviations (SS) values must be computed, along with three *df* values. Once we have these values we can compute the F-statistic; the following formula which consolidates everything can be used

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}.$$

To demonstrate all of the computations, let us revisit our original example where a researcher is interested in examining the differences among three systems with respect to performance (Figure 11.1). The original description had each subject use each of the three systems (within-subjects design). However, the basic ANOVA assumes that observations are independent and in order to use the ANOVA formula used in this section, we would need to have an experimental design where each subject only used one of the systems (between-subjects design). Otherwise, we would compute a *repeated-measures ANOVA*. Figure 11.5 illustrates the computation of the F-statistic with sample data assuming that each subject only uses one system. This example is divided into four parts:

- Part (I) shows the computation of the *within groups SS*. Deviations are taken using the individual group means. The sums of these squared deviations are then summed to form the *within groups SS*.
- Part (II) demonstrates the computation of the *total SS*. Deviations are taken using the grand mean instead of the individual group means.
- Part (III) demonstrates the computation of the *between groups SS*. This is computed by taking the deviations between the grand mean and each individual group mean. These values are then squared and multiplied by the n of each group. These values are summed to form the *between groups SS*.

Research Question: What is the relationship between system type and performance?
Sample Hypothesis: Subjects will perform better with System C than with Systems A or B.

System and Mean Performance		Part I: Score Deviations from Group Means d_w	Part I: d_w^2	Part II: Score Deviations from Grand Mean d_g	Part II: d_g^2	Part III: Group mean deviations from the Grand Mean d_b	Part III: d_b^2	Part III: Group n times d_b^2
System A	0.2347	0.0525	0.0028	0.1121	0.0126			
	0.3426	-0.0554	0.0031	0.0042	0.0001			
	0.2788	0.0084	0.0001	0.0680	0.0046			
	0.3046	-0.0174	0.0003	0.0422	0.0018			
	0.2753	0.0119	0.0001	0.0715	0.0051			
Mean	0.2872					0.0596	0.0036	0.0178
n	5							
System B	0.3123	-0.0685	0.0047	0.0345	0.0012			
	0.1258	0.1180	0.0139	0.2210	0.0488			
	0.2338	0.0100	0.0001	0.1130	0.0128			
	0.3104	-0.0666	0.0044	0.0364	0.0013			
	0.2368	0.0070	0.0001	0.1100	0.0121			
Mean	0.24382					0.1030	0.0106	0.0530
n	5							
System C	0.5477	-0.0383	0.0015	-0.2009	0.0404			
	0.4878	0.0216	0.0005	-0.1410	0.0199			
	0.5238	-0.0144	0.0002	-0.1710	0.0313			
	0.5011	0.0083	0.0001	-0.1543	0.0238			
	0.4866	0.0228	0.0005	-0.1398	0.0195			
Mean	0.5094					-0.1626	0.0264	0.1232
n	5							
Grand Mean	0.3468	**Within *SS*** (Σd_w^2)	0.0324	**Total *SS*** (Σd_g^2)	0.2354		**Between *SS*** ($\Sigma n d_b^2$)	0.1940

Part IV

Summary table

Source	*SS*	*d.f.*	*MS*	*F*
Between	0.1940	2	0.0970	35.9259
Within	0.0324	12	0.0027	
Total	0.2264	14		

$$F = \frac{0.0970}{0.0027}$$

= 35.9259

Fig. 11.5 Computation of the F-statistic using sample performance data for three systems (A, B and C).

- In Part (IV), the *MS* values are computed as the *SS* divided by the *df*, so that $MS_{\text{between}} = SS_{\text{between}}/df_{\text{between}}$. The quotients are then used to compute the F-statistic.

Recall that for our example, the critical F-values are equal to 5.14 for significance at the 0.05 alpha level or 10.92 for significance at the 0.01 alpha level. Thus, our F-statistic of 35.9259 is significant at the $p < 0.01$ level. Although our ANOVA is statistically significant, we do not know between which pairs of systems there were significant differences.

For instance, mean performance with System A might be significantly different from mean performance with System B and System C, but there might not be any significant difference between System B and System C. There are actually three pair-wise comparisons that we need to make — System A, System B; System A, System C; System B, and System C. To evaluate these pair-wise differences, *post-hoc* tests are conducted. A number of *post-hoc* tests can be used, including *Scheffé*, *Tukey HSD*, and *Bonferroni*. The difference among these tests is beyond the scope of this paper, but the *Scheffé* test is one of the safest and most conservative tests. Using a safe test (and by safe test it is meant a test that reduces the risk of a Type I error) is particularly important with *post-hoc* analysis because of the number of pair-wise comparisons being made. Essentially, one is conducting hypothesis testing for each pair. As one does more tests, the risk of a Type I error accumulates. This is called *experiment-wise alpha level*. The basic notion is that conducting more tests increases the risk that a statistically significant result will happen just by chance. In fact, this is one reason why it is better to conduct an ANOVA instead of multiple *t*-tests: the more tests you conduct, the greater the chance of a Type I error. What makes the *Scheffé* test conservative is that during the pair-wise comparisons, the between-treatments *df* from the ANOVA is used, even though only two groups are being compared. Thus, there are fewer *df*, which makes the test harder. The implication of this is that it is possible to have a statistically significant ANOVA, but no statistically significant *post-hoc* tests.

11.2.2.4 More ANOVAs

In the example above, we explored the effect of one independent variable (system type) on one dependent variable (performance). The basic one-way ANOVA described above also accommodates situations where a researcher examines a single independent variable in relation to multiple dependent variables. Sometimes researchers are interested in looking at more than one independent variable in relation to a single dependent variable. In this case, a *multi-factor ANOVA*[9] is conducted. When the researcher is interested in examining more than

[9] This is also called a *uni-variate ANOVA*, where *uni* refers to the dependent variable.

		System			
		A	B	C	Sex $\overline{X}$
Sex	Males	$\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$
	Females	$\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$
	System $\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$	Grand $\overline{X}$

Fig. 11.6 Example of a 3×2 factorial design using system type and sex. Numbers in cells represent means (of, for example, performance).

one independent variable in relation to more than one dependent variable, a *MANOVA*[10] is used.

Earlier in this paper, the factorial design was introduced. Studies that are designed in this way are typically appropriate for multi-factor ANOVAs. In fact, such representations are useful for understanding what is being compared in the multi-factor ANOVA and what types of computations are required. Figure 11.6 presents a 3×2 factorial representation of a relationship that could be studied from the IIR research scenario. In this example, the impact of two variables, *system type* and *sex*, on *performance* is investigated. Note that system type is an independent variable, since it was manipulated by the researcher, sex is a quasi-independent variable and performance is a dependent variable. (From this point forward, these independent variables will be called *factors* to coincide with the language of factorials.)

The impact of each one of these factors on the dependent variable is called a *main effect*. For instance, system type may have an impact on performance (e.g., all subjects perform better with System A, than Systems B or C, regardless of sex), and sex may have an impact on performance (e.g., female subjects perform better than male subjects regardless of system). In this example, there are two possible main effects. The number of possible main effects is equal to the number of factors.

It may also be the case that system type and sex interact. For instance, females may perform better with System A than System B or C, but males may perform better with System B than System A or C. This is called an *interaction effect*. The purpose of a multi-factor

[10] This is also called a *multi-variate ANOVA* where *multi* refers to the dependent variable.

ANOVA is to test the main effects and the interaction effects.[11] The computation behind a multi-factor ANOVA is nearly identical to that of a single factor ANOVA, expect that F-statistics are computed for each possible relationship (one for each factor and one for all possible interactions). Most multi-factor ANOVAs involve two or three factors. Anything beyond that generates a large number of possible main effects and interaction effects and such results can be extremely difficult to understand and interpret. Moreover, designs with larger numbers of factors require larger numbers of subjects.

Figure 11.6 is useful for conceptualizing the comparisons made during a multi-factor ANOVA. First, recall that ANOVA concentrates on variances as a way to determine whether differences between means are significant. The means are not compared directly, although they are used to compute *SS*. In Figure 11.6, the ANOVA testing for a main effect for sex would compare the row totals, the ANOVA testing for a main effect for system would compare the column totals and the ANOVA testing for the interaction would compare the values in the cells of the table. Thus, we can talk about row means, columns means, cell means, and the overall (grand) mean. Each of these values also has a variance associated with it, which is what is compared with the ANOVA. To calculate the F-ratio for the interaction, the ANOVA first identifies differences that cannot be explained by the main effects. These extra differences are then evaluated to determine whether there is a significant interaction effect. The entire computation of the multi-factor ANOVA will not presented here, because it involves a large number of steps. Instead, means and standard deviations for a sample of 30 subjects are added to the above example and presented in Figure 11.7.

In this example, there are significant main effects for both system and sex: females performed significantly better than males [$F(1,30) = 12.46,\ p < 0.01$)] and subjects performed significantly better with System C than with System A or B [$F(2,30) = 5.55, p < 0.01$)].[12] There

[11] A researcher is not required to have hypotheses about all possible effects. One may only be interested in the interaction effect, but not the main effects.

[12] We would technically need a *post-hoc* test to determine that the difference was C > A, B, but a visual inspection of the means suggests that this would be the only significant relationship.

		System			
		A	B	C	Sex $\overline{X}$
Sex	Males	0.336 (0.023)	0.132 (0.123)	0.214 (0.214)	0.227 (0.088)
	Females	0.130 (0.012)	0.338 (0.024)	0.334 (0.334)	0.267 (0.107)
	System $\overline{X}$	0.233 (0.110)	0.235 (0.110)	0.274 (0.077)	0.247 (0.099)

Fig. 11.7 3×2 factorial with one independent variable, system type, and one quasi-independent variable, sex. The values in the cells represent sample means (standard deviations) for performance ($n = 30$).

is also a significant interaction effect $[F(1,30) = 122.59,\ p < 0.01)]$: males performed best with System A, second best with System C and worst with System B (A > C > B), while females performed best (and about the same) with System B and C and worst with System A (B, C > A). We would need to conduct *post-hoc* tests to pinpoint between which pairs the significant differences occurred.

There are several other types of ANOVAs that will not be discussed in detail. The underlying formulas for computing these ANOVAs are similar to those presented above, but require more computations because there are more comparisons. *MANOVA* (Multiple Analysis of Variance) is basically a combination of the single factor and multi-factor ANOVAs discussed above in that it is used when the researcher is examining the effects of multiple factors on multiple dependent measures. There is also a special version of ANOVA that deals with between-subject independent variables. This is called repeated-measures*ANOVA*. Finally, g*eneralized linear modeling* (GLM) allows one to develop a function describing the relationship among the independent and dependent variables based on significant ANOVA results. This is similar in nature to linear regression, which is discussed briefly below.

11.2.2.5 Measures of Association

Another set of statistical techniques that are often used in IIR evaluations is correlation. There are many kinds of correlation coefficients including Pearson's r, Spearman's *rho*, and Kendall's *tau*. Correlation coefficients are measures of association; basically these coefficients

describe how scores on two variables co-vary.[13] For example, if a subject has a high performance score is the subject likely to give the system a positive usability rating? One important distinction to keep in mind is that correlation does not show causality. Correlation only shows that two things co-vary. Performance and age might be correlated, but this does not mean that performance causes a person's age or that age causes a person's performance. It only means that these things are systematically related. In this paper, we will look at Pearson's r and Spearman's *rho*. Kendall's *tau* is also used a lot in IIR and IR more broadly. Spearman's *rho* and Kendall's *tau* are used to test similar kinds of relationships. In the interest of space, only Spearman's *rho* is presented. Spearman's *rho* is technically a non-parametric statistic, so it will be presented in another section.

Correlation coefficients vary between -1 and $+1$. The magnitude of the coefficient indicates its strength, while the sign indicates if the relationship is positive or negative. A positive relationship indicates that increases in one variable are associated with increases in the other variable (or, conversely, that decreases in one variable are associated with decreases in the other variable). A negative relationship indicates that increases in one variable are associated with decreases in the other variable. This is also known as an inverse correlation. A value of zero indicates that there is no relationship between the two variables, while $+1$ and -1 indicate functional relationships.

Magnitude is very important for interpreting the meaningfulness of the correlation coefficient. It is possible (and common) to find statistically significant correlation coefficients that are actually quite weak, so one should always pay attention to the value of the coefficient. The real problem is that the correlation coefficient does not actually represent the accuracy with which predictions can be made. For instance, a correlation coefficient of 0.30 does not mean that given one variable the other variable can be predicted 30% of the time or with 30% accuracy. The strength of the relationship lies in the squared correlation coefficient (r^2), so that a correlation of 0.30 means that given one variable

[13] Correlation coefficients can be computed for more than two variables, but in this paper we will just consider the relationship between two variables.

Table 11.7 Cohen's and Guilford's guidelines for interpreting correlations.

Cohen (1988)	Guilford (1956)
0.10–0.29 small	< 0.20 slight, almost negligible
0.30–0.49 medium	0.21–0.40 low
0.50 + large	0.41–0.70 moderate
	0.71–0.90 high
	> 0.91 very high

the other variable can be predicted with 9% (0.30^2) accuracy. Clearly, this gives a very different view of the strength of coefficient. For correlation coefficients below 0.50, differences between the actual coefficient and the r^2 values are quite pronounced. Thus, one should be extremely cautious interpreting any statistically significant correlation coefficients, especially those whose values are small.

There are several guidelines for interpreting the magnitude of a correlation. Two such interpretations are given in Table 11.7. Both authors stress that these are guidelines, rather than absolutes. Guildford [113] offers more distinctions than Cohen [56]. Both use the absolute values of the coefficient.

Pearson's r

The Pearson's correlation (r) is one of the most common correlation coefficients. Traditionally, it is used with continuous data types (interval or ratio level data) and measures linear relationships. The calculation for Pearson's r examines the degree to which two variables vary together in relation to the degree to which they vary separately. The formula for Pearson's r is given in Figure 11.8, along with an example from the IIR Research Scenario which looks at the relationship between query length and performance. To calculate Pearson's r, we use the *sum of products of deviations*, which is similar in nature to the sum of squared deviations calculation that was used in the t-test and ANOVA.

The sum of products of deviations is illustrated in Figure 11.8. For any given subject with scores on variables X and Y, the deviations of each of these scores from their respective sample variable means are multiplied. After this is done for each subject, these values are summed to form the sum of products of deviations. This value represents the

	Raw Data		Deviations from Means				
Subject	X (Query Length)	Y (Performance)	x	y	xy	x^2	y^2
1	2	0.2445	−0.5000	−0.0219	0.0110	0.2500	0.0005
2	3	0.3022	0.5000	0.0458	0.0229	0.2500	0.0021
3	4	0.3387	1.5000	0.0723	0.1085	2.2500	0.0052
4	1	0.1804	−1.5000	−0.0860	0.1290	2.2500	0.0074
5	2	0.2556	−0.5000	−0.0108	0.0054	0.2500	0.0001
6	3	0.3433	0.5000	0.0769	0.0385	0.2500	0.0059
7	3	0.2990	0.5000	0.0326	0.0163	0.2500	0.0011
8	4	0.3711	1.5000	0.1047	0.1571	2.2500	0.0110
9	2	0.1915	−0.5000	−0.0749	0.0375	0.2500	0.0056
10	1	0.1277	−1.5000	−0.1387	0.2081	2.2500	0.0192
Σ	25	2.664			$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$
Mean	2.50	0.2664			0.7343	10.5000	0.0581

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2.}} = \frac{0.7343}{\sqrt{10.50 \cdot 0.0581}} = \frac{0.7343}{\sqrt{0.6101}} = \frac{0.7343}{0.7811} = +0.9401$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9401\sqrt{8}}{\sqrt{1-0.8838}} = \frac{2.6590}{\sqrt{0.1162}} = 7.8000$$

Fig. 11.8 Computation of Pearson's r with sample query length and performance data.

extent to which the two variables co-vary. To calculate the extent to which the two variables vary separately, each individual X and Y deviation score is squared and then summed within each variable. These values are then multiplied and square-rooted. This calculation is similar to the variance measures used in the previous statistics.

The Pearson's r for our example data yields a value of +0.9401, which we can see is quite large in magnitude. This suggests that there is a strong positive correlation between query length and performance and the r^2 is 0.8838 which demonstrates that query length explains quite a bit of the variability in performance.

Although our coefficient looks strong, we still need to evaluate it with respect to probability. The null hypothesis states that there is no correlation between the two variables, in which case we would expect $r = 0$. (Even when no relationship exists, the correlation coefficient is usually not zero. In most cases, it would be some non-zero value.) To evaluate the statistic, we need to compute a corresponding t-statistic and use the t-distribution to evaluate the likelihood of observing the coefficient. This formula is given in the bottom of Figure 11.8. This

requires knowledge of the Pearson's r, r^2 and the sample size. In the numerator, the r is multiplied by the square-root of the *df* $(n-2)$ for the sample. The *df* is also used to enter the t-distribution (along with the alpha level). Using Table 11.5, we can see that our critical value is 2.306. Our t-statistic is 7.80, which is significant at the 0.01 level, so we can reject the null hypothesis. Although the t-statistic is necessary in determining whether the null hypothesis can be rejected, many people do not realize that this is computed as part of correlation testing. It is not necessary to report this statistic; instead, we would report $r = 0.9401$, $p < 0.01$.

11.2.2.6 Regression

Correlation coefficients measure the extent to which two variables covary, but they do provide information about how to predict one value from another. Regression can be used to discover the function describing the relationship among two or more variables. Regression is a sophisticated set of statistical procedures and a discussion of these procedures is beyond the scope of this paper. There are several forms of regression including techniques for both linear and non-linear relationships, and for different data types. Regression is also useful for evaluating the importance of a set of predictor variables and determining which are the most useful for predicting a particular output variable.

11.2.2.7 Effect Size

Effect size measures the strength of a test statistic. Very often a researcher might want to go beyond just saying that statistically significant relationships were found, to a discussion and comparison of the strengths of such relationships. As was shown with the correlation coefficient, even when statistically significant relationships are found, they are not always meaningful. The value given by r^2 is a measure of *effect size*. It indicates the proportion of variability in one variable that can be determined by its relationship with another variable. Stated another way, it shows how much variance in one variable is explained by differences in another variable. In addition to r^2, there are several

other measures of effect size that can be used in conjunction with *t*-tests and ANOVAs. Two such measures will be presented here, *Cohen's d* for *t*-statistics and eta^2 for *F*-statistics. The interpretation of the values are taken from Cohen [56], who is very cautious about associating values with specific qualitative labels such as *small* and *large*. The use of Cohen's [56] interpretations of effect sizes are standard across a range of behavioral science disciplines and despite Cohen's cautiousness they provide useful heuristics for interpreting effect sizes.

Cohen's *d* can be computed in a number of ways, but the easiest is given below in the first formula, which uses the value of the *t*-statistic and the *df*. This formula assumes that sample sizes of the two groups that are being tested are equal. In cases where this is violated, another version of the formula can be used, which accounts for the differences. This version of the formula (b) includes values for the size of each sample (e.g., if males and females were being compared, one of these values would correspond to the number of male subjects while the other would correspond to the number of female subjects). Typical interpretations of Cohen's *d* are: small = 0.2, medium = 0.5 and large = 0.8.

$$d = \frac{2t}{\sqrt{df}} \quad \text{(a)} \qquad d = \frac{t\,(n_1 + n_2)}{\left(\sqrt{df}\right)\left(\sqrt{n_1 n_2}\right)} \quad \text{(b)}$$

(a) for equal sample sizes or (b) for unequal sample sizes

The computation for eta^2 is also relatively straight-forward: it is the ratio of the between-treatments SS to the total SS. These values are interpreted on a slightly different scale from Cohen's *d*: small = 0.10, medium = 0.25, and large = 0.40.

$$\eta^2 = \frac{SS_{bg}}{SS_T}.$$

11.2.2.8 Non-Parametric Tests

Unlike parametric tests, non-parametric tests make few assumptions about the distribution of variables in the population. Specifically, these tests do not rely on the assumption that variables are distributed normally in the population. Non-parametric tests are also useful for

analyzing discrete data types, such as nominal and ordinal measures. Non-parametric tests can also be considered as more robust since they make fewer assumptions than parametric tests and can be used in more situations. However, it is important to note that in general, non-parametric tests are not as sensitive as parametric tests and thus, the risk of Type II errors are greater. The important thing is for researchers to select the most appropriate test to ensure the credibility and integrity of their results rather than the significance.

There are a number of non-parametric tests that have been used in IIR including the *Mann-Whitney* test, *Wilcoxon Signed-Rank* test, *Kruskal-Wallis* test, *Spearman's Rho*, and *Chi-square*. We will look closely at *Spearman's Rho* and *Chi-square* because they test different data types and relationships than any of the previously discussed tests. The other non-parametric tests are not discussed since they offer non-parametric alternatives to other tests. The *Mann-Whitney* and *Wilcoxon Signed-Rank* tests offer alternatives to the *t*-test, while the *Kruskal-Wallis* test offers an alternative to ANOVA.

Spearman's rho

Spearman's *rho* is correlation coefficient which has been typically used to evaluate ordinal data and to test for relationships that are not necessarily linear. Thus, the Spearman correlation measures the consistency of the relationship between two variables, but it does not say anything about its form. Ordinal level data is often rank-level data. For instance, subjects in a study might be rank-ordered from the best performer to the worst performer. It was noted much earlier in this paper that the Likert-type scale data that is common in IIR evaluations is technically ordinal level data, although it is promoted to interval level status so that more sophisticated analysis can be performed with it. However, when exploring correlation, it is possible to study this data at its native level using Spearman's *rho*.

As a reminder, ordinal level data tells us that one thing is [better or worse] or [more or less] than another, but it does not tell us how much since the distances between points are not constant. This can be easily illustrated with the example above, where subjects are to be ranked according to how well they perform. Ratio level data (performance

score) could be used to create this ranking, but once the ranking was done, we would only know that Subject A was 1, Subject F was 2, Subject B was 3, etc. We would not know how much better Subject A was than Subject F, and there would be no guarantee that the difference in performance scores between Subject A and F was equal to the difference between Subject F and B. Thus, some information is lost when converting ratio level data into ordinal level data.

The calculation of Spearman's *rho* is displayed in Figure 11.9, along with sample data from the IIR Research Scenario that investigates the relationship between familiarity and usability. This formula is actually a simplification of the Pearson's r formula that assumes that scores are ranked. Thus, before using this formula, the raw data must be converted into ranked data. If the original measure is ranked data (e.g., subject ranking according to performance) then this step is not necessary. The data in the example come from two scales and so the scores first need to be transformed into rank values. Although this is not shown in the

	Raw data		Ranking values		Differences	
Subject	X (Familiarity)	Y (Usability)	x_r	y_r	D $(x_r - y_r)$	D^2
1	4	5	7.5	9.5	–2	4
2	1	1	1.5	1.5	0	0
3	3	4	5	6.5	–1.5	2.25
4	1	4	1.5	6.5	–5	25
5	3	1	5	1.5	3.5	12.25
6	3	2	5	3.5	1.5	2.25
7	2	4	3	6.5	–3.5	12.25
8	4	2	7.5	3.5	4	16
9	5	5	9.5	9.5	0	0
10	5	4	9.5	6.5	3	9
						$\Sigma = 83$

Formula:

$$\rho = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$
$$= 1 - \frac{6(83)}{10(100 - 1)}$$
$$= 1 - 0.503$$
$$\rho = 0.497$$

Fig. 11.9 Calculation of Spearman's *rho* using sample familiarity and usability data.

example, the easiest way to do this is to order the scores from smallest to largest and assign rank values to each position. In our example data, there are a few ties — e.g., four subjects used a usability rating of four. When two or more scores are tied, the mean of their ranked positions is computed and assigned to all scores with this value. In the example, two subjects used a usability rating of one. The corresponding rank value for these two subjects is 1.5, or (1+2/2). The differences between each subject's ranked X and Y scores are then computed and squared. The set of differences are then summed. This value is multiplied by six and forms the nominator of the *rho* formula. The denominator is a simple calculation using the sample size.

To determine whether the coefficient is significant and the null hypothesis can be rejected, the t-statistic formula displayed in Figure 11.8, along with the t-distribution is used. It is important to note that the Spearman's *rho* formula loses some of its accuracy when there are a lot of ties; if this is the case, then the researcher might want to explore an alternative coefficient.

Chi-Square

The chi-square test is used to compare the distribution of scores across two or more levels. Figure 11.10 illustrates some sample data that corresponds to the IIR Research Scenario where the researcher asks subjects to indicate which system they liked best. The numbers in this Figure represent the frequency of subjects who selected a particular system as their favorite. Is there a significant difference with respect to which system subjects prefer?

Since we are only looking at the distribution of one variable, our test is for *goodness of fit*, which examines how good the data fit the

	System			
	A	B	C	Total
Observed frequencies (O)	1	1	13	15
Expected frequencies (T)	5	5	5	15

$$\chi^2 = \sum \left[\frac{(O-T)^2}{T} \right] = \frac{(1-5)^2}{5} + \frac{(1-5)^2}{5} + \frac{(13-5)^2}{5} = 19.2$$

Fig. 11.10 Computation of χ^2 for sample system preference data.

distribution specified by the null hypothesis. The null hypothesis states that there are no differences in subjects' system preference. This is represented by the second row of Figure 11.10 — these values are referred to as the expected frequencies. If the null hypothesis is true, then the preference distributions should be roughly equal across system. Unless otherwise specified, the null hypothesis assumes the distributions will be equal across category. However, if it is known that the distributions in the population are unequal, then the expected frequencies can be adjusted. As mentioned earlier in the discussion of z-statistics, it is rarely the case in IIR that we know anything about the population, so the default null is almost always used. The purpose of the chi-square test is to compare the observed distributions with the null expectation. The alternative hypothesis, in this situation, simply states that the population is not divided equally among the various categories.

The formula for, and computation of, chi-square is shown at the bottom of Figure 11.10. This formula is equal to the sum of the squared differences between the observed and expected frequencies divided by the expected frequency. This formula basically measures the discrepancy between the observed frequencies and the expected (or theoretical) frequencies. The value of the chi-square statistic is directly related to the size of the discrepancy — the larger the discrepancy, the larger the chi-square value.

Similar to ANOVA, the chi-square distribution is positively skewed — the majority of scores will cluster around 0–1 — these values represent the null hypothesis. A statistically significant chi-square value will be out in the tail of the distribution. As with all statistical tests, the chi-square statistic also has an associated *df*, which is equal to the number of categories minus 1 $(C - 1)$ (in the example $3 - 1$). This value, along with an alpha level, allows one to enter the table of values that correspond to the chi-square distribution (Table 11.8) [27] to determine if a particular chi-square value is statistically significant. Table 11.8 tell us that the chi-square value for our sample data is beyond the critical value of 5.99 and is therefore, statistically significant. In fact, our chi-square statistic is significant at $p < 0.001$. We would report this as $\chi^2(2) = 19.2$, $p < 0.001$.

Table 11.8 Chi-square distribution.

	Probability values						
df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.3233	2.7055	3.8414	5.0238	6.6349	7.8794	10.828
2	2.7725	4.6051	5.9914	7.3777	9.2103	10.5966	13.816
3	4.1083	6.2513	7.8147	9.3484	11.3449	12.8381	16.266
4	5.3852	7.7794	9.4877	11.1433	13.2767	14.8602	18.467
5	6.6256	9.2363	11.0705	12.8325	15.0863	16.7496	20.515
6	7.8408	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.0371	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2188	13.3616	15.5073	17.5346	20.0902	21.9550	26.125
9	11.3887	14.6837	16.9190	19.0228	21.6660	23.5893	27.877

The chi-square test can also be used to test for independence when the distributions of two variables are being compared. Using our IIR example, we might examine whether there is a relationship among system preference and sex. This is similar in nature to correlation in that each subject has a value on two variables (system preference and sex) except that chi-square examines the frequency distributions since these variables are nominal. The null hypothesis in this case states that the distribution of system preferences will be the same for females as for males, or put another way, the frequency distributions will have the same shape for both females and males.

Some sample data is presented in Figure 11.11, along with the chi-square formula and the chi-square computation for our sample data. Note that when two variables are involved, the expected frequencies are

	System A	System B	System C	Row Totals
Males:				
Observed Frequencies (O)	7	26	7	40
Expected Frequencies (T)	8	19	13	40
Females:				
Observed Frequencies (O)	7	7	16	30
Expected Frequencies (T)	6	14	10	30
Column totals	14	33	23	70 (Grand total)

$$\chi^2 = \sum \left[\frac{(O-T)^2}{T} \right] = \frac{(7-8)^2}{8} + \frac{(26-19)^2}{19} + \frac{(7-13)^2}{13} + \frac{(7-6)^2}{6} + \frac{(7-14)^2}{14} + \frac{(16-10)^2}{10} = 12.74$$

Fig. 11.11 Calculation of χ^2 using sample system preference and sex data.

a function of the characteristics of the sample along these two variables. For instance, in the sample there are 40 males and 30 females. Thus, the distribution of system preferences cannot be equal in terms of an absolute value, but must be equal proportionately. To compute the expected frequencies, first the proportions of subjects selecting each system are computed: System A ($14/70 = 20\%$), System B ($33/70 = 47\%$) and System C ($23/70 = 33\%$). The null hypothesis assumes that these same proportions will be observed for both males and females: Males [System A ($0.20 * 40 = 8$), System B ($0.47 * 40 = 19$), System C ($0.33 * 40 = 13$)] and Females [System A ($0.20 * 30 = 6$), System B ($0.47 * 30 = 14$), System C ($0.33 * 30 = 10$)]. While the computation and interpretation of the chi-square test for independence is the same as that for goodness of fit, the calculation of *df* differ. For tests of independence, $df = (R - 1)(C - 1)$, where R = number of rows and C = number of columns. Using this value in conjunction with the chi-square distribution in Table 11.8, we see that the critical value is 5.99 with alpha = 0.05. Our chi-square statistic is well beyond this and is even significant at the 0.005 level, so we can reject the null hypothesis. We would report this as $\chi^2(2) = 12.74$, $p < 0.005$.

11.2.3 Cohen's Kappa

The final statistic that will be presented is Cohen's Kappa, a measure of inter-rater reliability. Inter-rater reliability (also known as inter-coder reliability) shows the extent to which two or more people agree on how to classify a set of objects. It can be used to check the reliability between relevance assessments made by two people and to check the reliability of how a researcher has analyzed and classified qualitative data. Inter-rater reliability measures provide a much stronger measure of rating consistency than simple percent agreement since they take into account the distribution of responses and the amount of agreement that would happen by chance. Percent agreement is an inflated index of agreement and can be especially misleading when the underlying distributions are skewed.

Cohen's Kappa is not an inferential statistic. Instead, it produces a value similar to a correlation coefficient. The values for Kappa

range from 0–1.00, with larger values indicating better reliability. Most researchers accept Kappa values greater than 0.70 as satisfactory. If the value is less than this, then researchers will often revise classification rules, solicit more raters to apply these rules and then re-assess the statistic. Thus, an important part of this exercise is ensuring that the rules for classification are clear, can be easily used to distinguish between objects, and can be understood and executed by multiple people. Raters do not have to necessarily agree with the classification rules they just have to execute them consistently.

The formula for Cohen's Kappa is given below. To execute this formula, one should first build a contingency table displaying the ratings made by the raters in relation to one another. The diagonal of this table will show the total agreements made by the two raters, while the off-sets will show the disagreements. Row and column totals should be computed as well as expected frequencies (*ef*) for each classification category:

$$\mathrm{K} = \frac{\Sigma a - \Sigma ef}{n - \Sigma ef}$$

where Σa = sum of the agreements (diagonal), n is the total number of objects and

$$ef = \frac{row\_total \cdot col\_total}{overall\_total}$$

The computation of Cohen's Kappa is shown in Figure 11.12. Imagine that two raters have used a four point scale to classify the relevance of 259 documents. In total, the raters have agreed on 167 ratings (see diagonal). A simple percent agreement would show that the raters have agreed 64% of the time. However, the Cohen's Kappa shows the inter-rater agreement to be 12% points less (52%). Since this value is less than the target agreement of 70%, it would be necessary to refine classification rules and perform the ratings again. An examination of the disagreements can help identify where raters are having the most problems. In the example, most disagreements happen between relevance categories two and three. Thus, refining rules for distinguishing among these two types of relevance is the best place to start with the revisions.

Contingency Table

		Rater 1				Row
		1	2	3	4	Totals
Rater 2	1	70	7	5	0	82
	2	5	25	28	2	60
	3	6	22	32	5	65
	4	2	4	6	40	52
Column Totals		83	58	71	47	259

Expected Frequencies (*ef*)

$$ef_1 = \frac{82 * 83}{259} = 26.28$$

$$ef_2 = \frac{60 * 58}{259} = 13.44$$

$$ef_3 = \frac{65 * 71}{259} = 17.82$$

$$ef_4 = \frac{52 * 47}{259} = 9.44$$

$$\Sigma ef = 66.98$$

$$\Sigma a = 167$$

Cohen's Kappa

$$K = \frac{167 - 66.98}{259 - 66.98}$$

$$= \frac{100.02}{192.02}$$

= .52 (52%)

Fig. 11.12 Calculation of Cohen's Kappa using sample data from two raters, who have labeled a set of documents according to four levels of relevance.

11.2.4 Statistics as Principled Argument

The title of this section takes its name from the book by Abelson [1], who describes five properties that determine the persuasiveness of a statistical argument: magnitude, articulation, generality, interestingness and credibility. These properties emphasize that statistics assist with analysis, but their messages have to be interpreted by humans. Abelson is careful to point out that single studies are not definitive, significance tests provide limited information, and interpretation of statistical results is just as important as the statistics themselves.

Magnitude is related to the strength of a result. Even in cases where statistical significance is found one must look critically at the strength of the relationship. In previous sections, it was shown that some statistically significant correlation coefficients are not meaningful. Calculations of effect size are the most common methods for assessing magnitude. If effect sizes are small, then researchers should be more conservative with their interpretations and conclusions.

Articulation is the degree of specificity with which results are presented. An example of poor articulation is when a researcher conducts

an ANOVA to test for differences between groups, but does not conduct follow-up tests to pinpoint differences. The main point about articulation is to understand that what is being studied is typically complex and it is sometimes necessary to look closely at individual cases to understand what is happening rather than relying on the overall statistic.

Generality is the applicability of study results and conclusions to other situations. Researchers typically use reductive methods to examine very narrow problems, which can impact the generality of the results. Abelson advocates for a wide range of investigations that center on the same phenomena. The implications are that a single study should be one of a larger body of research designed to investigate a particular problem. Single study results provide support for particular conclusions, but not definitive conclusions. The accumulation and analysis of data from many different studies designed to investigate the same problem enhance generality.

Interestingness and credibility are attributes of the research story in which statistical arguments are placed. Abelson [1, p. 13] adopts the viewpoint that "for a statistical story to be theoretically interesting, it must have the potential, through empirical analysis, to change what people believe about an important issue". Abselson [1, p. 13] notes that the importance of an issue contributes to its interestingness, where importance is defined as "the number of theoretical and applied propositions needing modification in light of the new result". The criterion of importance is typically used to evaluate the quality of research and good writing practice dictates that authors include some statement of importance in their research reports.

The credibility of research, according to Abelson, should be evaluated according to the soundness of the method and theoretical coherence. The soundness of the method is a quality of research design and data analyses. One of the most common problems in IIR reports is that researchers do not provide enough detail for readers to evaluate the credibility of the method. Without being able to understand the experimental design and procedures, it is nearly impossible to assess the credibility of the results. Credible results depend on a credible research design.

Theoretical coherence is a bit more difficult to discuss in the context of IIR research since at present much of the research is not theoretically based and there is a strong underlying current of applied science rather than basic science. However, if a research claim contrasts with prevailing theory or belief, then the researcher should be prepared to rule out alternative explanations of the findings and demonstrate why the alternative explanation is the most parsimonious. The researcher must also demonstrate the coherence of an alternative theory by showing that it can explain a number of interconnected findings. The burden of proof ultimately lies with the researcher and having statistically significant results does not provide proof, only evidence (provided the method is sound).

12

Validity and Reliability

Validity and reliability assessments can be applied to both the method used to conduct the study as well as to specific measures. Validity is the extent to which methods and measures allow a researcher to get at the essence of whatever it is that is being studied, while reliability is the extent to which the method and measures yield consistent findings. Validity and reliability assessments are particularly important for understanding the overall quality and limitations of a study, and ultimately, the extent to which research results are believable and generalize. All studies can be critiqued in terms of validity and reliability and no study will be free of validity and reliability issues. There is a tension between validity and reliability, so optimizing both of these in a single study is usually not possible. Although measures are technically part of the method, these two concepts will be discussed separately, since there are special validity and reliability issues related to measurement. Method will be used to refer to the specific procedures used to conduct the study and measures will refer to instruments and metrics.

There are two broad validity classes: internal validity and external validity. Internal validity is related to the quality of what happens during the study. One of the most common threats to internal validity is

instrumentation which is related to the quality of the instruments and measures. If an instrument yields poor or inaccurate data, then the results of the study are unlikely to be valid. External validity is the extent to which results from a study generalize to the real world. A study may have good internal validity, but the results may be meaningless outside the particular experimental situation. Thus, internal validity is a necessary but not sufficient condition for external validity.

Certain methods are associated with certain levels of validity and reliability, regardless of how they are executed. In addition, each individual study will have validity and reliability issues that are specific to how that particular study is executed. Laboratory studies are generally thought to be less valid, but more reliable than naturalistic studies. Laboratory studies typically involve artificial situations that are tightly controlled by the researcher. As a result, it is questionable whether the behavior exhibited by subjects in a laboratory study is the same as the behavior they would exhibit in a natural environment. However, with a detailed study protocol and certain number of controls, it is possible for each subject to experience the evaluation situation in a similar way, so laboratory studies are generally characterized as having high reliability. Naturalistic studies provide a more realistic view of subjects' searching behaviors, but such studies are not typically controlled and so it is impossible to ensure that each person experiences the study in the same way.

Demand effects and reactivity are important concerns in studies with human subjects. The research context can demand that subjects behave in particular ways. This includes who administers the experiment, where it is administered and how it is administered. One specific type of demand effect is experimenter demand effects where the experimenter (either consciously or unconsciously) communicates to subjects how they should behave. The experimenter, in most cases, knows a lot more about the study than the subjects, including the desired outcome. Thus, there is a danger that this knowledge is communicated either implicitly or explicitly during the experimenter–subject interactions. In some research, safeguards are put in place to protect against this. For instance, in medical research double-blind experiments are common. In this situation, neither the subject nor researcher knows

who is assigned to the experimental condition. In other disciplines, it is customary for those administering the study to be ignorant of the goals, objectives and hypotheses of the study. It is also the case that subjects might try to guess the purpose of the study and act accordingly to please the experimenter even if this does not reflect their behaviors and desires.

Reactivity refers to the situation where people know they are being observed so they modify their behavior. One specific type of reactivity that has been discussed and debated a lot in the behavioral sciences is the Hawthorne effect. This is a form of reactivity where subjects change their behaviors temporarily (usually in a positive way) because someone is paying attention to them. While it may not be possible to control demand and reactivity effects in all situations, the important thing is to be mindful that they occur and take steps to prevent them if possible since they can potentially impact both the validity and reliability of study results.

Validity and reliability are also related to study procedures. It may be the case that the order in which study activities are carried out changes the validity and/or reliability of the data that is collected. For instance, in the IIR Research Scenario, subjects were asked to indicate their familiarities with different search topics. *When* this question is asked is likely to impact subjects' responses. If this question is asked after subjects search, then their responses will likely be affected by their experiences searching. In a longitudinal naturalistic study, asking subjects to reflect on their searching activities at monthly intervals is unlikely to yield the same kind of data as asking them to do this at weekly intervals.

Validity and reliability can also be used to critique instruments and measures. In most cases, instruments are used to collect data that will then be used to create measures, so in some ways these things are inextricably linked. Thus, these terms will be used somewhat interchangeable in this section. Instruments that yield qualitative data are generally thought to be more valid, but less reliable than those that yield quantitative data, especially with respect to eliciting information from subjects. For instance, open-ended questions which might be used for interviewing or as part of a questionnaire do not suggest appropriate

answers and topics to subjects or force them to respond in a specific way. Instead, subjects are able to provide any information they feel is relevant and they are able to describe their attitudes and feelings in more ways than just a number. However, it is unlikely that subjects will respond to such questions in the same way at two points in time.

Instrumentation is one of the biggest threats to the internal validity of a study. Consider an example where a researcher uses a logger to record what a user does while searching, but is unaware that the logger is not really recording everything that is happening. Measures computed from the data collected via this logger will not be valid. The instrument and measures may actually be reliable; that is, they will yield consistently invalid results. Thus, it is possible to have an instrument or measure that yields reliable results, but not valid results. Reliability is a necessary (but not sufficient) condition for validity, but the converse is not true.

Instrumentation and measurement are two very big problems in IIR that need increased attention. Instrumentation and measurement are particularly tricky when studying user perceptions, attitudes and behaviors because these things can be influenced by the *process* of instrumentation and measurement (see previous discussion of method variance). In IIR, the questionnaire is one of the most widely used instruments for collecting data from subjects. It is well-known that people exhibit a number of biases when responding to questionnaire items, including social desirability responding and acquiescence. It is also well-known that people are sensitive to characteristics of measurement tools and the contexts in which they are used. Such biases are a huge source of measurement error, which poses a serious threat to the internal validity of a study. Despite this, most of the questionnaires and scales that are used in IIR do not have established validity and reliability and are often developed *ad-hoc*. While there are some items and scales that appear in many studies and could be characterized as a core question set, most of these items have not undergone any significant validity and reliability testing. They have become the core by default rather than because of their specific properties.

Theoretically, the validity and reliability of all measures should be established. This is not a trivial endeavor and requires a number of

studies designed exclusively around the measure. Practically speaking, the majority of IIR measures do not have demonstrated validity and reliability, although many have *face validity*. With respect to measurement, four major types of validity can be evaluated: face validity, predictive validity, construct validity and content validity. *Face validity* is not evaluated formally, but is related to whether the measure makes sense and is acceptable to a community of researchers. For instance, using shoe size as a measure of system usability has no face validity — it does not make sense to use such a measure as a surrogate for usability. Face validity, in many ways, is socially constructed and dependent on researcher consensus.

Predictive validity is the extent to which a measure predicts a person's behavior. For instance, if a person scores high on a college entrance exam, we would expect that person to do well in college. Thus, predictive validity looks at the relationship between the item used to measure a behavior and the behavior itself. *Construct validity* is the extent to which a measure makes sense within the context of other measures that are related to it. For instance, if a researcher develops a new item for measuring ease of use, responses to this item should be in accord with responses to other related items about ease of use. If there are five other ease of use items, responses to the new item should classify the system in a way similar to the other items. Otherwise, they are probably not all measuring the same thing. Finally, *content validity* is related to the extent to which a measure covers the range of possible meanings of the concept that it is purported to measure. Usability offers another good example to illustrate this type of validity — if we only used a single item to measure usability this measure would not have good content validity because the concept of usability is known to be complex and multi-faceted.

While validity is primarily concerned with whether or not a measure adequately captures the essence of a concept, reliability is primarily concerned with whether or not the measure yields consistent findings. Issues of reliability are complicated in situations where data is self-reported by subjects and where the researcher is the instrument. In these situations, personal bias, response bias, memory and demand effects can impact reliability. One of the best approaches to

measurement is to use instruments with established reliability. Many behavioral science disciplines have collections of established measures, some of which will be more or less appropriate to IIR depending on the focus of the study. There are some usability scales and measures that have established reliability in the human–computer interaction and business information system literatures. However, the appropriateness of these measures to IIR systems should be closely evaluated; it is likely that there are other things that need to be evaluated in the IIR situation, so the measures may actually lose some of their content validity when applied to IIR.

There are many ways to explore and establish the validity and reliability of measures, but a discussion of these techniques is beyond the scope of this paper. If a researcher is interested in establishing a new method or measure, further reading about different techniques is recommended (see [105, 286]).

13

Human Research Ethics[1]

Any research that involves human subjects necessarily requires some special ethical considerations. While ethical guidelines vary greatly from country to country and even from institution to institution, the goal of this section is to present general ethical issues that are relevant to IIR research and to stimulate more serious discussion of these issues.

Discussions of research ethics, especially in an international context, are difficult because ideas of what is right and wrong are fundamentally social constructs and can vary widely from culture to culture. Many countries do not have ethics review boards. Moreover, formalized ethics review processes are often viewed negatively by many researchers. Some view the ethics review process as mere institutional bureaucracy. Some assume that there are no ethical issues associated with IIR

[1] Much of what is presented in this section is based on my participation on the Institutional Review Board (IRB) at the University of North Carolina. Through this participation, I engage regularly in discussion, debate and reflection about the ethics of a wide-variety of research studies and procedures. It should come as no surprise that I am a strong proponent and defender of human research ethics and IRBs. This perspective will be clear in this section. It is also important to note that this paper is written from the perspective of an academic researcher working in the United States.

research since subjects are not being injected with fluids or consuming experimental medicines. In most cases, chances of physical harm to subjects in IIR evaluations are practically non-existent, but this does not mean that the ethics of IIR research should not be reviewed or that no risks exist. The principle risk to subjects in IIR evaluations is psychological harm. For example, a subject who is unable to use a retrieval system successfully may become distressed or leave the study feeling like a failure.

Providing the best possible protection to human subjects should be taken seriously since research would not be possible without them. We should be proactive with respect to evaluating the ethics of our own research and not wait for someone else to identify possible problems. As researchers, we have a responsibility to monitor our actions critically and develop ethical standards and principles for our specific research context.

13.1 Who is a Human Subject?

All of the studies depicted in Figure 2.1 with the exception of those at the systems end involve human subjects. A human subject is defined by US Federal Regulation as "a living individual about whom an investigator conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information" (45 CFR 46.102(f)). Thus, according to the first part, all studies that involve humans — whether humans are studied directly or helping to develop research infrastructure by providing relevance assessments — are obliged to be reviewed. One might argue that a study where people provide relevance assessments is not the same as a study where humans are given experimental drugs or a study where humans' search behaviors are logged. Of course, these types of studies are not the same, but what is common is that each study cannot be done without the participation of humans. One might also argue that the point of one's IIR study is to evaluate the system and not the person, but this is missing the point — if you are using people to evaluate your system then you are necessarily studying search behavior and therefore have

an obligation to protect your participants' welfare.[2] Even the rights of TREC assessors must be protected!

13.2 Institutional Review Boards

The primary way that research involving human subjects is reviewed in the United States at academic institutions is via Institutional Review Boards (IRBs). The main purpose of IRBs is to "protect the rights and welfare of research subjects and to function as a kind of ethics committee focusing on what is right and wrong and on what is desirable or undesirable. The IRB is thus required to make judgments about what individuals and groups ought to do and how they ought to do it" [8, pp. 7–8].

Many countries have bodies similar to the IRB, especially for medical research, and there are two international documents that discuss research ethics, the Nuremberg Code[3] and the Declaration of Helsinki.[4] Although the Nuremberg Code was a response to so-called medical research by Nazi physicians, it has some important things to say about how subjects should be treated generally. Specifically, the Code articulated the requirement for voluntary and informed consent, a risk/benefit analysis that favored benefits, and the right to withdraw from the study without penalty. These three requirements form the basis of the current review process and researchers are obliged to make provisions to ensure that these things are met, regardless of the risks involved through participation. The Declaration of Helsinki, which was drafted in 1964 by the World Medical Association, extended the Nuremberg Code. This Declaration extended the Code by stating that the interests of the subject should always be given a higher priority than those of society.

In the United States, there is great variability in how different Boards operate, but there are some fundamental policies and procedures that IRBs must adhere to that are mandated by the

[2] Review happens at many different levels and research that only has humans make relevance assessments would likely be exempt from extensive review.
[3] http://www.ushmm.org/research/doctors/Nuremberg_Code.htm
[4] http://www.wma.net/e/policy/17-c_e.html

National Research Act of 1974. IRBs are formally defined by a federal regulation that describes how government agencies must operate. Since many institutions accept funding from the federal government, they are required to follow the federal regulation to institute ethics review boards to ensure the welfare and protection of human subjects. Regardless of whether a researcher accepts federal funding or not, most academic institutions require review of all research involving human subjects.

The National Research Act of 1974 was passed in response to a series of unethical events involving research with human subjects and the growing concern that human subjects were being exploited and harmed regularly in research. This response was not only motivated by events in medical research, but also research in social and behavioral sciences. The National Research Act of 1974 defined IRBs as they are currently known and established the policies and procedures that such Boards must follow when reviewing research proposals. The Act also resulted in the creation of the Belmont Report which identified three primary ethical principles that should guide the conduct of research with human subjects [228]. These principles form the basis of rules and regulations IRBs in the United States use to evaluate the ethics of research proposals.

13.3 Guiding Ethical Principles

The Belmont Report established three principles: (1) respect for persons (2) beneficence, and (3) justice. These principles are only presented briefly in this paper and it is important to note that the issues related to each principle are much more complex and nuanced than this presentation permits.

Of the three principles, *respect for persons* is the one with which IIR researchers should be most concerned. Respect for persons incorporates two ethical convictions. First, individuals should be treated as autonomous agents, and second, persons with diminished autonomy are entitled to special protection. In other words, people should be able to exercise free-will with respect to both joining and exiting a study, and people who are in positions where this free-will might be compromised

should be given special attention (e.g., those who have cognitive disabilities, children, prisoners, and those with very low education). Diminished autonomy can also be a function of the person's relationship with the researcher. For instance, if a researcher is also a teacher, students in the researcher's class are said to have diminished autonomy because of the researcher's role as teacher — the teacher has a conflict of interest, and the student may feel coerced. Asking one's students to be study participants is viewed by many as unethical. Coercion is anything that compromises a person's ability to exercise free-will and act voluntarily. This further includes offering unjustifiably large sums of money to study subjects.

Many of the issues related to respect for persons are codified in the consent form that subjects sign. The consent that is obtained from subjects is referred to as *informed consent*, which means two things: subjects are told explicitly what will happen during the study and subjects agree to this. Being informed, of course, means that subjects are provided with enough information to make a reasonable choice about whether to participate. This includes letting subjects know how many others will participate (e.g., being 1 of 60 is different from being 1 of 5). It is also important to let people know that they are free to withdraw their participation at any time without penalty and ask that their data be deleted. If compensation is provided, then it should be made clear how this will be handled. Although it is rarely the case that subjects withdraw from IIR studies, it is nevertheless important to let them know that they have this right.

The consent form should also describe to subjects how their privacy will be protected and how confidentiality will be maintained. Researchers have an obligation to protect the privacy of their subjects. A breach of such privacy has the potential to put subjects in harm's way by compromising "a person's reputation, financial status, employability, or insurability, or in some way result in stigmatization or discrimination" [8, p. 28]. It should be clear to subjects in the consent form how their data will be used and how it will be protected since a breach may have serious consequences. It is often difficult for a researcher (or subject) to predict how a breach will be harmful, so great care should be taken to protect data even when it is viewed as

relatively harmless. Things change over time, which is another thing that makes research ethics complex and continual review necessary.

Two of the best ways to protect subjects' privacy is not to associate their names with the data (i.e., do not write a subject's name on the top of a questionnaire), and as soon as the study is finished delete all records containing subjects' names and other identifiable information, such as email addresses. In most IIR studies, there is no reason to maintain records containing identifiable data. In many cases, the only record linking subjects to a specific study is the signed consent form. Because of this, many IRBs grant a *waiver of written consent* for studies that involve minimal risk to subjects. Instead of completing a written consent form, subjects are presented with what is called an *information sheet*, which looks nearly identical to a written consent form, except that it does not have a place for the subject and researcher to sign their names. Researchers can afford subjects with the most protection by not keeping any record of their participation.

Beneficence is concerned with the well-being of research subjects. At the core of this principle is the notion that subjects should not be harmed. A risk/benefit analysis of the potential risk of harm to subjects and the potential benefits of the research is usually conducted. Often, individual subjects do not benefit directly from participating in a study (note that receiving compensation is not a benefit) and benefits are often discussed at a societal level. It is uncommon in IIR evaluations for individual subjects to benefit directly from their participation, although it is conceivable that subjects might learn how to be better searchers. Societal benefits — i.e., helping to create better information access systems — are usually the biggest benefit of IIR evaluations. Although most IIR studies do not involve great risk of harm, it is important to consider the psychological risks associated with participating in a study, as well as the risks involved with a breach of privacy once the data have been collected.

Another issue related to beneficence is the quality of a research project: if a project is so poorly designed that it results in invalid and unreliable data, is this an ethical concern? Although the answer to this question is debatable, it can be viewed as an ethical concern since subjects have been put at some risk (however minimal) through their

participation with no possibility of benefit (even to society) and, more importantly, their time has been wasted. Of course, all studies have flaws. Many studies fail to support research hypotheses, so "waste" both subjects' and researchers' time *inadvertently*. The issue is whether the researcher has done everything to ensure that the study design is the best possible before commencing. When little or no effort or thought has been put into planning and testing a study method, then the question of whether the researcher has violated ethical obligations to subjects can be asked.

The final principle is *justice*, which is related to who bears the burden of the research and who is able to benefit from it. This principle addresses the practice of targeting and recruiting vulnerable subjects, such as poor, uneducated people or people with diminished autonomy, such as prisoners. Although this principle is usually not an issue in IIR studies, we can consider the implications of only studying a particular group of subjects — for instance, university students. While university students arguably bear the heaviest burden in IIR studies (especially for those conducted in academic settings), they are also likely to benefit from this research in the long-term. However, other, less frequently studied groups might also be able to contribute to what we know about IIR. If these users are not studied, then we may miss the needs of these groups and eventually fail to incorporate these into the IIR systems and techniques we develop.

13.4 Some Specific Concerns for IIR Researchers

In addition to some of the general concerns for researchers presented in the preceding section, there are special issues related to the IIR research context that deserve mention. These are issues which will likely be discussed in more detail in the upcoming years. The first issue is related to *user privacy and search logs* [136]. Each type of logging situation presents a different set of concerns. Although many readers will automatically think of the large search logs amassed by search engine companies, logging is a part of most IIR studies and logs contain varying amounts of personal information, which when released might cause harm to users. The logs generated by users who are evaluating

an experimental IR system on a closed corpus with assigned tasks are probably the least risky. As the nature of the study changes from closed corpora and assigned tasks to the open Web and natural tasks, the risk of harm and privacy violations increases and the researcher has a greater responsibility to protect the data and subjects.

In 2006, AOL released search log data to the public and it was immediately clear that individual users could be identified through deductive disclosure [118]. The people at AOL probably had good intentions — they were, after all, releasing a large data set that could be used by researchers — but this case highlighted the sensitivity of search log data and how relatively easy it is to reconstruct individual identities by putting together lots of smaller pieces of information. The problem of deductive disclosure is challenging and was addressed in part by a workshop on query log analysis at the *World Wide Web Conference* in 2007 [10]. Example research from this workshop include a discussion of query log anonymization and solutions [2], and an analysis of query logs with a focus on privacy and the applicability of using existing privacy guidelines such as HIPAA[5] [298]. There is a desire to share collections and shared collections have traditionally been an important part of IR research. However, the de-identification process is much more complicated with personal data and it is likely that more work will need to be done on the anonymization process before such data can be made publicly available.

There are other ethical issues related to logging user interactions that are less obvious. First, consider the principle of informed consent — when users accept cookies or agree to conditions of use for search engine plug-ins, are they really giving *informed* consent [196]? Are such agreements clear about how much privacy the user gives up when using particular Web sites or search engines? Furthermore, while users typically grant permission for researchers to log their interactions, there are many interactions that involve more than one person and third-party disclosure becomes an issue. For instance, a researcher who studies retrieval of email messages may obtain permission from a set of users to log their email, but it is unlikely that the researcher will

[5] http://www.hhs.gov/ocr/hipaa/

obtain permission from all people sending that user email. A similar thing can be said for the use of personal photo and video collections — such items typically contain images of a lot of people. There are also other issues with regard to what people search for and look at online. Researchers who do survey research are required to provide assistance to subjects if they respond in certain ways to certain questions, for example, questions assessing suicidal ideation. What if a subject looks at documents about suicide or bomb-making? Do we have any special ethical obligations to the subject or to society?

Another set of ethical issues stems from crawling and using postings, comments, messages, reviews, interactions, etc. that have been created by users — for instance, as part of myspace.com, amazon.com, and in other environments that promote social interaction. Some of the first researchers to study internet groups and online communication were from the communication and sociology fields [85, 98, 214, 252, 289]). These researchers discovered that collecting and analyzing data that people post online in "public" venues was not as straightforward as it seemed; many people who posted messages became upset with how their messages were used and repurposed. Other controversial methods involved researchers joining groups as legitimate participants only to gather data for research. Although it is likely to change in the future, current review boards first question whether the site or service has any policy that forbids research-related activities, and then looks at the extent to which users have to authenticate to participate. There is often a distinction made between public and private when authentication is required because it can potentially change users' expectations about how their information will be used and who will consume it.

Another important issue that has not been discussed much in IIR is the life and death of data. Historically in IIR, data sets have been retained infinitude because collecting data from users has always been very time-consuming. Indeed, one of the benefits of conducting an IIR study is the collection of a data set that can be used for future investigations. One implication of keeping data sets for perpetuity is that data sets will outlive the researchers who assemble them. Most researchers do not have a specific plan for what will happen to their data sets over the course of their lifetime and subjects do not always have clear

expectations about how long their data will exist and how this may affect them in the future. Ultimately, the researcher is responsible for protecting the data throughout its lifetime (even if this exceeds the lifetime of the researcher) and should articulate in writing a life-plan for the data. With the emerging institutional repository movement at universities, researchers should also think of the implications of turning-over data sets to be stored in such repositories, especially with respect to ownership. If a researcher leaves a particular institution, the data set might have to be left behind. The emergence of institutional repositories and cyber-infrastructure will likely change how proposals are reviewed, how researchers communicate with subjects and how data are stored and protected.

Videotaping subjects also presents special ethical issues [189]. In the past, traditional video recorders were used to capture computer screens while subjects engaged in IIR. Typically such cameras were placed behind the subject and the back of the subject's head was usually visible on the recording. In some cases, the side of the subject's face was also visible. This necessarily changes the risks because the videos contain the likeness of the individual and as long as the videos exist, subjects will be identifiable. Most review boards require researchers to get additional permission from subjects to videotape, which allows subjects to opt-out of the videotaping. Today, most screen recording software allows researchers to record the screen of the computer without recording the subject's likeness. However, video recordings are still in use in many studies for other types of observations. Special considerations also need to be made with respect to audio recordings which are often made during interviews since a person's voice is recorded and it could be used to identify the person. Thus, if audio recordings are captured transcriptions should be made as soon as possible and the tapes destroyed to provide the most protection.

There have been a number of IIR studies that involve *deception* — subjects are lead to believe one thing about what is going on in the study, but in reality another thing is happening. These studies are not usually traditional IIR evaluations, but experimental studies that are more focused on behavior. Examples of deception include manipulating the order of search results, telling subjects that they are using a

particular system when they are not and giving false feedback. Most of the deception involved in IIR studies can be considered minor. Regardless, study proposals that involve deception are looked at more carefully by IRBs because they involve more risks. Moreover, the *informed* part of the consent process is compromised since researchers are unable to obtain consent regarding deception. One essential element in studies that involve deception is a *debriefing* at the end of the study to let subjects know the real purpose of the study and to describe the deception that was involved. Subjects should always be given an opportunity to ask questions at the end of the study and to withdraw their data if they wish.

A final ethical concern for IIR researchers is the extent to which ethics should be considered when reviewing manuscripts for publication. In some disciplines, it is common for researchers to include in published manuscripts the IRB approval number or a statement about how the ethics of the study were evaluated. In IIR, it is taken in good faith that researchers abide by ethical principles and requirements, but where, when and how are such ethical principles communicated and acquired? What guidelines are taken to be the standard? How are young researchers educated about ethics? Ethical responsibilities in general have not extended beyond local review of individual studies. However, it may be the case that ethics review will need to happen at different levels of the research process. Consider research that has been published using the data set released (and retracted) by AOL. Is it ethical to use this data? Should our community publish reports that use this data? If the answer is no, then how should we monitor this?

14

Outstanding Challenges and Future Directions

This paper provided background information about IIR and guidance about how to design and conduct IIR evaluations. However, there were many topics that could not be discussed directly and there are also many outstanding challenges.

14.1 Other Types of Systems

This paper focused on traditional text retrieval systems, although there are currently a great number of systems that support retrieval of different types of information objects including images, video, audio, and personal information; varying types of textual units such as answers and passages; varying languages; and varying devices. There are also a number of systems that use experimental visualization techniques and offer support for a broader range of information-seeking activities (e.g., saving and sorting results). While many of the basic techniques presented in this paper can be applied to different types of IIR systems and use scenarios, each have their own special set of concerns that must be addressed since the nature of the objects being retrieved and the information needs and purposes behind search vary.

Multimedia search includes image, video and audio. One interaction that is very different in these types of systems from traditional text retrieval systems is querying. While in text retrieval systems the objects that are being retrieved match the method of querying, in multimedia systems querying is still often limited to text which does not necessarily match the form of the objects being retrieved. Furthermore, notions of aboutness and relevance can be even more problematic in these settings than in traditional text-based settings. Although not technically a part of multimedia search, interfaces that use experimental visualization techniques also require special consideration [156].

Personal information management (PIM) focuses on a variety of media types, including self-created objects and email, and a variety of specialized tasks such a re-finding and information organization [86, 155]. PIM is an ongoing activity often done in anticipation of future actions (such as re-finding information objects) or expected uses (such as sharing information objects). Because PIM is concerned with information classification and retrieval, it has many things in common with IIR. However, two important differences are that a variety of types of information objects and systems might be examined in a single study and that the information is personal. The implications of this are that evaluations often have to be more flexible and tailored to individuals and carried out in naturalistic settings. This makes it more difficult to study causal relationships and to identify findings that generalize.

There are also many sub-areas of IIR that investigate how users search in environments where smaller units of information are retrieved, such as XML fragments, answers, passages and summaries, and where larger units of information are retrieved such as books. Interactive retrieval of documents marked with XML has received considerable attention in the past few years, most notably through the INEX workshop series (e.g., [176, 273]). In these studies, researchers must accommodate search of parts of documents rather than the whole and understand how users make sense of these various fragments. Research investigating interactive question–answering is also emerging as an area that has many interesting opportunities for IIR researchers. Some studies have been done to investigate users' preferences for answer sizes

[179] and to develop evaluation methods and measure for interactive QA systems [127, 164, 169], but in general, less is known about how people interact with and use QA systems. Many QA systems use natural language dialogue to facilitate interaction, which adds another dimension to the evaluation. Summarization technology is developing rapidly and there is no reason to expect that interactive, user-centered evaluations will not be of interest to researchers in this community. Finally, the success of a recent SIGIR workshop on book search [160] has demonstrated a renewed interest in a domain that is rooted in interactive IR evaluation [83].

Another area where there is quite a lot of systems-centered research, but not much user-centered research is cross-language retrieval, although several researchers have made contributions to this area [204, 211]. Cross-language retrieval is not as widespread and common as the other types of retrieval discussed in the preceding paragraphs, so it is hard to identify search tasks and contexts. However, cross-language retrieval is an important and relevant task to many, most notably government intelligence officers. In addition, there are research programs whose goal is to bring together numerous technologies, including summarization, multimedia and cross-language into a single system and some preliminary reports of user-centered evaluations of these systems [300].

This paper did not address evaluation issues associated with adaptive systems and other systems designed to personalize interactions. These types of systems are particularly difficult to evaluate because usually they are designed to be used over long periods of time. A single search session of the kind that typically happens in a standard IIR study simply does not allow such systems to realize their potential. Since search is personalized to individuals, it is also difficult to set-up a general evaluation framework for all subjects. Social search systems are also showing great promise, but introduce additional considerations. In particular, the cold-start and data sparsity problems must be addressed before evaluation can take place. However, once these problems are addressed, many aspects of the standard IIR evaluation model can be applied. With respect to experimental studies of search behaviors, social search creates many opportunities for researchers to test and

apply theories from social psychology to better understand behavior in this context.

Collaborative IR systems that support group information-seeking and retrieval have emerged recently as a popular area in IR (see [102] for an overview; [153, 199]). Researchers in computer supportive cooperative work (CSCW) and educational technologies have studied systems that support collaborative work for some time. While the research from these areas can provide guidance on the design of studies for collaborative IR, there are also a number of issues specific to the IR situation that will need to be addressed. Again, some elements of the standard IIR evaluation model might be effective in this context, but the danger is always that an overuse of such models prevents the development of more appropriate models. There is also an additional type of interaction that must be accounted for — the interaction between the people engaged in collaboration. The future will likely involve not only the development of novel systems for collaborative IR, but also novel evaluation methods and measures, which might be rooted in communication theory and social psychology.

Finally, the evaluation and study of mobile information-seeking and retrieval also introduces its own special issues [109]. The information-seeking needs and behaviors of users, the situations in which searching takes place and the nature of the device and hardware make this type of retrieval different from standard, non-mobile text search.

14.2 Collections

Sharable collections have played an important role in IR and IIR evaluation, but most of the collections that have been used in IIR studies test their limits in terms of generalizability and usability. TREC collections have been used widely in IIR evaluations, but as described earlier, researchers must make some simplifying assumptions about the nature of relevance, the generalizability of relevance assessments and appropriateness of assigned search tasks.

There are several possible directions that IIR research can take with respect to developing shared collections. The first is to determine how collections developed for systems-centered evaluation can be

better used in IIR evaluations. This involves engaging with a number of perennial problems in IIR, including the nature of relevance. The second direction is to create new collections that contain some elements of traditional collections, such as a corpus of documents, but that also contain new elements that are specific to the interactive retrieval situation. Voorhees [287] discusses the difficulties of creating a test collection for adaptive information retrieval.

The third direction is to develop task sets that can be used in different situations. While researchers often use TREC topics as search tasks, a larger variety of tasks that systematically vary across a number of attributes (e.g., difficulty and specificity) would greatly facilitate evaluation and experimentation. The development of shared tasks is more than just penning them. Like any instrument, tasks should undergo a number of tests to ensure that they are representative of the attributes they are purported to embody, that they can be used consistently across a number of situations and that users interpret the tasks in expected ways.

A final direction towards shared collections is shared data sets, which may or may not conform to the traditional definition of a collection. This includes large scale query log data collected by search engines and other large organizations, as well as data collected by researchers who focus on smaller-scale laboratory experiments. Search engine companies in particular, have made some efforts to share log data. More successful examples involve controlled sharing through personal relationships, competitive grant programs for academic researchers and more recently, specialized workshops (e.g., *Workshop on Web Search Click Data*[1]). Sharing with fewer restrictions and across more circumstances may be on the horizon, but privacy issues will likely dictate that some restrictions will always apply.

Many academic researchers have collected detailed log data, often supplemented with self-report and interview data, from subjects which can also act as a type of shared collection. These data sets typically involve few subjects, but contain rich contextual data about needs and behaviors. Some of these data sets are collected in the context of Web

[1] http://research.microsoft.com/∼nickcr/wscd09/

search, while others are collected as part of evaluations of experimental systems. Although there is not a strong tradition or incentive to share such data sets[2] and no real infrastructure to support sharing, a data repository would support numerous kinds of research, including classic IR and IIR, as well as meta-analysis, systematic review, and comparative and historical analysis.

14.3 Measures

One of the most significant measurement challenges is developing performance measures that can be used in interactive search scenarios. There are a number of standard evaluation measures available to those conducting systems-centered IR studies. Many of these have been used in IIR evaluations, but in most cases, the assumptions underlying the measures do not match what happens during interactive searching. Most of these measures assume stable, independent, binary relevance. In interactive search situations, relevance assessments can change throughout the search session and vary based on the presentation order of documents, as well as other contextual factors such as the user's familiarity with the search topic. In interactive situations, relevance assessments are rarely binary and are made by many users (not just a single assessor) who do not always agree on what is relevant. Furthermore, many standard IR evaluation measures are based on a one-to-one correspondence between a query and topic, and only accommodate a single search result list for any given topic. In interactive search scenarios, users typically enter many queries during a search session and view a number of search results lists. This situation introduces a variety of issues including duplicate search results. The work on discounted cumulated gain [149, 150] represents an important step towards the development of performance measures that are better suited to interactive searching, but more measures are needed, especially those that reflect session-based performance.

[2] The American Psychology Association [9, p. 354] includes as an ethical principle of scientific publishing that researchers maintain their data sets for at least five years after publication and make them available to journal publishers and other researchers who might question the findings and/or want to replicate the study.

Better methods for eliciting evaluative data from subjects are also needed. One approach to obtaining better evaluative data from subjects is to identify indirect measures or subsidiary measures that are highly correlated with measures of interest. An example of this is Czerwinski et al.'s [66] subjective duration assessment which was described earlier. Indirect measures are a potentially useful way to address some of the problems with self-report measures, even though many actually rely on self-report. For instance, Czerwinski et al. [66] asked subjects to estimate how long it took them to complete tasks (self-report) and used the differences between these estimates and the real time to determine task difficulty. The underlying assumption is that if subjects were asked to directly respond to a question about task difficulty, that these responses would likely be subject to response bias. It is important to note that the establishment of indirect measures requires careful and thorough investigation. Such measures must be evaluated rigorously against some gold standard, which is often challenging to elicit and/or determine. For instance, to link the discrepancy between a user's time estimate and the actual time with task difficulty requires some baseline measure of task difficulty. The use of the dual monitoring task by Dennis et al. [72] also represents an attempt to use an indirect method for understanding more about subjects' experiences.

Eye-tracking data provides another source of information that can be used to create evaluation measures. Researchers have studied eye movements for well over 100 years — initially in the fields of cognitive psychology and physiology and later in the field of human–computer interaction [143]. Jacob and Karn [143] attribute the first use of eye-tracking in human–computer interaction to Fitts et al.'s [96] study of the movements of airplane pilots' eyes as they used cockpit controls; eye movements were recorded using motion picture cameras. Today, there is better equipment and better theoretical frameworks for understanding eye movements, although there have only been a few studies that have used eye-tracking in IIR research (e.g., [152, 222]). Although the equipment is clearly better today than in the past, it is still expensive and awkward for subjects. Even with the best equipment, subjects often need to sit very still, which can be difficult when searching for multiple tasks during a one hour period. Another difficulty is that large

amounts of data are generated — it can be difficult to analyze and make-sense of this data. However, eye-trackers provide more refined information about how a subject experiences an IIR system and conducts information-seeking and retrieval. This includes more detailed information about which parts of documents subjects view and if subjects cognitively engage with a particular feature or object even when there is no observable log action such as a click. Lorigo et al. [184] provide an overview of eye-tracking and online search and identifies future research directions.

Emotional and affective measures are likely to play an increasingly important role in IIR evaluation. The notion of affective computing has been around for quite some time in the human–computer interaction literature [212], but has not made its way to IIR research [157]. There are certainly studies of users' emotions and affective states during the information-seeking process (e.g., [174]), but researchers have yet to tie specific emotional responses to particular IR interactions and states. Arapakis and Jose [12] recently conducted a study which documented the range of emotions that subjects experience while engaged in laboratory IR tasks. While the ultimate goal of this work is to use emotions as feedback for retrieval, the work suggests that emotional or affective measures might be useful for evaluative purposes. What remains is for someone to develop a theoretical framework for understanding how emotions and affective responses can be used as evaluative feedback and how one might reliably capture such information, whether through facial recognition software or self-report. In addition to measures of affect and emotion, Hassenzahl et al. [124] discuss the notion of hedonic quality. Hedonic qualities are qualities such as originality, aesthetic appeal and innovativeness that do not necessarily have any relation to the task the system is designed to support or the system's performance, but that still contribute to people's experiences and evaluations.

Physiological signals, such as heart rate and perspiration are also potentially rich sources of data about users' experiences and reactions during IIR. Researchers in many disciplines have investigated the relationship between human physiological signals and emotional and mental states. In IIR, such signals might be used as evaluation measures or implicit relevance feedback. Equipment for measuring basic signals such

as skin response and heart rate are relatively inexpensive. Although such equipment is not a normal part of most users' workspaces, physiological sensors are increasingly available and it is not difficult to imagine a world where these types of sensors are a normal part of people's lives. As with eye-tracking data, the biggest challenge with physiological data is analyzing the large number of signals and understanding what they really mean.

15

Conclusion

Reflecting on three decades of IR research, Robertson [220, p. 447] notes, "In the end, any experimental design is a compromise, a matter of balancing those aspects which the experimenter feels should be realistic with those controls which are felt to be necessary for a good result". Similarly, research design in IIR is about making choices; the primary goal of this paper was to catalog and compile material related to methods for the evaluation of interactive information retrieval systems into a single source to help researchers make more informed design decisions. Robertson [220, p. 447] continues, "a field advances not by deciding on a single best compromise, but through different researchers taking different decisions, and the resulting dialectic". The intent of this paper is not to suggest that there is a single best evaluation method or even that evaluation is the only useful type of IIR research — IIR is more than system evaluation and retrieval effectiveness. IIR requires pluralistic approaches and methods. A single, prescribed model would be deleterious.

Despite the length of this paper, many of the presentations were brief; it is hoped that this paper will provide a foundation around which others can discuss methods for studying IIR. This includes the creation

of more detailed reviews of some of the topics discussed in this paper such as IIR history, measures and ethics. People have varying opinions about how IIR evaluation should be conducted. The content of this paper represents one such opinion that is informed heavily by the literature, the author's research experiences and an academic background that is rooted in the behavioral sciences. IIR blends behavioral and computer sciences in an effort to study very complex activities: information search and retrieval. It can be difficult to negotiate these two research traditions and uphold their respective research standards all while maintaining scientific integrity. The length of this paper reflects the complexity, difficulties and nuances of studying IIR and demonstrates why more serious scholarship devoted specifically to methods and measures is needed to further IIR research.

Acknowledgments

I would like to thank Nick Belkin and Paul Kantor for their training and guidance; Justin Zobel, Barbara Wildemuth and Cassidy Sugimoto for their feedback and discussion about this paper; Fabrizio Sebastiani and Jamie Callan for their great patience and encouragement; and three anonymous reviewers for their careful and thoughtful comments.

References

[1] R. P. Abelson, *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum Publishers, 1995.

[2] E. Adar, "User 4XXXXX9: Anonymizing query logs," in *Proceedings of the Workshop on Query Log Analysis: Social and Technological Challenges, at the 16th International World Wide Web Conference*, Banff, Canada, 2007.

[3] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior," in *Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '06*), pp. 3–10, Seattle, WA, 2006.

[4] J. Allan, "HARD track overview in TREC 2003: High accuracy retrieval from documents," in *TREC2003, Proceedings of the 12th Text Retrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Washington, DC: GPO, 2004.

[5] J. Allan, "HARD track overview in TREC 2005: High accuracy retrieval from documents," in *TREC2005, Proceedings of the 14th Text Retrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Washington, DC: GPO, 2006.

[6] B. Allen, "Information needs: A person-in-situation approach," in *Information Seeking in Context: Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts*, (P. Vakkari, R. Savolainen, and B. Dervin, eds.), pp. 111–122, Tampere, Finland, 1997.

[7] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *SIGIR Forum*, vol. 42, pp. 10–16, 2008.

[8] R. Amdur, *Institutional Review Board Member Handbook*. Sudbury, Massachusetts: Jones and Bartlett Publishers, 2003.

[9] American Psychological Association, *Publication Manual of the American Psychological Association*. Washington, DC: APA, Fifth ed., 2001.

[10] E. Amitay, G. C. Murray, and J. Teevan, "Workshop on query log analysis: Social and technological challenges," in *Proceedings of the 16th International World Wide Web Conference*, Banff, Canada, 2007.

[11] P. Anick, "Using terminological feedback for web search refinement: A log based study," in *Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '03*), pp. 88–95, Toronto, CA, 2003.

[12] I. Arapakis and J. Jose, "Affective feedback: An investigation of the role of emotions during an information seeking process," in *Proceedings of the 31st Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '08*), pp. 395–402, Singapore, Malaysia, 2008.

[13] E. Babbie, *The Practice of Social Research*. CA, Wadsworth, 10 ed., 2004.

[14] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz, "Relevance assessment: Are judges exchangeable and does it matter," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '08*), pp. 667–674, Singapore, Malaysia, 2008.

[15] M. J. Bates, "Information search tactics," *Journal of the American Society for Information Science*, vol. 30, pp. 205–214, 1979.

[16] M. M. Beaulieu, "Interaction in information searching and retrieval," *Journal of Documentation*, vol. 56, pp. 431–439, 2000.

[17] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, and P. Williams, "Okapi at TREC-5," in *Proceedings of the 5th Text Retrieval Conference* (*TREC-5*), (E. M. Voorhees and D. K. Harman, eds.), pp. 143–165, Washington, DC: GPO, 1997.

[18] B. Bederson, "Interfaces for staying in the flow," *Ubiquity*, vol. 5, 2004.

[19] N. J. Belkin, "Anomalous states of knowledge as a basis for information retrieval," *Canadian Journal of Information Science*, vol. 5, pp. 133–143, 1980.

[20] N. J. Belkin, "Helping people find what they don't know," *Communications of the ACM*, vol. 43, pp. 58–61, 2000.

[21] N. J. Belkin, A. Cabezas, C. Cool, K. Kim, K. B. Ng, S. Park, R. Pressman, S. Rieh, P. Savage, and I. Xie, "Rutgers interactive track at TREC-5," M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, and P. Williams, "Okapi at TREC-5," in *Proceedings of the 5th Text Retrieval Conference* (*TREC-5*), (E. M. Voorhees and D. K. Harman, eds.), pp. 257–265, Washington, DC: GPO, 1997, 1997.

[22] N. J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool, "Query length in interactive information retrieval," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 205–212, Toronto, Canada, 2003.

[23] N. J. Belkin and A. Vickery, *Interaction in Information Systems: A Review of Research from Document Retrieval to Knowledge-Based Systems*. Library and Information Research Report 35: The British Library, University Press, Cambridge, 1985.

[24] D. J. Bell and I. Ruthven, "Searchers' assessments of task complexity for Web searching," in *Proceedings of the 26th Annual International European Conference on Information Retrieval* (*ECIR 2004*), pp. 57–71, Sunderland, UK, 2004.

[25] J. L. Bennett, "The user interface in interactive systems," *Annual Review of Information Science and Technology*, vol. 7, pp. 159–196, 1972.

[26] J. D. Bernal, "Preliminary analysis of pilot questionnaires on the use of scientific literature," *The Royal Society Scientific Information Conference*, pp. 589–637, 1948.

[27] W. H. Beyer, *Handbook of Tables for Probability and Statistics*. Cleveland, OH: Chemical Rubber Co. Publishers, 1966.

[28] M. Bilenko and R. W. White, "Mining the search trails of surfing crowds: Identifying relevant websites from user activity," in *Proceedings of the 17th International Conference on the World Wide Web* (*WWW '08*), pp. 51–60, Beijing, China, 2008.

[29] A. Blandford, A. Adams, S. Attfield, G. Buchanan, J. Gow, S. Makri, J. Rimmer, and C. Warwick, "The PRET A Rapporter framework: Evaluating digital libraries from the perspective of information work," *Information Processing and Management*, vol. 44, pp. 4–21, 2008.

[30] C. L. Borgman, "End user behavior on an online information retrieval system: A computer monitoring study," in *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '83*), pp. 162–176, Bethesda, MD, 1983.

[31] C. L. Borgman, "All users of information retrieval systems are not created equal: An exploration into individual differences," *Information Processing and Management*, vol. 25, pp. 237–251, 1989.

[32] P. Borlund, "Experimental components for the evaluation of interactive information retrieval systems," *Journal of Documentation*, vol. 56, pp. 71–90, 2000.

[33] P. Borlund, "The concept of relevance in IR," *Journal of the American Society for Information Science*, vol. 54, pp. 913–925, 2003a.

[34] P. Borlund, "The IIR evaluation model: A framework for evaluation of interactive information retrieval systems," *Information Research*, vol. 8, p. 152, 2003b.

[35] P. Borlund and P. Ingwersen, "The development of a method for evaluating interactive information retrieval systems," *Journal of Documentation*, vol. 53, pp. 225–250, 1997.

[36] P. Borlund and P. Ingwersen, "Measure of relative relevance and ranked half-life: Performance indicators for interactive information retrieval," in *Proceedings of the 21st ACM SIGIR Conference on Research and Development of Information Retrieval* (*SIGIR '98*), pp. 324–331, Melbourne, Australia, 1998.

[37] P. Borlund and I. Ruthven, "Introduction to the special issue on evaluating interactive information retrieval systems," *Information Processing and Management*, vol. 44, pp. 1–3, 2008.

[38] P. J. Borlund, W. Schneider, M. Lalmas, A. Tombros, J. Feather, D. Kelly, A. P. de Vries, and L. Azzopardi, *Proceedings of the 2nd International Symposium on Information Interaction in Context*. London, UK, 2008.

[39] B. R. Boyce, C. T. Meadow, and D. H. Kraft, *Measurement in Information Science*. London, UK: Academic Press, Inc, 1994.

[40] J. Bradley, "Methodological issues and practices in qualitative research," *Library Quarterly*, vol. 63, pp. 431–449, 1993.

[41] J. Budzik and K. J. Hammond, "User interactions with everyday applications as context for just-in-time information access," in *Proceedings of the 5th International Conference on Intelligent User Interfaces* (*IUI '00*), pp. 44–51, New Orleans, LA, 2000.

[42] M. Bulmer, *Questionnaires V. 1.* Thousand Oaks, CA: Sage Publications, 2004.

[43] K. Byström, "Information and information sources in tasks of varying complexity," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 581–591, 2002.

[44] K. Byström and P. Hansen, "Conceptual framework for tasks in information studies," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 1050–1061, 2005.

[45] K. Byström and K. Järvelin, "Task complexity affects information seeking and use," *Information Processing and Management*, vol. 31, pp. 191–213, 1995.

[46] J. Callan, J. Allan, J. L. A. Clarke, S. Dumais, D. A. Evans, M. Sanderson, and C. Zhai, "Meeting of the MINDS: An information retrieval research agenda," *SIGIR Forum*, vol. 41, pp. 25–34, 2007.

[47] D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1966.

[48] R. Capra, "Studying elapsed time and task factors in re-finding electronic information," *Personal Information Management, CHI 2008 Workshop*, Florence, Italy, 2008.

[49] D. O. Case, *Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior*. Lexington, KY: Academic Press, 2002.

[50] K. Charmaz, "Qualitative interviewing and grounded theory analysis," in *Handbook of Interview Research: Context and Method*, (J. F. Gubrium and J. A. Holstein, eds.), CA: Sage Publications, 2002.

[51] E. Chatman, "The impoverished life-world of outsiders," *Journal of the American Society for Information Science*, vol. 47, pp. 193–206, 1996.

[52] H. Chen, R. Wigand, and M. Nilan, "Exploring Web users' optimal flow experiences," *Information Technology and People*, vol. 13, pp. 263–281, 2000.

[53] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human–computer interface," in *Proceedings of ACM Human Factors in Computing Systems Conference* (*CHI 1988*), pp. 213–218, 1988.

[54] C. W. Cleverdon, "The Cranfield tests on index language devices," in *Readings in Information Retrieval*, (K. Spark-Jones and P. Willett, eds.), (Reprinted from *Aslib Proceedings*, pp. 173–192.) San Francisco: Morgan Kaufman Publishers, 1997/1967.

[55] C. W. Cleverdon, L. Mills, and M. Keen, *Factors Determining the Performance of Indexing Systems, vol. 1 — Design*. Cranfield, England: Aslib Cranfield Research Project, 1966.

[56] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates, Second ed., 1988.

[57] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS Quarterly*, vol. 19, pp. 189–211, 1995.

[58] C. Cool, "The concept of situation in information science," *Annual Review of Information Science and Technology*, pp. 5–42, 2001.

[59] W. S. Cooper, "Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems," *American Documentation*, vol. 19, pp. 30–41, 1968.

[60] W. S. Cooper, "On selecting a measure of retrieval effectiveness, part 1: The "subjective" philosophy of evaluation," *Journal of the American Society for Information Science*, vol. 24, pp. 87–100, 1973.

[61] P. Cowley, J. Haack, R. Littlefield, and E. Hampson, "Glass Box: Capturing, archiving and retrieving workstation activities," in *Proceedings of the 2nd ACM Workshop on Continuous Archival and Retrieval of Personal Experiences* (*CARPE '05*), pp. 13–18, Santa Barbara, CA, 2006.

[62] F. Crestani and H. Du, "Written versus spoken queries: A qualitative and quantitative comparative analysis," *Journal of the American Society for Information Science and Technology*, vol. 57, pp. 881–890, 2006.

[63] M. Csikszentmihalyi, *Finding Flow: The Psychology of Engagement with Everyday Life*. New York: Basic Books, 1997.

[64] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the experience-sampling method," *Journal of Nervous and Mental Disease*, vol. 175, pp. 526–536, 1987.

[65] M. Czerwinski, E. Horvitz, and E. Cutrell, "Subjective duration assessment: An implicit probe for software usability," in *Proceedings of IHM-HCI 2001 Conference*, pp. 167–170, Lille, France, 2001.

[66] M. Czerwinski, E. Horvitz, and S. Wilhite, "A diary study of task switching and interruptions," in *Proceedings of ACM Human Factors in Computing Systems Conference* (*CHI '04*), pp. 175–182, Vienna, Austria, 2004.

[67] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies: Why and how," in *Proceedings for the 1st International Conference on Intelligent User Interfaces* (*IUI '93*), pp. 193–200, Orlando, FL, 1993.

[68] H. Dang, D. Kelly, and J. Lin, "Overview of the TREC 2007 question answering track," in *TREC2007, Proceedings of the 16th Text Retrieval Conference*, (E. Voorhees and L. P. Buckland, eds.), Washington DC: GPO, 2008.

[69] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, pp. 319–340, 1989.

[70] M. De Mey, "The cognitive viewpoint: Its development and its scope," in *CC77: International Workshop on the Cognitive Viewpoint,* (M. De Mey et al., eds.), pp. xvi–xxxi, Ghent, Belgium: University of Ghent Press, 1977.

[71] S. Debowski, R. Wood, and A. Bandura, "The impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic enquiry," *Journal of Applied Psychology*, vol. 86, pp. 1129–1141, 2001.

[72] S. Dennis, P. Bruza, and R. McArthur, "Web searching: A process-oriented experimental study of three interactive search paradigms," *Journal of the*

American Society for Information Science and Technology, vol. 53, pp. 120–133, 2002.

[73] N. K. Denzin and Y. S. Lincoln, *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications, 2000.

[74] B. Dervin, "From the mind's eye of the user: The sense-making qualitative-quantitative methodology," in *Qualitative Research in Information Management*, (R. Glazier, ed.), pp. 61–84, Englewood, CO: Libraries Unlimited, 1992.

[75] B. Dervin, "Given a context by any other name: Methodological tools for taming the unruly beast," in *Information Seeking in Context: Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts*, pp. 13–38, Tampere, Finland, 1996.

[76] A. Dillon, "User analysis in HCI: The historical lesson from individual differences research," *International Journal of Human–Computer Studies*, vol. 45, pp. 619–637, 1996.

[77] W. D. Dominick and W. D. Penniman, "Automated monitoring to support the analysis and evaluation of information systems," in *Proceedings of the 2nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '79*), pp. 2–9, Dallas, TX, 1979.

[78] P. Dourish, "What we talk about when we talk about context," *Personal and Ubiquitous Computing*, vol. 8, pp. 19–30, 2004.

[79] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff I've Seen: A system for personal information retrieval and re-use," in *Proceedings of the ACM Conference on Research and Development in Information Retrieval* (*SIGIR '03*), pp. 72–79, Toronto, Canada, 2003.

[80] S. T. Dumais and N. J. Belkin, "The TREC interactive tracks: Putting the user into search," in *TREC: Experiment and Evaluation in Information Retrieval*, (E. M. Voorhees and D. K. Harman, eds.), pp. 123–153, Cambridge, MA: MIT Press, 2005.

[81] M. Dunlop, "Time, relevance and interaction modeling for information retrieval," in *Proceedings of the 20th ACM Conference on Research and Development in Information Retrieval* (*SIGIR '97*), pp. 206–213, Philadelphia, PA, 1997.

[82] M. D. Dunlop, C. W. Johnson, and J. Reid, "Exploring the layers of information retrieval evaluation," *Interacting with Computers*, vol. 10, pp. 225–236, 1998.

[83] D. E. Egan, J. R. Remde, L. M. Gomez, T. K. Landauer, J. Eberhardt, and C. C. Lochbaum, "Formative design-evaluation of SuperBook," *ACM Transactions on Information Systems*, vol. 7, pp. 30–57, 1989.

[84] M. Eisenberg, "Measuring relevance judgments," *Information Processing and Management*, vol. 24, pp. 373–389, 1988.

[85] D. Elgesem, "What is special about the ethical issues in online research?," *Ethics and Information Technology*, vol. 4, pp. 195–203, 2002.

[86] D. Elsweiler and I. Ruthven, "Towards task-based personal information management evaluations," in *Proceedings of the 30th ACM Conference on Research and Development in Information Retrieval* (*SIGIR '03*), pp. 22–30, Amsterdam, The Netherlands, 2007.

[87] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: The MIT Press, Revised ed., 1993.

[88] C. H. Fenichel, "Online searching: Measures that discriminate among users with different types of experience," *Journal of the American Society for Information Science*, vol. 32, pp. 23–32, 1981.

[89] K. D. Fenstermacher and M. Ginsburg, "Client-side monitoring for Web mining," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 625–637, 2003.

[90] R. Fidel, "Online searching styles: A case-study-based model of searching behavior," *Journal of the American Society for Information Science*, vol. 35, pp. 211–221, 1984.

[91] R. Fidel, "Qualitative methods in information retrieval research," *Library and Information Science Research*, vol. 15, pp. 219–247, 1993.

[92] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," *Transactions on Information Systems*, vol. 20, pp. 116–131, 2002.

[93] K. E. Fisher, S. Erdelez, and L. E. F. McKenchnie, *Theories of Information Behavior*. Medford, NJ: Information Today, 2005.

[94] D. W. Fiske, "Convergent-discriminant validation in measurements and research strategies," in *Forms of Validity in Research*, (D. Brinbirg and L. H. Kidder, eds.), pp. 77–92, San Francisco: Jossey-Bass, 1982.

[95] S. T. Fiske, "Mind the gap: In praise of informal sources of formal theory," *Personality and Social Psychology Review*, vol. 8, pp. 132–137, 2004.

[96] P. M. Fitts, R. E. Jones, and J. L. Milton, "Eye movements of aircraft pilots during instrument-landing approaches," *Aeronautical Engineering Review*, vol. 9, pp. 24–29, 1950.

[97] B. N. Flagg, *Formative Evaluation for Educational Technologies*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.

[98] S. Flicker, D. Haans, and H. Skinner, "Ethical dilemmas in research on Internet communities," *Qualitative Health Research*, vol. 14, pp. 124–134, 2004.

[99] N. Ford, D. Miller, and N. Moss, "The role of individual differences in Internet searching: An empirical study," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 1049–1066, 2001.

[100] N. Ford, D. Miller, and N. Moss, "Web search strategies and human individual differences: Cognitive and demographic factors, Internet attitudes and approaches," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 741–756, 2005.

[101] N. Ford, T. D. Wilson, A. Foster, D. Ellis, and A. Spink, "Information seeking and mediated searching. Part 4: Cognitive styles in information seeking," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 728–735, 2002.

[102] J. Foster, "Collaborative information seeking and retrieval," *Annual Review of Information Science and Technology*, vol. 40, pp. 329–356, 2006.

[103] L. Freund and E. Toms, "Revisiting informativeness as a process measure for information interaction," in *Proceedings of the Web Information-Seeking and Interaction* (*WISI*) *Workshop at the 30th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval (*SIGIR '07*), pp. 33–36, Amsterdam, The Netherlands, 2007.

[104] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human–system communication: An analysis and a solution," *Communications of the ACM*, vol. 30, pp. 964–971, 1987.

[105] R. M. Furr and V. R. Bacharach, *Psychometrics: An Introduction*. Sage Publications, Inc, 2007.

[106] J. A. Ghani and S. P. Deshpande, "Task characteristics and the experience of optimal flow in Human–Computer interaction," *Journal of Psychology*, vol. 128, pp. 381–391, 1994.

[107] J. A. Ghani, R. Supnick, and P. Rooney, "The experience of flow in computer-mediated and in face-to-face groups," in *Proceedings of International Conference on Information Systems* (*ICIS 1991*), pp. 229–237, New York, NY, 1991.

[108] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine, 1967.

[109] A. Göker and H. Myrhaug, "Evaluation of a mobile information system in context," *Information Processing and Management*, vol. 44, pp. 39–65, 2008.

[110] F. J. Gravetter and L. B. Wallnau, *Statistics for the Behavioral Sciences*. Thomson Learning, Fifth ed., 1999.

[111] C. Grimes, D. Tang, and D. M. Russell, "Query logs alone are not enough," in *Proceedings of the Workshop on Query Log Analysis: Social and Technology Challenges at the 16th International World Wide Web Conference*, Banff, Canada, 2007.

[112] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey, "The validity of the stimulated retrospective think-aloud method as measured by eye tracking," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 1253–1262, Montreal, Canada, 2006.

[113] J. P. Guilford, *Fundamental Statistics in Psychology and Education*. New York: McGraw Hill, 1956.

[114] J. Gwizdka, "Revisiting search task difficulty: Behavioral and individual difference measures," *Proceedings of the 71th Annual Meeting of the American Society for Information Science and Technology* (*ASIS and T*), 2008.

[115] D. F. Haas and D. H. Kraft, "Experimental and quasi-experimental designs for research in information science," *Information Processing and Management*, vol. 20, pp. 229–237, 1984.

[116] K. Halttunen and K. Järvelin, "Assessing learning outcomes in two information retrieval learning environments," *Information Processing and Management*, vol. 41, pp. 949–972, 2005.

[117] P. A. Hancock and N. Meshkati, *Human Mental Workload*. The Netherlands: Elsevier Science Publishers, 1988.

[118] S. Hansell, AOL removes search data on vase group of Web users, New York Times, Friday, March 14, 2008. Business Section. http://query.nytimes.com/gst/fullpage.html?res=9504E5D81E3FF93BA3575BC0A9609C8B63, 2006.

[119] S. Harada, M. Naaman, Y. J. Song, Q.-Y. Wang, and A. Paepcke, "Lost in memories: Interacting with photo collections on PDAs," in *Proceedings of*

the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (*JCDL '04*), pp. 325–333, Tuscon, AZ, 2004.

[120] D. K. Harman, "Introduction to special issue on evaluation issues in information retrieval," *Information Processing and Management*, vol. 28, pp. 439–440, 1992.

[121] D. K. Harman, "The TREC test collection," in *TREC: Experiment and Evaluation in Information Retrieval*, (E. M. Voorhees and D. K. Harman, eds.), pp. 21–52, Cambridge, MA: MIT Press, 2005.

[122] S. G. Hart and L. E. Staveland, "Development of a NASA-TLX (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, (P. Hancock and N. Meshkati, eds.), pp. 139–183, The Netherlands: Elsevier Science Publishers, 1988.

[123] S. P. Harter and C. A. Hert, "Evaluation of information retrieval systems: Approaches, issues and methods," *Annual Review of Information Science and Technology*, vol. 32, pp. 3–94, 1997.

[124] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner, "Hedonic and ergonomic quality aspects determine a software's appeal," *Proceedings of the ACM Conference on Human Factors in Computing Systems* (*CHI '00*), pp. 201–208, 2000.

[125] D. Hawking, P. Bailey, and N. Craswell, "Efficient and flexible search using text and metadata," *CSIRO Mathematical and Information Sciences Tech Report, 2000-83*, available online at http://es.csiro.au/pubs/hawking_tr00b.pdf, 2000.

[126] M. D. Heine, "Simulation, and simulation experiments," in *Information Retrieval Experiment*, (K. Spärck-Jones, ed.), pp. 179–198, London, UK: Butterworths and Co. Ltd, 1981.

[127] W. Hersh, "Evaluating interactive question answering," in *Advances in Open Domain Question Answering*, (T. Strzalkowski and S. Harabagiu, eds.), pp. 431–455, Dordrecht, The Netherlands: Springer, 2006.

[128] W. Hersh and P. Over, "Introduction to a special issue on interactivity at the Text Retrieval Conference (TREC)," *Information Processing and Management*, vol. 37, pp. 365–367, 2001.

[129] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson, "Do batch and user evaluations give the same results?," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '00*), pp. 17–24, Athens, Greece, 2000.

[130] W. R. Hersh, D. L. Elliot, D. H. Hickam, S. L. Wolf, A. Molnar, and C. Leichtenstein, "Towards new measures of information retrieval evaluation," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '95*), pp. 164–170, Seattle, WA, 1995.

[131] S. Hirsh and J. Dinkelacker, "Seeking information in order to produce information: An empirical study at Hewlett Packard Labs," *Journal of the American Society for Information Science and Technology*, vol. 55, pp. 807–817, 2004.

[132] K. Hornbaek, "Current practice in measuring usability: Challenges to usability studies and research," *International Journal of Human–Computer Studies*, vol. 64, pp. 79–102, 2005.

[133] K. Hornbæk and E. L.-C. Law, "Meta-analysis of correlations among usability measures," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 617–626, San Jose, CA, 2007.

[134] I. Hsieh-Yee, "Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers," *Journal of the American Society for Information Science*, vol. 44, pp. 161–174, 1993.

[135] M. Huang and H. Wang, "The influence of document presentation order and number of documents judged on users' judgments of relevance," *Journal of the American Society for Information Science*, vol. 55, pp. 970–979, 2004.

[136] G. Iachello and J. Hong, "End-user privacy in human–computer interaction," *Foundations and Trends in Human–Computer Interaction*, vol. 1, pp. 1–137, 2007.

[137] P. Ingwersen, *Information Retrieval Interaction*. London: Taylor Graham, 1992.

[138] P. Ingwersen, "Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory," *Journal of Documentation*, vol. 52, pp. 3–50, 1996.

[139] P. Ingwersen and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context*. Dordrecht, The Netherlands: Springer, 2005.

[140] P. Ingwersen and P. Willett, "An introduction to algorithmic and cognitive approaches for information retrieval," *Libri*, vol. 45, pp. 160–177, 1995.

[141] International Standards Office, *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part II, Guidance on Usability* (ISO 9241-11:1998). 1998.

[142] E. Jacob, "Qualitative research traditions: A review," *Review of Educational Research*, vol. 57, pp. 1–50, 1987.

[143] R. J. K. Jacob and K. S. Karn, "Eye tracking in human–Computer interaction and usability research: Ready to deliver the promises (section commentary)," in *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, (J. Hyona, R. Radach, and H. Deubel, eds.), pp. 573–605, Amsterdam, Elsevier Science, 2003.

[144] B. J. Jansen, "Search log analysis: What it is, what's been done, how to do it," *Library and Information Science Research*, vol. 28, pp. 407–432, 2006.

[145] B. J. Jansen, R. Ramadoss, M. Zhang, and N. Zang, "Wrapper: An application for evaluating exploratory searching outside of the lab," in *Workshop on Evaluating Exploratory Search Systems at the 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR '06)*, Seattle, WA, 2006.

[146] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing and Management*, vol. 42, pp. 248–263, 2005.

[147] K. Järvelin, "An analysis of two approaches in information retrieval: From frameworks to study designs," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 971–986, 2007.

[148] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *Proceedings of the 23rd ACM SIGIR Conference on Research and Development of Information Retrieval* (*SIGIR '00*), pp. 41–48, Athens, Greece, 2000.

[149] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems* (*TOIS*), vol. 20, pp. 422–446, 2002.

[150] K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen, "Discounted cumulated gain based evaluation of multiple-query IR sessions," in *Proceedings of the 30th European Conference on Information Retrieval*, Glasgow, Scotland, 2008.

[151] H. R. Jex, "Measuring mental workload: Problems, process and promises," in *Human Mental Workload*, (P. Hancock and N. Meshkati, eds.), pp. 5–39, The Netherlands: Elsevier Science Publishers, 1988.

[152] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the 28th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '05*), pp. 154–161, Salvador, Brazil, 2005.

[153] H. Joho, D. Hannah, and J. M. Jose, "Comparing collaborative and independent search in a recall-oriented task," in *Proceedings of the 2nd IIiX Symposium on Information Interaction in Context*, pp. 89–96, London, UK, 2008.

[154] H. Joho and J. M. Jose, "Effectiveness of additional representations for the search result presentation on the Web," *Information Processing and Management*, vol. 44, pp. 226–241, 2008.

[155] W. Jones and J. Teevan, *Personal Information Management*. Seattle, WA: University of Washington Press, 2007.

[156] C.-A. Julien, J. E. Leide, and F. Bouthillier, "Controlled user evaluations of information visualization interfaces for text retrieval: Literature review and meta-analysis," *Journal of the American Society for Information Science and Technology*, vol. 59, pp. 1012–1024, 2008.

[157] H. Julien, L. E. F. McKechnie, and S. Hart, "Affective issues in library and information science systems work: A content analysis," *Library and Information Science Research*, vol. 27, pp. 453–466, 2005.

[158] M. Käki and A. Aula, "Controlling the complexity in comparing search user interfaces via user studies," *Information Processing and Management*, vol. 44, pp. 82–91, 2008.

[159] J. Kalgren and K. Franzen, "Verbosity and interface design," Retrieved on 01 February 2008 at http://www.ling.su.se/staff/franzen/irinterface.html, 1997.

[160] P. Kantor, G. Kazai, N. Milic-Frayling, and R. Wilkinson, "Proceedings of the 2008 ACM workshop on research advances in large digital book repositories," in *Proceedings of the Conference on Information and Knowledge Management* (*CIKM '08*), Napa, CA, 2008.

[161] P. B. Kantor, "Evaluation of and feedback in information storage and retrieval systems," *Annual Review of Information Science and Technology*, vol. 17, pp. 99–120, 1982.

[162] J. Kekäläinen and K. Järvelin, "Using graded relevance assessments in IR evaluation," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 1120–1129, 2002.

[163] M. Kellar, C. Watters, and M. Shepherd, "A field study characterizing Web-based information-seeking tasks," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 999–1018, 2007.

[164] D. Kelly and X. Fu, "Elicitation of term relevance feedback: An investigation of term source and context," in *Proceedings of the 29th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '06*), pp. 453–460, Seattle, WA, 2006.

[165] D. Kelly, J. Harper, and B. Landau, "Questionnaire mode effects in interactive information retrieval experiments," *Information Processing and Management*, 2008.

[166] D. Kelly, P. B. Kantor, E. L. Morse, J. Scholtz, and Y. Sun, "User-centered evaluation of interactive question answering systems," in *Proceedings of the Workshop on Interactive Question Answering at the Human Language Technology Conference* (*HLT-NAACL '06*), pp. New York, NY, 2006.

[167] D. Kelly and J. Lin, "Overview of the TREC 2006 ciQA task," *SIGIR Forum*, vol. 41, pp. 107–116, 2007.

[168] D. Kelly, C. Shah, C. R. Sugimoto, E. W. Bailey, R. A. Clemens, A. K. Irvine, N. A. Johnson, W. Ke, S. Oh, A. Poljakova, M. A. Rodriguez, M. G. van Noord, and Y. Zhang, "Effects of performance feedback on users' evaluations of an interactive IR system," in *Proceedings of the 2nd Symposium on Information Interaction in Context* (*IIiX*), pp. 75–82, London, UK, 2008.

[169] D. Kelly, N. Wacholder, R. Rittman, Y. Sun, P. Kantor, S. Small, and T. Strzalkowski, "Using interview data to identify evaluation criteria for interactive, analytical question answering systems," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 1032–1043, 2007.

[170] J. Kim, "Task as a predictable indicator for information seeking behavior on the Web," Ph.D. Dissertation, Rutgers University, 2006.

[171] K.-S. Kim and B. Allen, "Cognitive and task influences on Web searching behavior," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 109–119, 2001.

[172] S. Kim and D. Soergel, "Selecting and measuring task characteristics as independent variables," *Proceedings of the American Society for Information Science and Technology Conference*, vol. 42, 2006.

[173] D. W. King, "Design and evaluation of information systems," *Annual Review of Information Science and Technology*, vol. 3, pp. 61–103, 1968.

[174] C. C. Kuhlthau, *Seeking Meaning: A Process Approach to Library and Information Services*. Norwood, NJ: Ablex, 1993.

[175] E. Lagergren and P. Over, "Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment," in *Proceedings of the 21st Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, pp. 164–172, Melbourne, Australia, 1998.

[176] B. Larsen, S. Malik, and A. Tombros, "The interactive track at INEX2005," in *INEX 2005*, (N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, eds.), pp. 398–410, Berlin: Springer, 2006.

[177] Y. Li and N. J. Belkin, "A faceted approach to conceptualizing tasks information seeking," *Information Processing and Management*, vol. 44, pp. 1822–1837, 2008.

[178] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, pp. 1–55, 1932.

[179] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger, "What makes a good answer? The role of context in question answering," in *Proceedings of the 9th IFIP TC13 International Conference on Human–Computer Interaction* (*INTERACT 2003*), (M. Rauterberg, M. Menozzi, and J. Wesson, eds.), Zurich, Switzerland, 2003.

[180] J. Lin and M. Smucker, "How do users find things with PubMed? Towards automatic utility evaluation with user simulations," in *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR 08*), pp. 19–26, Singapore, Malaysia, 2008.

[181] S. J. Lin and N. J. Belkin, "Modeling multiple information seeking episodes," in *Proceedings of the Annual Conference of the American Society for Information Science* (*ASIS '00*), pp. 133–146, USA, 2000.

[182] S. Littlejohn, *Theories of Human Communication*. Belmont, CA: Wadsworth, 1992.

[183] I. Lopatovska and H. B. Mokros, "Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts," *Information Processing and Management*, vol. 44, pp. 92–104, 2008.

[184] L. Lorigo, M. Haridasan, H. Brynjarsdottir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan, "Eye tracking and online search: Lessons learned and challenges ahead," *Journal of the American Society for Information Science and Technology*, vol. 59, pp. 1041–1052, 2008.

[185] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay, "The influence of task and gender on search and evaluation behavior using Google," *Information Processing and Management*, vol. 42, pp. 1123–1131, 2006.

[186] R. M. Losee, "Evaluating retrieval performance given database and query characteristics: Analytical determination of performance surfaces," *Journal of the American Society for Information Science*, vol. 47, pp. 95–105, 1996.

[187] H. P. Luhn, "A business intelligence system," in *H. P. Luhn: Pioneer of Information Science. Selected Works*, (C. K. Shultz, ed.), pp. 132–139, NY: Spartan Books, 1958.

[188] A. M. Lund, "Measuring usability with the USE Questionnaire," *Usability and User Experience*, vol. 8, Available online: http://www.stcsig.org/ usability/newsletter/0110_measuring_with_use.html, 2001.

[189] W. E. Mackay, "Ethics, lies and videotape," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 138–145, Denver, CO, 1995.

[190] G. Marchionini, "Exploratory search: From finding to understanding," *Communications of the ACM*, vol. 49, pp. 41–46, 2006a.

[191] G. Marchionini, "Toward human–Computer information retrieval," *Bulletin of the American Society for Information Science*, online: http://www.asis.org/Bulletin/Jun-06/marchionini.html (retrieved November 7, 2008), 2006b.

[192] G. Marchionini and G. Crane, "Evaluating hypermedia and learning: Methods and results from the Perseus Project," *ACM Transactions on Information Systems*, vol. 12, pp. 5–34, 1994.

[193] J. A. Maxwell, *Qualitative Research Design: An Interactive Approach*. CA: Sage Publications, 1996.

[194] D. A. Michel, "What is used during cognitive processing in information retrieval and library searching? Eleven sources of search information," *Journal of the American Society for Information Science*, vol. 45, pp. 498–514, 1994.

[195] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis: A Sourcebook of New Methods*. Newbury Park: Sage, 1984.

[196] L. I. Millett, B. Friedman, and E. Felten, "Cookies and web browser design: Toward realizing informed consent online," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 46–52, Seattle, WA, 2001.

[197] S. Mizzaro, "Relevance: The whole history," *Journal of the American Society for Information Science*, vol. 48, pp. 810–832, 1997.

[198] J. Morahan-Martin, "Males, females, and the Internet," in *Psychology and the Internet: Intrapersonal, Interpersonal, and Transpersonal Implications*, (J. Gackenback, ed.), pp. 169–197, San Diego: Academic Press, 1998.

[199] M. R. Morris and E. Horvitz, "SearchTogether: An interface for collaborative Web search," in *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST '07)*, pp. 3–12, New York, NY, 2007.

[200] J. L. Myers and A. D. Well, *Research Design and Statistical Analysis*. Mahway, NJ: Lawrence Erlbaum Associates Inc., Publishers, Second ed., 2003.

[201] K. A. Neuendorf, *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage Publications, 2002.

[202] J. Nielsen, "Usability 101: Introduction to usability," *Jakob Nielsen's Alertbox*, Retrieved May 05, 2008 from http://www.useit.com/alertbox/20030825.html, 2003.

[203] J. Nielsen and J. Levy, "Measuring usability: Preference vs performance," *Communications of the ACM*, vol. 37, pp. 66–75, 1994.

[204] D. Oard, "Evaluating interactive cross-language information retrieval: Document selection," in *Proceedings of the 1st Cross-Language Evaluation Forum*, Lisbon, Portugal, 2000.

[205] H. O'Brien and E. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, vol. 59, pp. 938–955, 2008.

[206] H. L. O'Brien, E. G. Toms, E. K. Kelloway, and E. Kelley, "Developing and evaluating a reliable measure of user engagement," in *Proceedings of the American Society for Information Science and Technology*, Columbus, Ohio, 2008.

[207] S. L. Payne, *The Art of Asking Questions*. Princeton, NJ: Princeton University Press, 1951.

[208] E. J. Pedhazur and L. P. Schmelkin, *Measurement, Design and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

[209] T. A. Peters, "The history and development of transaction log analysis," *Library Hi Tech*, vol. 11, pp. 41–66, 1993.

[210] T. A. Peters, M. Kurth, P. Flaherty, B. Sandore, and N. K. Kaske, "An introduction to the special section on transaction log analysis," *Library Hi Tech*, vol. 11, pp. 38–40, 1993.

[211] D. Petrelli, "On the role of user-centred evaluation in the advancement of interactive information retrieval," *Information Processing and Management*, vol. 44, pp. 22–38, 2008.

[212] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[213] E. M. Pilke, "Flow experiences in information technology use," *International Journal of Human–Computer Studies*, vol. 61, pp. 347–357, 2004.

[214] D. J. Pittenger, "Internet research: An opportunity to revisit classic ethical problems in behavioral research," *Ethics and Behavior*, vol. 13, pp. 45–60, 2003.

[215] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies," *Journal of Applied Psychology*, vol. 88, pp. 879–903, 2003.

[216] R. J. Riding and I. Cheema, "Cognitive styles — An overview and integration," *Education Psychology*, vol. 11, pp. 193–215, 1991.

[217] S. Y. Rieh, "Judgment of information quality and cognitive authority in the Web," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 145–161, 2002.

[218] S. E. Robertson, "The probability ranking principle in IR," *Journal of Documentation*, vol. 33, pp. 294–304, 1977.

[219] S. E. Robertson, "On GMAP — And other transformations," in *Proceedings of the 15th ACM international Conference on information and Knowledge Management* (*CIKM'06*), pp. 78–83, Arlington, VA, 2006.

[220] S. E. Robertson, "On the history of evaluation in IR," *Journal of Information Science*, vol. 34, pp. 439–456, 2008.

[221] S. E. Robertson and M. M. Hancock-Beaulieu, "On the evaluation of IR systems," *Information Processing and Management*, vol. 28, pp. 457–466, 1992.

[222] K. Rodden and X. Fu, "Exploring how mouse movements relate to eye movements on Web search results pages," in *Proceedings of the Web Information-Seeking and Interaction* (*WISI*) *Workshop at the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '07*), pp. 29–32, Amsterdam, The Netherlands, 2007.

[223] M. E. Rorvig, "Psychometric measurement and information retrieval," *Annual Review of Information Science and Technology*, vol. 23, pp. 157–189, 1988.

[224] I. Ruthven, "Integrating approaches to relevance," in *New Directions in Cognitive Information Retrieval*, (A. Spink and C. Cole, eds.), pp. 61–80, Netherlands: Springer, 2005.

[225] I. Ruthven, "Interactive information retrieval," *Annual Review of Information Science and Technology*, vol. 42, pp. 43–91, 2008.

[226] I. Ruthven, M. Baillie, and D. Elsweiler, "The relative effects of knowledge, interest and confidence in assessing relevance," *Journal of Documentation*, vol. 63, pp. 482–504, 2007.

[227] I. Ruthven, P. Borlund, P. Ingwersen, N. J. Belkin, A. Tombros, and P. Vakkari in *Proceedings of the 1st International Conference on Information Interaction in Context*, Copenhagen, Denmark, 2006.

[228] K. J. Ryan, J. V. Brady, R. E. Cooke, D. I. Height, A. R. Jonsen, P. King, K. Lebacqz, D. W. Louisell, D. Seldin, E. Stellar, and R. H. Turtle, "The Belmont Report," Available at http://www.hhs.gov/ohrp/humansubjects/ guidance/belmont.htm, 1979.

[229] G. Salton, "Evaluation problems in interactive information retrieval," *Information Storage and Retrieval*, vol. 6, pp. 29–44, 1970.

[230] G. Salton, "The state of retrieval system evaluation," *Information Processing and Management*, vol. 28, pp. 441–449, 1992.

[231] T. Saracevic, "Quo vadis test and evaluation," in *Proceedings of the Annual Meeting of the American Documentation Institute, 4*, pp. 100–104, New York, NY, 1967.

[232] T. Saracevic, "Relevance: A review of and a framework for the thinking on the notion in information science," *Journal of the American Society for Information Science*, vol. 26, pp. 321–343, 1975.

[233] T. Saracevic, "Evaluation of evaluation in information retrieval," in *Proceedings of the 18th Annual ACM SIGIR Conference on Research and Development of Information Retrieval*, pp. 138–146, Seattle, WA, 1995.

[234] T. Saracevic, "The stratified model of information retrieval interaction: Extension and applications," *Proceedings of the American Society for Information Science Conference*, vol. 34, pp. 313–327, 1997.

[235] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 1915–1933, 2007a.

[236] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 2126–2144, 2007b.

[237] T. Saracevic and P. B. Kantor, "A study of information seeking and retrieving. II. Users, questions and effectiveness," *Journal of the American Society for Information Science*, vol. 39, pp. 177–196, 1988a.

[238] T. Saracevic and P. B. Kantor, "A study of information seeking and retrieving. III. Searchers, searches and overlap," *Journal of the American Society for Information Science*, vol. 39, pp. 197–216, 1988b.

[239] T. Saracevic, P. B. Kantor, A. Y. Chamis, and D. Trivison, "A study of information seeking and retrieving: Part 1, background and methodology,"

Journal of the American Society for Information Science, vol. 39, pp. 161–176, 1988.

[240] P. A. Savage-Knepshield and N. J. Belkin, "Interaction in information retrieval: Trends over time," *Journal of the American Society for Information Science*, vol. 50, pp. 1067–1082, 1999.

[241] R. Savolainen, "Everyday life information seeking: Approaching information seeking in the context of way of life," *Library and Information Science Research*, vol. 17, pp. 259–294, 1995.

[242] R. Siatri, "The evolution of user studies," *Libri*, vol. 49, pp. 132–141, 1999.

[243] C. L. Smith and P. B. Kantor, "User adaptation: Good results from poor systems," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '08*), pp. 147–154, Singapore, 2008.

[244] P. Solomon, "Discovering information in context," *Annual Review of Information Science and Technology*, vol. 36, pp. 229–264, 2002.

[245] E. Sormunen, "Liberal relevance criteria of TREC: Counting on negligible documents?," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '02*), pp. 324–330, Tampere, Finland, 2002.

[246] K. Spärck-Jones, *Information Retrieval Experiment*. London, UK: Butterworths and Co. Ltd, 1981.

[247] A. Spink, "Multiple search session model of end-user behavior: An exploratory study," *Journal of the American Society for Information Science*, vol. 47, pp. 603–609, 1996.

[248] A. Spink, "Study of interactive feedback during mediated information retrieval," *Journal of the American Society for Information Science*, vol. 48, pp. 382–394, 1997.

[249] A. Spink and H. Greisdorf, "Regions and levels: Measuring and mapping users' relevance judgments," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 161–173, 2001.

[250] A. Spink and R. M. Losee, "Feedback in information retrieval," *Annual Review of Information Science and Technology*, vol. 31, pp. 33–78, 1996.

[251] A. Spink, H. C. Ozmutlu, and S. Ozmutlu, "Multitasking information seeking and searching processes," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 639–652, 2002.

[252] S. R. Stern, "Encountering distressing information in online research: A consideration of legal and ethical responsibilities," *New Media and Society*, vol. 5, pp. 249–266, 2003.

[253] A. Strauss and J. Corbin, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. CA: Sage Publications, 1990.

[254] L. T. Su, "Evaluation measures for interactive information retrieval," *Information Processing and Management*, vol. 28, pp. 503–516, 1992.

[255] W. Sugar, "User-centered perspectives of information retrieval research and analysis methods," *Annual Review of Information Science and Technology*, vol. 30, pp. 77–109, 1995.

[256] SUMI Questionnaire. (retrieved November 05, 2008), http://www.ucc.ie/hfrg/questionnaires/sumi/index.html.

[257] Y. Sun and P. Kantor, "Cross-evaluation: A new model for information system evaluation," *Journal of American Society for Information Science and Technology*, vol. 56, pp. 614–628, 2006.

[258] J. Tague, "Informativeness as an ordinal utility function for information retrieval," *SIGIR Forum*, vol. 21, pp. 10–17, 1987.

[259] J. Tague and R. Schultz, "Evaluation of the user interface in an information retrieval system: A model," *Information Processing and Management*, vol. 25, pp. 377–389, 1988.

[260] J. M. Tague, "The pragmatics of information retrieval experimentation," in *Information Retrieval Experiment*, (K. S. Jones, ed.), pp. 59–104, London, UK: Butterworths and Co. Ltd, 1981.

[261] J. M. Tague-Sutcliffe, "The pragmatics of information retrieval experimentation, revisted," *Information Processing and Management*, vol. 28, pp. 467–490, 1992a.

[262] J. M. Tague-Sutcliffe, "Measuring the informativeness of a retrieval process," in *Proceedings of the 15th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '92*), pp. 23–36, Copenhagen, Denmark, 1992b.

[263] J. M. Tague-Sutcliffe, *Measuring Information: An Information Services Perspective*. San Diego, California: Academic Press, 1995.

[264] J. M. Tague-Sutcliffe, "Some perspectives on the evaluation of information retrieval systems," *Journal of the American Society for Information Science*, vol. 47, pp. 1–3, 1996.

[265] R. Tang, M. Shaw, and J. L. Vevea, "Towards the identification of the optimal number of relevance categories," *Journal of the American Society for Information Science*, vol. 50, pp. 254–264, 1999.

[266] A. R. Taylor, C. Cool, N. J. Belkin, and W. J. Amadio, "Relationships between categories of relevance criteria and stage in task completion," *Information Processing and Management*, vol. 43, pp. 1071–1084, 2007.

[267] R. S. Taylor, "Question negotiation and information seeking in libraries," *College and Research Libraries*, vol. 29, pp. 178–194, 1968.

[268] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger, "The perfect search engine is not enough: A study of orienteering behavior in directed search," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (*SIGCHI '04*), pp. 415–422, Vienna, Austria, 2004.

[269] J. Teevan, S. T. Dumais, and E. Horvitz, "Personalizing search via automated analysis of interests and activities," in *Proceedings of the 28th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '05*), pp. 449–456, Salvador, Brazil, 2005.

[270] P. Thomas and D. Hawking, "Evaluation by comparing result sets in context," in *Proceedings of the 15th Annual Conference on Information and Knowledge Management* (*CIKM '06*), pp. 94–101, Arlington, VA, 2006.

[271] A. Tombros, I. Ruthven, and J. M. Jose, "How users assess Web pages for information seeking," *Journal of the American Society for Information Science and Technology*, vol. 56, pp. 327–344, 2005.

[272] E. G. Toms, L. Freund, and C. Li, "W*i*IRE: The Web interactive information retrieval experimentation system prototype," *Information Processing and Management*, vol. 40, pp. 655–675, 2004.

[273] E. G. Toms, H. L. O'Brien, T. Mackenzie, C. Jordan, L. Freund, S. Toze, E. Dawe, and A. MacNutt, "Task effects on interactive search: The query factor," *Proceedings of INEX 2007*, pp. 359–372, 2007.

[274] R. Tourangeau, L. J. Rips, and K. Rasinski, *The Psychology of Survey Response*. New York, NY: Cambridge University Press, 2000.

[275] A. Turpin and W. Hersh, "Why batch and user evaluations do not give the same results," in *Proceedings of the 24th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '01*), pp. 225–231, New Orleans, LA, 2001.

[276] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proceedings of the 29th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '06*), pp. 11–18, Seattle, WA, 2006.

[277] H. Turtle, W. D. Penniman, and T. Hickey, "Data entry/display devices for interactive information retrieval," *Annual Review of Information Science and Technology*, vol. 16, pp. 55–83, 1981.

[278] D. J. Urquhart, "The distribution and use of scientific and technical information," *The Royal Society Scientific Information Conference*, pp. 408–419, 1948.

[279] P. Vakkari, "Task-based information searching," *Annual Review of Information Science and Technology*, vol. 37, pp. 413–464, 2003.

[280] P. Vakkari, "Changes in search tactics and relevance judgments when preparing a research proposal: A summary of the findings of a longitudinal study," *Information Retrieval*, vol. 4, pp. 295–310, 2004.

[281] P. Vakkari and N. Hakala, "Changes in relevance criteria and problem stages in task performance," *Journal of Documentation*, vol. 56, pp. 540–562, 2000.

[282] P. Vakkari and K. Järvelin, "Explanation in information seeking and retrieval," in *New Directions in Cognitive Information Retrieval*, (A. Spink and C. Cole, eds.), pp. 113–138, Berlin: Springer, The Information Retrieval Series, 2005.

[283] P. Vakkari and E. Sormunen, "The influence of relevance levels on the effectiveness of interactive information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 55, pp. 963–969, 2004.

[284] A. Veerasamy and N. J. Belkin, "Evaluation of a tool for visualization of information retrieval results," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 85–92, Zurich, Switzerland, 1996.

[285] A. Veerasamy and R. Heikes, "Effectiveness of a graphical display of retrieval results," *SIGIR Forum*, vol. 31, pp. 236–245, 1997.

[286] M. Viswanathan, *Measurement Error and Research Design*. Sage Publications, Inc, 2005.

[287] E. M. Voorhees, "On test collections for adaptive information retrieval," *Information Processing and Management*, vol. 44, pp. 1879–1885, 2008.

[288] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press, 2005.

[289] J. B. Walther, "Research ethics in Internet-enabled research: Human subjects issues and methodological myopia," *Ethics and Information Technology*, vol. 4, pp. 205–216, 2002.

[290] P. Wang, "Methodologies and methods for user behavior research," *Annual Review of Information Science and Technology*, vol. 34, pp. 53–99, 1999.

[291] L. Wen, I. Ruthven, and P. Borlund, "The effects on topic familiarity on online search behaviour and use of relevance criteria," in *Proceedings of the 28th European Conference in Information Retrieval* (*ECIR 2006*), London, UK, 2006.

[292] R. W. White, M. Bilenko, and S. Cucerzan, "Studying the use of popular destinations to enhance web search interaction," in *Proceedings of the 30th Annual ACM International Conference on Research and Development in Information Retrieval* (*SIGIR '07*), pp. 159–166, Amsterdam, The Netherlands, 2007.

[293] R. W. White, I. Ruthven, J. M. Jose, and C. J. van Rijsbergen, "Evaluating implicit feedback models using searcher simulations," *ACM Transactions on Information Systems*, vol. 23, pp. 325–261, 2005.

[294] B. M. Wildemuth, *Applications of Social Research Methods to Questions in Information and Library Science*. Libraries Unlimited, (in press).

[295] B. M. Wildemuth, M. Yang, A. Hughes, R. Gruss, G. Geisler, and G. Marchionini, "Access via features versus access via transcripts: User performance and satisfaction," *TREC VID 2003 Notebook Paper*, 2003.

[296] F. Williams, *Reasoning with Statistics: How to Read Quantitative Research*. Orlando, Florida: Holt, Rinehart and Winston, Inc, Fourth ed., 1992.

[297] T. D. Wilson, "On user studies and information needs," *Journal of Documentation*, vol. 37, pp. 3–15, 1981.

[298] L. Xiong and E. Agichtein, "Towards privacy-preserving query log publishing," in *Proceedings of the Workshop on Query Log Analysis: Social and Technology Challenges at the 16th International World Wide Web Conference*, Banff, Canada, 2007.

[299] W. Yuan and C. T. Meadow, "A study of the use of variables in information retrieval user studies," *Journal of the American Society for Information Science*, vol. 50, pp. 140–150, 1999.

[300] P. Zhang, L. Plettenberg, J. L. Klavans, D. W. Oard, and D. Soergel, "Task-based interaction with an integrated multilingual, multimedia information system: A formative evaluation," in *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries* (*JCDL '07*), pp. 117–126, Vancouver, BC, 2007.

[301] X. Zhang, "Collaborative relevance judgments: A group consensus method for evaluating user search performance," *Journal of the American Society for Information Science and Technology*, vol. 53, pp. 220–231, 2002.

ge Models for Information
n Human Language Technol

00/S00158ED1V01Y200811HLT001)

Champaign

dramatically, search engines such as Google are playing a more and more itical to all search engines is the problem of designing an effective retrieval ents accurately for a given query. This has been a central research problem in eral decades. In the past ten years, a new generation of retrieval models, often guage models, has been successfully applied to solve many different trieval problems. Compared with the traditional models such as the vector space model, dels have a more sound statistical foundation and can leverage statistical estimation to val parameters. They can also be more easily adapted to model non-traditional and complex ems. Empirically, they tend to achieve comparable or better performance than a traditional s effort on parameter tuning. This book systematically reviews the large body of literature on stical language models to information retrieval with an emphasis on the underlying principles, ective language models, and language models developed for non-traditional retrieval tasks. t literature has been synthesized to make it easy for a reader to digest the research eved so far and see the frontier of research in this area. The book also offers practitioners an roduction to a set of practically useful language models that can effectively solve a variety of ems. No prior knowledge about information retrieval is required, but some basic knowledge lity and statistics would be useful for fully digesting all the details.

PDF (812 KB) PDF Plus (813 KB)

awi, Mohamad Medhat Gaber, Mihaela Cocea. (2014) Text stream mining for Massive Open
es: review and perspectives. *Systems Science & Control Engineering* 2, 664-676.
tion date: 1-Dec-2014.

eke. (2013) Information Retrieval Models: Foundations and Relationships. *Synthesis Lectures on*
epts, Retrieval, and Services 5:3, 1-163.
tion date: 26-Jul-2013.
F (6811 KB) | PDF Plus (2322 KB) | Supplementary Material
homas, Bram Adams, Ahmed E. Hassan, Dorothea Blostein. (2012) Studying software
g topic models. *Science of Computer Programming*.
tion date: 1-Sep-2012.

Maosong Sun. (2011) Can prior knowledge help graph-based methods for keyword
ontiers of Electrical and Electronic Engineering in China.
tion date: 28-Nov-2011.

Quick

- Alert me when
  New articles ci
- Download to c
- Related article
  Morgan & Clay
- View Most Dow
  Articles

Quick S

Morgan & Cla

Author:

ChengX

Keyword

Informati
search e
retrieval
language
smoothin
topic mo

SEARCH

College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

INST 734
Information Retrieval Systems
Fall 2014 (Online)
Schedule

For each module, this syllabus shows the start date, the end date, and a link to the module. Click on the module title in the **Module** column to get to the module. **Module Previews** are available for modules that are not yet completely ready. Modules will normally be ready several weeks before their start date.

Most modules span 7 days. To accommodate university holidays, Module 1 spans 6 days; Module 13 spans 14 days. In every case, assignments are due at midnight on evening of the final day of the module. This is a sharp deadline; ELMS will not allow assignments to be submitted after midnight. You are welcome to start modules early, but you can not finish them late!

In Modules 2-13, five students will be asked to create one-page summaries of specific additional readings that are due at midnight on Thursday night (for Module 13, this is midnight of the second Thursday night). Links to these additional readings can be found in the **Assigned Summaries** column, and the assignment of additional readings to specific students for modules 2-5 will be available on ELMS by September 9.

Module Number	Start Date	End Date	Module	Module Preview	Assignment	Assigned Summaries	Intro From	Guest Cameo
1	Sep 2	Sep 7	Structure of IR systems		E1		Tenerife	Hussein Suleman
2	Sep 8	Sep 14	Evidence from content		E2	2	SIGIR	Mark Sanderson

3	Sep 15	Sep 21	Ranked retrieval		E3	3	Melbourne	David Lewis
4	Sep 22	Sep 28	Interaction		P4	4	Hokkaido	Kalervo Järvelin
5	Sep 29	Oct 5	Evaluation		E5	5	Kyushu	Jamie Callan
6	Oct 6	Oct 12	Web search		P6	6	SAA	
7	Oct 13	Oct 19	Evidence from behavior			7	CLIP Lab	Jaime Teevan
8	Oct 20	Oct 26	Evidence from metadata		P8	8	Smithsonian	
9	Oct 27	Nov 2	Filtering and recommendation			9	Library of Congress	Julio Gonzalo
10	Nov 3	Nov 9	Scanned documents			10	National Archives	
11	Nov 10	Nov 16	Cross-language search		P11	11	ASIS&T	
12	Nov 17	Nov 23	Speech and music			12	Culpeper	
13	Nov 24	Dec 7	Images and video		P13	13	TREC	
14	Dec 8	Dec 14	Future of IR		P14		NTCIR	
15	Dec 15	Dec 20	Final Exam					

Assignment Key:
E = Exercise
P = Project component

Doug Oard
Last modified: Tue Dec 16 06:26:53 2014

cs446, search engines James Allan Spring 2015

CMPSCI 446 is an undergraduate-level course in search engines and in Information Retrieval, the the science and engineering of indexing, organizing, searching, and making sense of unstructured or mostly unstructured information, particularly text. The class provides an overview of the important issues in information retrieval, and how those issues affect the design and implementation of search engines. The course emphasizes the technology used in Web search engines, and the information retrieval theories and concepts that underlie all search applications. Mathematical experience (as provided by CMPSCI 240) is required. You should also be able to program in Java (or some other closely related language).

For Spring 2015, cs446 is using the University Moodle system. Most of the information on this page is repeated in the class' Moodle site (which will not be available until the start of the semester). Your UMass userid and password will be required for access. Guest access is likely if needed; the password will be announced in class if that happens.

Registering for Spring 2015?

If enrollment in the class reaches the capacity of the classroom but you are interested in taking the class, you may request an override using the on-line overrides form. You will receive an acknowledgement message that indicates when you will be notified of a decision. These decisions are incredibly complex and involve lots of moving targets, so it may take awhile before one is made.

In addition to filling out that form (if enrollment is stopped), you are also welcome to come to the first class to get a better sense of the class. Once students drop the class (and it is likely that a few will), students who have requested an override will be admitted until the class fills again.

Topics

The following topics will be covered, though the order will be determined in part by student needs and interests:

- Overview of search engines and information retrieval
- Architecture of a search engine
- Acquiring data
- Processing text
- Ranking with indexes
- Queries and interfaces
- Retrieval models
- Evaluating search engines

- Classification and clustering
- Social search

Meeting times

The course will meet for two lectures a week: Tuesday and Thursday mornings, 10:00-11:15am, location TBD (tentatively in the Computer Science Building, room 140. That building is located smack in the middle of B2 on this map.)

Textbook

The following text is required for this course.

- B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Addison Wesley, February 2009.

Assignments and exams

Your grade in this class will be based upon homeworks, projects, a midterm exam, and a final.

Collaboration and help

You may discuss the ideas behind assignments with others. You may ask for help understanding class and search engine concepts. You may study with friends. However...

The work that you submit must be your own. It may not be copied from the web, from another student in the class, or from anyone else. If you stumble upon and use a solution from the textbook or from class, you are expected to acknowledge the source of the work (for example, "// The following way of solving this problem is on page 215 of the textbook").

Your effort on exams (midterms or final) must be your own.

Your homework submissions must be your own work and not in collaboration with anyone.

Your project work must be your own work and not a copy of someone else's work, nor done in collaboration with anyone.

UMass Amherst

UMass Amherst

Web Login

moodle.umass.edu

Once you log in, you have access to this and other services that use **Web Login**.

About your NetID & Password | Forgot your password?