Foundations and Trends® in Information Retrieval

Information Retrieval: The Early Years

Suggested Citation: Donna Harman (2019), "Information Retrieval: The Early Years", Foundations and Trends® in Information Retrieval: Vol. 13, No. 5, pp 425–577. DOI: 10.1561/1500000065.

Donna Harman
National Institute of Standards and Technology, USA
donna.harman@nist.gov

now
the essence of knowledge
Boston — Delft

Contents

Information Retrieval: The Early Years

Donna Harman

National Institute of Standards and Technology, USA; donna.harman@nist.gov

ABSTRACT

Information retrieval, the science behind search engines, had its birth in the late 1950s. Its forbearers came from library science, mathematics and linguistics, with later input from computer science. The early work dealt with finding better ways to index text, and then using new algorithms to search these (mostly) automatically built indexes. Like all computer applications, however, the theory and ideas were limited by lack of computer power, and additionally by lack of machine-readable text. But each decade saw progress, and by the 1990s, it had flowered. This monograph tells the story of the early history of information retrieval (up until 2000) in a manner that presents the technical context, the research and the early commercialization efforts.

Donna Harman (2019), "Information Retrieval: The Early Years", Foundations and Trends® in Information Retrieval: Vol. 13, No. 5, pp 425–577. DOI: 10.1561/1500000065.

1

Introduction

This monograph traces the evolution of information retrieval, telling a story starting before 1960 and ending in 2000. This evolution was pushed by an ever-growing demand for better ways of finding information, a push that led to new (mostly) government services and to accelerated research. Part of the story is how the early operational systems developed and were able to leverage the minimal technology to produce usable services. Another part is how technology improved, including computer speed, memory, and networking. But the largest part is how the research grew from a small nucleus of ideas drawn from library science, mathematics, and linguistics to the powerhouse that exists today.

I was very privileged to have worked in the SMART lab at Cornell with Prof. Salton in the mid to late 1960s. This included watching Michael Keen do the first in-depth analysis of experimental results on the IRE, ADI, and Cranfield collections. It also included helping to build and analyze the initial MEDLARS test collection, specifically built to compare ranking and Boolean retrieval. They were exciting times–we thought that the new ranking algorithms would shortly revolutionize the retrieval scene.

Alas, when I next got involved in information retrieval in the mid 1980s, Boolean retrieval dominated the commercial world and the core information retrieval research community was very marginalized. This was thankfully about to change as technology improved.

In late 1989 I was asked to build the new large TIPSTER test collection and I initiated the TREC conferences in 1992. The early 1990s brought the excitement of watching the flowering of the ranking techniques in many diverse retrieval models. The arrival of the Internet and the search engines, all using some type of ranking, was indeed a triumph for the retrieval community.

It is clearly impossible to tell the full history of forty plus years in a limited monograph. Instead I have concentrated on the story about how the field got going, starting with the critical issues in indexing and then the early experiments in searching. The story then follows these early themes as they develop. This is set in the context of the available computer technology and the separate but parallel story of the early operational systems, such as MEDLARS.

The emphasis is on how the ideas built on one another, with papers selected that reflect the main experimental themes. My own biases influence the story in that most papers have an experimental flavor as opposed to ones more theoretical in nature.

Chapter 2 deals with the early "automatic" indexing experiments, which laid the groundwork for the field. Chapter 3 (the 1960s) continues with further work on indexing, such as the Cranfield experiments, and then looks at the early experiments with automatic searching. Chapter 4 (the 1970s) introduces the work in probabilistic models, including the many experiments in term weighting, and discusses the expansion of the operational systems. By the 1980s (Chapter 5) the (mostly commercial) operational systems dominate the scene, but research continued, extending the work of the 1970s. Chapter 6 (the 1990s) illustrates the explosion of the research, both in the diversity of topics and the diversity of researchers (hence only a sample of the research is discussed, mostly following the themes from earlier chapters).

2

In the Beginning (Pre-1960)

2.1 There Have Always Been Libraries

There have always been libraries to store the precious books, manuscripts, scrolls and other information-rich artifacts. In addition to preserving these objects, however, the libraries also needed to know what they had (catalogues) and how to find things in their collections. It is said (Eliot and Rose, 2009) that Callimachus, a Greek poet in 3 BC, was the first to build a catalogue. Francis Bacon in 1605 proposed the organization of all knowledge into categories in his Advancement of Learning. The top categories were Memory, Reason, and Imagination, with each of these then broken down into a second level of categories. For example, Memory had four subcategories: natural, civil, ecclesiastical and literary.

Bacon's categories were picked up and modified by Thomas Jefferson for the 6,700 books he had accumulated by 1815. He created three top-level categories of History, Philosophy, and the Fine Arts. The Fine Arts category covered literary works, but also gardening, painting, architecture and music. Equally importantly, he added 42 "chapters" under these headings that were much more specific, such as the one for chemistry (Gilreath and Wilson, 1989). When he sold his collection

to the Library of Congress in 1815 (for $23,900), his organization was then used by the Library as the "shelf" order of the books so that they could be more easily found (most small libraries at that time shelved books by height or date of acquisition).

In 1876 Melvil Dewey published and copyrighted the first edition of a new method of organizing books, the Dewey Decimal System. He had developed this new classification system while working in the library at Amherst College, using that library as a case study. As Bacon and Jefferson had done before, he divided literature into disciplines, such as philosophy, social sciences, history, religion, pure science, etc. Each category was given an Arabic numerals (000, 100, 200, etc.), with subclasses within these main classes, which were then further broken down into decimal places. His first edition had 44 pages and 2000 entries; the second edition, published in 1885 had 314 pages and 10,000 entries. An abridged edition was published in 1894 for small libraries. By 1927, this system was in use by 96% of U.S. libraries responding to a survey. An international version of this system (the Universal Decimal Classification) was started in 1895 by the Belgian Paul Otlet, originally as a translation of the Dewey system into French, but eventually expanded into other languages, and extended over the years into an international standard that is used by around 150,000 libraries in 130 countries, both for organization and for indexing.

Meanwhile back at the Library of Congress (LOC), the new head Herbert Putnam (1899) brought his own system influenced both by the Dewey Decimal and the Cutter Expansive Classification. This system started with the 26 English alphabet letters, and then switched to numbers and he felt it would be more appropriate for the huge (and fast growing) LOC collection. In addition to the new classification for shelf order, he introduced the use of subject headings, starting from the dictionary of subject headings published by the American Library Association. So as books arrived for cataloguing, they were assigned an LOC classification code (shelf order), but also multiple subject headings taken from the controlled vocabulary of the LOC subject headings (which of course continued to expand). These subject headings provided a way of indexing the books that allowed more flexibility than just the classification codes.

The final library innovation at this time was the card catalog. When the French in 1791 confiscated the library holdings of all religious houses, the inventory was taken on the blank backs of playing cards! But the idea of having a card for each book, rather than a fixed printed catalog, was picked up in many places. For example in 1862 there was a card catalog made for public use at Harvard with hand written cards for 35,762 books. This idea was eventually standardized by the American Library Association; in 1901 the LOC started selling copies of its cards leading to a widespread use of this type of catalog. Each card had the book title, author name, the classification code (Dewey or other), various metadata about publisher, pages, etc., and subject headings. There could be multiple cards per book–title cards, author cards, and multiple cards for different subject headings. Figure 2.1 shows a subject heading card.

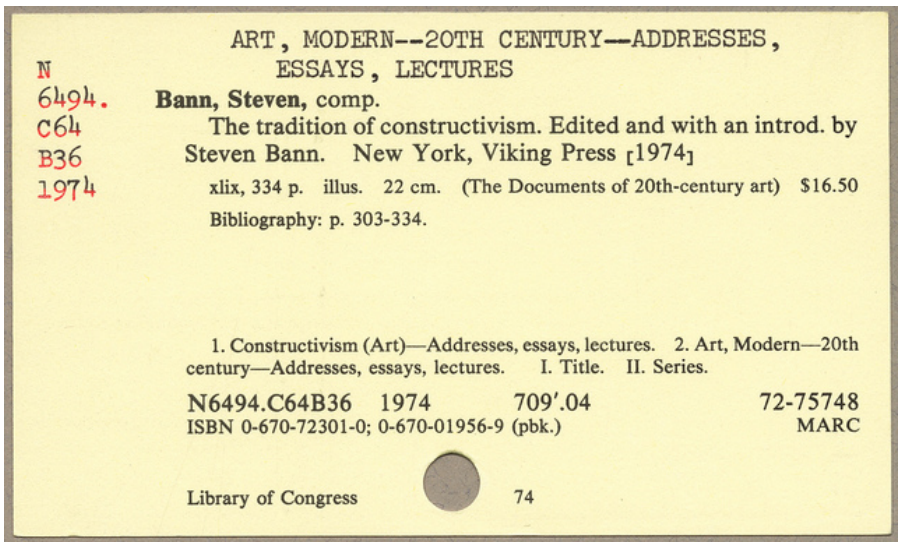
ART, MODERN--20TH CENTURY--ADDRESSES,
ESSAYS, LECTURES
N
6494.
C64
B36
1974
Bann, Steven, comp.
The tradition of constructivism. Edited and with an introd. by Steven Bann. New York, Viking Press [1974]
xlix, 334 p. illus. 22 cm. (The Documents of 20th-century art) $16.50
Bibliography: p. 303-334.
1. Constructivism (Art)—Addresses, essays, lectures. 2. Art, Modern—20th century—Addresses, essays, lectures. I. Title. II. Series.
N6494.C64B36 1974 709'.04 72-75748
ISBN 0-670-72301-0; 0-670-01956-9 (pbk.) MARC
Library of Congress 74

Figure 2.1: A typical card from a card catalog.

However library cataloging deals with unique entities only; magazines, periodicals, journals, etc. are catalogued only as a set. As an example, the Journal of the American Society of Information Science and Technology (formerly American Documentation) is given the Library of Congress classification of Z1007.A477 and subject headings of

Information Science; Periodicals and Documentation; Periodicals. It is left up to the individual journals to produce an index of their contents, which was usually done on an annual basis. As the number of technical journals grew in the late 1800s, it became difficult for researchers and practitioners to keep informed of important new information. This led to the development of independent indexing and abstracting services, mostly connected with professional societies.

One of the first of these in the U.S. was the Engineering Index, a series of annual "Index Notes" initiated by Dr. John Butler Johnson, a professor of Civil Engineering at Washington University in St. Louis, Missouri. Originally published in the Journal of the Association of Engineering Societies in 1884, these short descriptions of the articles collected from across multiple relevant journals became the Engineering Index in 1896. The 1920 edition of the index had 586 pages, with nearly 14,000 items referring to articles in over 700 engineering and allied technical publications.

Another such independent indexing publication was started for medicine by John Shaw Billings, head of the Library of the Surgeon General's Office, United States Army. It was meant to be a monthly supplement for the Index-Catalogue published by the Surgeon General's Office, since it took 15 years to publish the first Index-Catalogue in 1895 and another 20 to publish the second! Index Medicus covered new articles from selected journals, books and theses and was privately published until 1916 when it was merged with the quarterly index published by the American Medical Association and then eventually published by the National Library of Medicine.

As a final example of these independent indexing/abstracting projects, the first issue of Chemical Abstracts (CA) was published in 1907. The editor, William A. Noyes, Sr., used volunteers as abstractors, a tradition that continued until 1994.

Fast forward through a world-wide depression and two World Wars to 1945. In May of 1945 the ENIAC computer was completed at the Moore School, University of Pennsylvania. It was not only 1000 times faster than the Harvard Mark I (operational only one year earlier), but was based entirely on electronic components (vacuum tubes, etc.) as opposed to electromechanical relays.

The year 1945 also saw the publication of Vannevar Bush's "As We May Think" (Bush, 1945), a remarkable insight into how information could be stored and accessed. This was his vision:

> Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

Bush based his concept both on current technical capabilities, such as the use of microfilm to store books, and on his personal desires for a researcher's workstation. So there would be books, periodicals, newspapers, etc., all commercially available, along with copies of various personal manuscripts. Then there would be a "transparent platen" that would be used as input for items such as notes and photographs to be added to the collection. But the most revolutionary part for which this article is cited (8185 times as of early 2018) is his concept of "trails" as a type of associative indexing. A user would build a trail by first finding one item of interest, and then permanently link it to a second item, giving this trail a code name. This code name would be stored with each item and could be extended to as many items as desired. There would be multiple trails of course, including user comments, creating a mesh of interlinked concepts. Bush even envisioned that trails could be shared, creating sources of encyclopedic knowledge for multiple users.

For an interesting discussion of the Memex and how its predictions were realized, see Lesk (1995).

2.2 But More was Needed; Early Mechanical Devices

The exploding amount of scientific literature after 1945 created major problems for researchers. Elizabeth Smith (Smith, 1993) summarized it well.

> These problems were the direct outcome of tremendous increases in research during and after World War II, which

> generated massive volumes of scientific literature for which the existing indexing and retrieval methods were inadequate. New disciplines, new technologies, and new terminology evolved that did not fit into the existing hierarchical subject-heading classification schemes. The existing periodical indexes, which were either alphabetical or classified, proved inadequate to accommodate the broad and complex newly emerging knowledge. It became evident that new methods of indexing and retrieving information had to be developed, including machines that would store and search for information.

Governments and societies stepped in to help. The National Science Foundation (NSF) was established in the U.S. in 1950 to coordinate the exchange of scientific and technical information. In 1948 the Royal Society of Great Britain held the Scientific Information Conference in London, with four sections, each having a specific problem to address (publication and distribution of papers, abstracting services, indexing and other library services, and reviews). Section III on indexing was led by Dr. J. E. Holmstrom (Holmstrom, 1948), who addressed not only the issues of traditional library classification schemes (in particular the Universal Decimal System) but also talked about current and future mechanical devices for storing (and finding) information, including the upcoming UNIVAC machine. He centered his discussion on three types of machines: punched cards with slots, punched (Hollerith) cards for mechanical sorting machines, and microfilm machines.

The punched cards with slots, also called edge-notched cards, had been around since early in the 1900s under names such as Copeland-Chatterson/Paramount or McBee Keysort cards. These cards were paperboard, often 5 by 8 inches, with a blank space in the middle (for writing), and holes around all four edges. To create a record, the appropriate holes were punched so that they became open notches and the described item was written in the blank center of the card. Kilgour (Kilgour, 1997) gives an illustrated example of these types of cards being used for sorting tree specimens, where each card represented a specific tree and the 76 holes each represented a different feature.

The tree features were carefully selected as important characteristics, such as geographical region, growth rings, etc. A steel or wooden needle could then be passed through the card deck at one feature hole and the notched edged cards for that feature would drop out. This subset of dropped cards could be then needled for a second feature (a Boolean "and"). Note that this type of card system could easily be set up by individual scientists on a much smaller scale to help with their research, where features could be authors, chemical names, journal names, etc.

A more sophisticated version of the edge-notched cards was developed by Calvin Mooers when he founded his Zator card company specifically for document searching in 1947. Calvin Mooers is better known as the first person who is said to have used the term "information retrieval" in a publication. It may have been used in his Master's Thesis from M.I.T., but in 1950 he wrote (Mooers, 1950):

> The problem of directing a user to stored information, some of which may be unknown to him, is the problem of "information retrieval"... In information retrieval, the addressee or receiver rather than the sender is the active party.

The Zator cards had holes only on the top and were meant to be used in a special machine. Each card represented a specific document and the 40 holes at the top represented the "descriptors" which were encoded to allow very complex patterns (Mooers, 1951). Mooers marketed his systems to various businesses/professional societies, customizing the descriptors for each customer.

A second type of simple hole punching was the Batten card, also known as the peek-a-boo system. This type of system had been used as early as 1915 for identification of birds (Kilgour, 1997), but Batten's cards had a 20 by 20 matrix containing 400 squares. Unlike the edge-notched systems, this system had one card for each subject heading and the holes punched in the 400 squares were based on the sequentially-numbered documents that matched that subject heading (the inverse of the edge-notched systems). To find all the documents that corresponded to two different subject headings (again a Boolean "and") the two cards were held in front of a light source and the light "peeked" through the

holes for the documents containing both subject headings. This idea was eventually expanded to 800-hole IBM cards.

The idea of using punched cards in mechanical sorting machines was also not new. The U.S. 1890 census had been run on a card sorter/tabulator built by Herman Hollerith using punched cards with round holes, 12 rows and 24 columns. Counters could be used on a single hole, or on a combination of holes. Eventually his company was joined with others to form the IBM company and in 1928 the card format was changed to 12 rows by 80 columns, almost doubling the amount of data that could be stored.

When this was extended to information retrieval, each 80 column card represented an item/article and some of the columns of the card carried the item's serial number. The rest of the card carried the indexing information, usually based on subject headings turned into some type of code. Unlike the 1890 census data, or most business data, the information to be encoded was complex and various schemes had to be designed in order to translate the information into simple codes that could be used for sorting. For most of the experimental systems there were fixed columns used for specific categories.

An example would be the experiment starting in 1948 at the Welch Medical Library at John Hopkins University where the first 7 columns contained the serial number, column 8 indicated how many cards were coded for that serial number, and columns 24–32 was allocated for the category chemistry, with columns 28–31 for the code for a specific chemical compound. These cards would be passed through the sorter, and based the columns chosen for sorting, cards would be selected, and then this selected subset could be run through a second pass to sort on a different set of columns. This was tedious and of course very time-consuming. Bohnert (Bohnert, 1955) reports on seven different experimental systems, mostly using mechanical sorters, but in general the experiments did not result in the new methods being adopted.

A major advocate of these new systems was the chemical industry. Malcolm Dyson, a U.K. chemist and J. W. Perry, a U.S. chemist, asked Thomas Watson Sr., IBM president, for help in 1947, resulting in Watson assigning one of his top engineers, Hans Peter Luhn, to work on improved versions of the sorted card technique. The result was the

Luhn scanner, first demonstrated at the World Chemical Conclave in New York City, September 1951. Instead of a sorting machine, Luhn's scanner ran the cards vertically through a specially designed machine. The punched codes here could be in any of the 80 vertical columns. The same codes were used to input a question to the scanner, which used pattern matching via photo-electric cells. The Luhn scanner was expensive to build and was never a commercial success, however it did get Luhn interested in the problems of information retrieval.

Holmstrom also talked about the microfilm machines, and in particular the Rapid Selector. This machine had a very complicated history, with the first prototype built by Vannevar Bush's team at M.I.T. between 1938 and 1940 for cryptanalysis (Buckland, 1992). Ralph Shaw, head librarian of the U.S. Department of Agriculture, wrote to Bush in 1946 asking to borrow the prototype for information retrieval; this eventually resulted in a contract from the Department of Commerce to Bush's former students at Engineering Research Associates to build an operational Rapid Selector, and a contract to Ralph Shaw to build testing materials to investigate the usefulness of the machine (Varlejs, 1999).

The Rapid Selector (and other related microfilm machines) carried the document abstracts on half of the frame and a series of coded dots on the other half, the codes corresponding to information about the document, such as encoded subject headings, metadata, etc. The Rapid Selector delivered to Shaw in 1949 had six selection entries of seven digits each (Bagg and Stevens, 1961). The "final" version built for the Bureau of Ships by the National Bureau of Standards had 45 code positions across the film horizontally, and as many positions as needed vertically. The film would be run through the machine, looking for the dot pattern that matched the question (similar to the Luhn scanner), and at that point would automatically copy the abstract to a second film for later reading.

The Rapid Selector was never a commercial success. Bourne (1961) lists twelve different advertised products, none of which ever worked for document retrieval. The slow speed, the fact that only one criterion code could be searched per run (similar to the card sorting machines), and the difficulty of coding the document information into appropriate patterns

made the machine too expensive except for specialized applications like the Bureau of Ships. However the fascinating history of the machine, including patents by Emanuel Goldberg in 1931 (Buckland, 1992) and the complex story behind its further development by Bush (Burke, 1992) make for interesting reading, illustrating the difficult path from research to production.

For more on the history of these mechanical devices, see Sanderson and Croft (2012).

J. W. Perry continued his interest in the use of punched cards, writing a heavily-used textbook in 1951. During 1954 and 1955, he (and Allen Kent and M. M. Berry) jointly published a series of ten articles in American Documentation. These articles defined the indexing problem for machines, starting with the problems of complex subject headings being encoded into short patterns, and continuing into how to deal with synonyms and more generic terms, and the limitations of the different mechanical devices in terms of searching. One of the articles (Kent *et al.*, 1955) dealt with evaluation, including definitions for "pertinency" (precision) and "recall", including a comment that both were important.

Perry and Kent founded the Center for Documentation and Communication Research at Case Western University in 1955, the first academic program for the use of mechanical devices for information retrieval. They embarked on a 5-year project for the American Society of Metals (ASM), designing and constructing the WRU Searching Selector based on ideas in their earlier papers. The Selector could search 25,000 current papers (abstracts) from metallurgy using a punched paper tape database and five concurrent Boolean searches. In 1960, after running pilot searches based on questions from their members (Rees and Kent, 1958), ASM initiated the first fee-based bibliographic search service.

2.3 Indexing Wars: Round 1

Computers were fast becoming more of a usable reality.

- In 1949 the EDSAC at Cambridge University began operation as the first large scale, stored program electronic digital computer.

- In 1951 the first UNIVAC computer was delivered to the U.S. Bureau of Census by Mauchley and Eckert and became commercially available in 1954.

- The first IBM 701 came in 1953, with programs stored in internal and addressable electronic memory. It came with two tape drives, a magnetic drum memory, along with a card reader, a printer, and an operator's station.

The opportunities to expand the earlier mechanical devices to much faster and flexible machines seemed endless. However the complex encoding of subject headings, classifications, etc. still provided a big bottleneck to progress. The painstaking work done by Kent and Perry (resulting in an indexing vocabulary of over 30,000 terms organized into a classification scheme (Kent *et al.*, 1954)), or by Calvin Mooers, could not scale beyond handling small domains. There was no shortage of proposals for new (and equally complex) indexing schemes, but implementing any one of these would have been an enormous effort with no guaranteed payoff.

An editorial (Shera, 1955) in American Documentation addressed the problem.

> May the age-old controversies that arouse from the conventional concepts of classification not be reborn in the mechanized searching systems of the future. There is hope for the avoidance of such error if we will but regard documentation systems as useful devices the benefits of which must be determined, not by polemics but by the intelligent measurement of such benefits in relation to needs and costs. The machines of the future can make us free, but only if we are willing to subject them and ourselves, to the most rigid intellectual discipline.

There had been some testing, although inconclusive. In May of 1951 the Armed Services Technical Information Agency (ASTIA) was established to "provide an integrated program of scientific and technical services to the Department of Defense and its contractors". This required that

the document processing centers of the Navy Research Section (at the Library of Congress) and the Central Air Documents Office (in Dayton) be combined into a single service. Unfortunately the two organizations had subject heading lists based on different philosophies (Gull, 1956) and these would have to be combined in some way.

Mortimer Taube was the chief of the Navy Research Section, trained as a librarian, and had a long interest in better ways of indexing documents. In September of 1951 he and Alberto F. Thompson of the Atomic Energy Commission Technical Information Service had presented the paper "The Coordinate Indexing of Scientific Fields" (see Appendix A in Gull (1987)) before the Symposium on Mechanical Aids to Chemical Documentation. This paper was the first use of term "coordinate indexing", although it was basically what the sequential card sorting or edge-notched systems were doing. However in this talk Taube added the idea that the subject headings did not have to be so complex if coordinate indexing was used.

> In a coordinate index, detail and specificity, as well as complex ideas, are achieved by coordinating two or more elements or terms. This means that the number of terms in any field can be enormously reduced as contrasted with the number of ways in which information can be stored or found.

In the spring of 1952 Taube founded Documentation, Inc. (DI), the "first private organization anywhere devoted to research and development in the field of documentation", starting with a contract from the Air Force and ASTIA to research better methods of indexing and specifically ones that would allow a principled combination of the two different subject headings. Cloyd Dake Gull (who had worked under Taube at the Navy Research Center) also joined the company, along with a mathematician Irma S. Wachtel. These three published a paper in 1952 (Taube *et al.*, 1952) discussing their initial investigations into this new type of indexing. They started by looking at the existing subject headings for a large sample of reports from both organizations and attempting to index these with single terms which could then be coordinated. Looking at 40,000 subject headings, they found that only some 7000 different words were used (Cleverdon, 1991). Issues such as multi-term descriptions

("bone cancer" or "radio network") were debated and rules constructed for their handling. Some examples of their indexing are shown in a later paper (Taube and Associates, 1955). Once these problems had been resolved, they launched into a larger scale indexing of the documents themselves, with the goal of running a full test between a search using the old subject headings and a search using "Uniterms", the proprietary form for the phrase "unit terms".

Gull's 1987 paper gives many of the early difficulties that Taube encountered in his goal to make a commercial success of this method. It should be noted that the "Uniterm System of Coordinate Indexing" was completely foreign to librarians at this point and readers found it hard to understand. Gull's answer to one reviewer's question about the use of single terms was as follows:

> There was no confusion in the understanding of Uniterrns in 1952–1954 by the staff of DI. Uniterms were single words, a subset of the logical sum of all the words found in the language of the index; they were derived from the texts in the collection which was indexed. It mattered not that some words represented concrete objects and others abstract ideas. It did matter how the searcher combined the single words into search questions to achieve satisfactory search results.

Searching for documents using the Uniterm system was done manually in 1952, using a variation of the edge-punched cards. There was one card for each Uniterm, and all of the sequentially numbered documents that contained that term were listed on that card in 10 columns of 40 rows. But there were no notches and no holes; the document numbers had to be manually "eyeballed" using a clever arrangement by Taube (see Kilgour (1997) for details and a sample card).

In 1953 the ASTIA headquarters approved a comparison of the Uniterm system with the traditional subject headings (Gull, 1956). There were 30,000 newly arrived documents selected for the test, with 15,000 to be indexed by Documentation Inc. using their Uniterm system and 15,000 to be indexed by ASTIA at the Library of Congress using their regular subject headings. The indexing was done and there were 93 current requests gathered. Each team then searched their system

for all 93 requests and gathered their set of relevant documents. The two different relevant sets were compared, showing that there were 580 documents in common out of 3200 documents selected across the two groups. Discussion confirmed agreement that most of the opposite group's documents were indeed relevant, but that did not resolve the issue of which system was the best. An earlier suggestion to submit both sets of results to the requestor had not been followed. A look at the failure analysis for each group showed similar errors. For the 492 documents missed by the ASTIA group, the two largest categories were a poor search (indexing was correct), and that the subject heading was too broad. The 318 documents missed by the Documentation Inc. group were mainly searching errors, although there were 97 that were missed because of inconsistent indexing. Although this test was inconclusive (and certainly did not impress the library community), it did show that it was possible to index and search using the Uniterm system.

Taube continued to lecture about his system and in 1953 managed to get the librarians at the Naval Ordnance Test Station (NOTS) in China Lake, California to start using it for indexing reports. By 1954 there were searches being manually conducted using the cards (Segesta and Reid-Green, 2002). NOTS had an IBM 701 and the manager of the computing section, Harley Tillitt, was anxious to find new uses for it. During 1954 his team started to implement the Uniterm system on the 701; it was a major challenge since the IBM 701 did not have a built-in alphameric character representation. This was resolved using "building blocks", i.e. subroutines that performed a series of specific operations to convert the alphameric characters to a six bit binary code (Bracken and Oldfield, 1956). The 701 system basically mimicked the manual search (Bracken and Tillitt, 1957), working from a library tape of 14,000 report numbers and 9,600 descriptors. Tillitt claimed that there were approximately 16 library searches made 3 times a week, making the NOTS installation the first computerized information retrieval system.

2.4 Automatic Indexing/Abstracting: Part 1 (Luhn)

By showing that it was possible to categorize documents via a series of single terms, Taube had demonstrated that complex subject headings were not necessary for indexing. However this basically put the burden on the searcher to pick the best terms for coordination, and this was not easy (note that the majority of errors using Uniterms in the ASTIA tests had been in searching). Luhn's paper "A New Method of Recording and Searching Information" (Luhn, 1953) addressed this problem. His concern was that it was difficult for both the indexer and the searcher to pick the most important concepts and that the use of strict coordination often missed "subordinated aspects". He proposed broadening the concept range in the following manner.

> If we consider a concept as being a field in a multi-dimensional array, we may then visualize a topic as being located in that space which is common to all the concept fields stated. It may further be visualized that related topics are located more or less adjacent to each other depending on the degree of similarity and that this is so because they agree in some of the identifying terms and therefore share some of the concept fields.

The paper showed several Venn diagrams to illustrate the point but did not specify exactly how the searching might be done. Luhn continued these ideas in his 1957 paper "A Statistical Approach to Mechanized Encoding and Searching of Literary Information" (Luhn, 1957). Here he not only addressed the problems of searching but also the initial indexing, where he felt that better use could be made of the increasing power of computers. So instead of subject heading type indexing, or a large number of Uniterms, he proposed the use of a dictionary of "notions". The construction of this dictionary would be "undertaken by experts thoroughly associated with the special field of the subject". The notions would be selected so as to "resolve the material in terms of an optimal number of equally weighted elements". His example suggests that "electricity" is too common (in that field), whereas "butterfly" would be useless because it would so rarely occur. Notions that were

too broad would be broken into subnotions; the ones that were too rare would be built into notional families. This dictionary would be built incrementally, adding new notions, subnotions, and families when necessary.

Once this dictionary existed, Luhn went on to propose that the occurrence of the notions within the documents could be automatically used by the computer to weight the notions.

> A notion occurring at least twice in the same paragraph would be considered a major notion. A notion which also occurs in the immediately preceding or succeeding paragraph would be considered a major notion even though it appears only once in the paragraph under consideration.

This was based on his observation that "the more frequently a notion or a combination of notions occur, the more importance the author attaches to them as reflecting the essence of his overall idea."

Luhn also suggested uses of the document structure in the weighting of the notions, such as occurrences of notions in captions, titles and resumes (abstracts). The search operation would then similarly encode an essay-length information request into a set of weighted notions and use statistical methods (such as requiring a given degree of similarity) to match the requests to the encoded documents. Luhn concludes the paper with "the salient contribution of the system described in this paper lies in its realization of the completely *automatic* encoding of the documents" and that "the ideas of the author would not be narrowed, biased, or distorted through the intervention of an interpreter".

Luhn quickly applied these ideas to automatic abstracting (Luhn, 1958a), where he went beyond the idea of notions to that of using the words themselves. The goal was to extract the most important sentences in a document to serve as its automatic abstract.

> It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnishes a useful measurement for determining the

> significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements.

His ideas for implementing this included using a stoplist or word frequency cutoff to reduce noise, using a simple method of conflating stemmed words, and then looking for sentences in the technical article that had the "the greatest number of frequently occurring different words found in greatest physical proximity to each other" (his suggested limit was no more than four or five intervening non-significant words). The sentences were then ranked in significance order and one or two were used as the extracted abstract.

Although Luhn never built these systems (the computers were not ready for them and indeed there was no machine-readable text), they contain the seeds of many of the big research ideas in the 1960s (vector space model, term frequency weighting, and statistical matching rather than strict Boolean), and even those in the 1970s (term discrimination weighting and IDF (see Section 4.1.4)).

However two of Luhn's innovations were quickly converted into heavily-used systems. Luhn and Herbert Ohlman from the System Development Corporation demonstrated the first machine-generated Key-Word-In-Context (KWIC) indexes at the International Conference on Scientific Information in 1958, using the actual titles, subtitles, figure and table captions, etc. of the papers presented at that important conference (Stevens, 1970). The system worked on a text stream of at most 60 characters of separated words by first removing the non-significant words (taken from a list) and then showing each of the significant words embedded (starting in column 25) within the surrounding text (Luhn, 1960). Figure 2.2 shows a piece of a KWIC index, with the embedded significant words in alphabetical order. Note the centering of the significant words, such as "ABSENT", with the surrounding text on the right and/or left side (the far right columns have metadata about the article). Within several years these KWIC indexes were in use by over 20 companies and organizations, often issued monthly (Stevens, 1970).

The year 1958 also saw the announcement of the SDI (selected dissemination of information) concept and system for the IBM Library.

Portion of the 1964 KWIC Index to the *American Political Science Review*

KEYWORD	Title of Article	Reference Code: Author	Year of Pubn.	Ident. Number
GREECE	ABANDONS PROPORTIONAL REPRESENTATION.=	POLYZO	AT29	961
BOARDS AND COMMISSIONS CREATED AND	ABOLISHED IN 1913.= STATE OFFICERS,	BATES	FG14	292
THE MONROE DOCTRINE	ABROAD IN 1823-24.=	ROBERT	WS12	225
JUDICIAL	ABROGATION OF COUNTY HOME RULE IN OHIO.=	SHOUP	EL36	1343
	ABSENT - VOTING IN NORWAY.=	SABY	RS18	474
MILITARY	ABSENT - VOTING LAWS.=	RAY	PO18	481
	ABSENT - VOTING LAWS, 1917.=	RAY	PO18	468
	ABSENT - VOTING LEGISLATION, 1924-1925.=	RAY	PO26	1910
	ABSENT VOTERS (LEGISLATION).=	RAY	PO14	294
	ABSENT VOTING (LEGISLATION).=	LAPP	JA16	374
	ABSENT VOTING LAWS.=	RAY	PO24	702
	ABSENT VOTING.=	KETTLE	C 17	426
	ABSENTEE VOTING IN THE UNITED STATES.=	STEINB	PG38	1456
D POLITICS.=	ABSOLUTISM AND RELATIVISM IN PHILOSOPHY AN	KELSEN	H 48	1941
RELATIVISM,	ABSOLUTISM, AND DEMOCRACY.=	OPPENH	F 50	2049
Y-- THE NATIONAL INTEREST VS. MORAL	ABSTRACTIONS.= THE MAINSPRINGS OF AMERICAN	MORGEN	HJ50	2039
SOCIAL SCIENCE	ABSTRACTS-- AN INSTITUTION IN THE MAKING.=	CHAPIN	FS30	1035
ON OF CASE STUDIES-- THE PROBLEM OF	ABUNDANCE (PUBLIC ADMINISTRATION).= PREP	STEIN	H 51	2071
PREME COURT DECISIONS-- THE USE AND	ABUSE OF QUANTITATIVE METHODS.= THE MATHEM	FISHER	FM58	2389

Figure 2.2: Keyword in context (KWIC).

This was a piece of a vision by Luhn (1958b) on how an intelligent business system (for researchers) should operate and contained both the auto-abstracting and encoding ideas, along with automatic creation and updating of "action-point" profiles (the SDI systems). For more on the details of his SDI process, see Luhn (1961).

2.5 Automatic Indexing/Abstracting: Part 2 (Maron and Kuhns)

Computers and storage were improving. The IBM 305, a random access disk storage system, was shipped on the IBM 350 RAMAC system for the first time in 1957. There were 50 disks, each with 100 concentric recording tracks reached by a moving arm. At 100,000 characters per disk, or about 5 MB total, these cost about $50,000. In 1959 the Control Data Corporation released the CDC 1604, the first fully transistorized computer (the NOTS system had only about a 20 minute uptime due to vacuum tube failures). The Institute of Information Scientists in the U.K. was founded in 1958, as was the U.S. Department of Defense Advanced Research Projects Agency (ARPA).

The International Conference on Scientific Information took place in the autumn of 1958 in Washington D.C. Although it was a followup to the 1948 London meeting, this time the research community was

involved, including a session on "The Organization of Information for Storage and Retrospective Search: Intellectual problems and equipment considerations in the design of new systems". Taube presented his Comac system, Margaret Masterman, Roger Needham and Karen Spärck Jones presented a paper entitled "The Analogy between Mechanical Translation and Library Retrieval" and Cyril Cleverdon described his just-funded Cranfield 1 project on evaluation. This was the conference for which Luhn had built the KWIC index; the organizers had specifically set up the papers such that the sentence separation was easy (two periods), and distributed the publication monotape two months before the conference so that researchers could have specific items keypunched for machine reading and research (Stevens, 1970).

Another researcher, M. E. Maron at Ramo-Wooldridge Corp (which later became TRW) was concerned about the issue of strict coordinate matching. He felt that "there should be degrees of aboutness" and in August 1958 sent an internal document to his management suggesting research using probabilities to express this aboutness. They approved, he enlisted a mathematician friend Larry Kuhns (Maron, 2008) and together they produced two tech reports (Maron *et al.*, 1959) detailing the theory, ways of implementing that theory, and finally results of an experiment showing reasonable results. A more formal version of the report was published in 1960 (Maron and Kuhns, 1960).

The central theme of the Maron/Kuhns report was this notion of probability: index tags should be expressed as "the probability that if an individual desires information of the type contained in the document then he will use the tag in question in requesting that information". Additionally they wrote that:

> the ideal search system is one that computes the distribution which describes the probability that a document will satisfy a requestor. This means that given a request, a class of documents is selected (namely those whose index terms are logically compatible with the terms and logic of the request) and for each document in this class, the system will have to compute a number, called the "relevance number" which will be a measure of the expected degree of relevance of

> the document for the requestor. The notion of weighting the index terms that are assigned to documents and using these weights to compute relevance numbers is basic to the technique which we call "Probabilistic Indexing".

The use of "relevance number" can be confusing here, but it refers to what could be called the relevance score today and the documents could be ranked by this relevance number.

Given this basic notion, the authors went on to elaborate in several directions. The obvious one was the ability of the requestor to add probability weights to the request, so there were not only weights on the index tags but weights on the request terms. Less obvious was the use the á priori of probability distributions associated with the documents themselves, i.e., the "relative frequency of usage" of the document within this library or collection.

Maron and Kuhns were also concerned with the problem of constructing the requests, given all the known problems with searching.

> As a remedy for this situation, Probabilistic Indexing includes methods for automatically elaborating upon any arbitrary request so as to improve its selectivity. That is to say, included among the methods of Probabilistic Indexing are mechanical rules for automatically relating index terms (and documents) so that given a request for a particular set of index terms a computer can determine what other terms are most closely related to the request and thereby automatically elaborate upon it in the most probable direction, in order to improve the selection.

The methods they suggested will be discussed here using words rather than mathematical notation in order to focus on the core ideas. Readers are referred to the 1960 paper for more specifics on the probabilistic formulas, and additionally to the 1959 report for an extensive mathematical derivation of their ideas. Note however that the mathematical notation systems have changed somewhat since 1959.

The authors based their rules for request elaboration on various types of closeness:

1. closeness of the request to other requests in the request space, i.e., modify the request itself based on this information;

2. closeness of selected documents to other documents in the document space, i.e., modify the selected document set by similarity (dissimilarity) to other documents in the document space;

3. closeness of index terms to other index terms in the index space, possibly semantic or syntactic closeness, but the paper (and report) goes into great and interesting detail on the multiple ways of looking at "closeness" here.

The 1960 paper contained the following overall search strategy (slightly modified to fit with comments added for clarity). This is included here in detail to show that many of the research ideas developed later were part of the thinking of Maron and Kuhns very early on. In addition to the weighting of terms, they introduced a notion similar to the IDF (see Section 4.1.4), the use of the "popularity" of documents in the collection to increase the probability of them being retrieved, and multiple ways of automatic improvement of the initial request (query) that are echoed today in the web world. The overall search strategy with the various control variables would not be out of place in textbooks today!

```
First we list the variables involved:
1. Input
     (a) The request R
     (b) The request weights
2. The Probabilistic Matrix
    (a) Dissimilarity measures between documents
    (b) Significance measures for index terms.  (An
         index term applied to every document in the
         library will have no significance, while an index
         term applied to only one document will be
         highly significant.  Thus significance  measures
         are related to the "extension number" for each
         term; i.e., to the number  of documents tagged
         with the term, the smaller this number,  the
         greater the significance of the index term.)
     (c) "Closeness" measures between index terms
```

```
3. The A Priori Probability Distribution (comment: how
     often this document has been accessed in the past)
4. Output (by means of the basic selection process;
     i.e.,the logical match plus Bayes' schema
     with all of its ramifications and refinements).
     (a) The class of retrieved documents; call this "C".
     (b) n, the number of documents in C.
     (c) Relevance numbers (comment: the "relevancy score").
5. Control Numbers
     (a) the maximum number of documents that
           we wish to retrieve.
     (b) Relevance number control; e.g., we may
          ignore documents with relevance number
          (comment: score) less than a specified value.
     (c) Generalized relevance number control.
     (d) Request weight control; i.e.,we elaborate
          on index terms in  the request if their request
          weight is higher than some specified value.
     (e) Significance number of index term control; i.e.,
           we give index terms of certain significance
           ....special attention.
6.Operations
     (a) Basic selection process
     (b) Elaboration of the request by using "closeness"
          in the request space.
     (c) Adjoining new documents to the class of retrieved
          documents by using "distance" in the document
          space.  trim C to documents having relevance
          number greater than the control number.
     (d) Merge: any merging operation between the
          two classes
```

The 1960 paper ends with an experiment to test these ideas (see the 1959 report for many more details on the experiment). The collection they built used 100 articles from the Science News Letter; brief and not too technical. They built their own index, a set of 47 well-defined categories based on 577 different keywords they identified across the set. Once the categories were defined, the documents were manually indexed using weights at intervals of $8(\frac{1}{8}, \ldots, \frac{8}{8})$ based on indexer choice. Figure 2.3 shows the guidance given to indexers for doing the weighting.

WEIGHT	DESCRIPTION	WHEN USED
8/8	Major Subject	The term is highly specific and covers an entire major subject of the document.
7/8	Major Subject	The term is specific and covers most of a major subject of the document.
6/8	More Generic Subject	The term is too broad and covers a major subject.
5/8	Other Important Terms	Terms that would be used in a binary indexing but not a major subject.
4/8	Less Generic Subject	The term relates to but is too narrow to cover a major subject.
3/8	Minor Subject	Includes such terms as relate to results of experiments, intermediate methods, possible uses, etc.
2/8	Other Subjects	Other relevant tags.
1/8	Barely relevant	Subjects classifier would not want to use but feels that some users might consider them relevant.

Figure 2.3: Maron's guide for assignment of weights.

The requests were built using the source document method, i.e., a set of random articles were selected and test subjects were asked to generate a request for which that document would be considered the answer. The search was then run, the documents ranked, and in 27 out of 40 requests the answer document was retrieved, usually at a high relevancy number. When the test subjects were asked to rank all the documents received and rate them for relevancy (on a 5 point scale), it was found that the relevancy numbers (scores) calculated by the machine had the same ranking as the document ratings, i.e. the very relevant documents had a mean relevancy score of 0.81 whereas the only slightly relevant ones had a score of 0.40.

Three different types of request elaboration were tried, all using modification of the index terms; each method retrieved an additional six answer documents. Failure analysis showed that for the seven answer documents not found, three were initial indexing failures and three were requests that were so poorly formulated that elaboration did not help.

Unlike Luhn, Maron and Kuhns were able to run an experiment and show that their theory worked. Even if the indexing was still done manually, there were weights, and these weights could be used to rank the retrieved document set. Like Luhn, their ideas became the seeds for research in the next decade, particularly in the probabilistic work that mainly started in the 1970s.

3

Full Steam Ahead (1960s)

3.1 Automatic Indexing/Abstracting: Part 3

In March 1959 at the Western Joint Computer Conference, Calvin Mooers (Mooers, 1960) presented a paper looking at the current state of automatic indexing and making a fascinating set of predictions. These predictions reveal of some of the thinking of the times, and also some of the wilder hopes for the use of computers. Mooers suggested that if the Luhn automatic abstracting (using frequency of words) was extended to just use the words themselves to index, then some type of normalization would be necessary to handle issues like equivalence classes of words and phrases. This thinking would be reflected in much of the research in the 1960s where "thesaurii" would be built to deal with this normalization issue. Mooers went on to suggest that one could use "inductive inference machines" to learn assignments of rudimentary subject headings using previously assigned manual subject headings, classifications, etc. (i.e., machine learning). He also discussed the problems users have in building the requests and suggested that machines could help in the learning process. "One could see that the process of input request formulation and the process of giving out information will merge into a sustained communication between the customer and the machine". It took over 30

years to see something like this start to happen and it is still a major area of research both commercially and academically.

Meanwhile the world's publication numbers were rapidly rising, with estimates in 1962 of 15,000 significant journals containing over 1 million papers per year (Bourne, 1962). About 60% of these were in English, with around 10% each in French, German, and Russian. Even with over 3500 abstracting and indexing services (worldwide), Bourne noted that only a few fields such as metallurgy, chemistry and medicine/biology were well covered. Some areas were critical enough that more resources could be employed. For example, in 1960 the National Library of Medicine introduced Medical Subject Headings (MeSH), a controlled vocabulary for indexing journal articles and books in the life sciences. But for most areas, there still needed to be some type of automatic indexing as there were not enough funds to manually assign index terms.

Maron (by then at Rand Corporation) wanted to find some way of automatically assigning probability weights to terms (Maron, 1961). He saw this as a two step process: first locate what he called "clue" words in a given document, and then figure out how to assign the probability weights to these words. His application was not information retrieval but the ability to automatically classify the documents into categories based on their terms, and he did this by building training and test collections, similar to what is done today. His corpus was 405 abstracts from the IRE Transactions on Electronic Computers, with 260 from March and June used for learning the weights (group 1), and 145 from September used for testing (group 2). These abstracts came with classifications into 10 major categories (with some subcategories), however Maron decided to work with a finer set of 32 categories, which he had manually assigned to the 405 abstracts (only 20% were assigned to more than one category).

Once this was done, the abstracts were keypunched. For group 1 (training) there were over 20,000 word occurrences (average 79 per abstract) of 3263 unique words. These needed to be winnowed down to the clue words. First the 55 most frequently occurring "function words" were removed (8,402 of the total 20,515 occurrences). Then a set of the most frequently occurring domain-specific words such as "computer" were removed as being too common to be clues, along with those words

only appearing once or twice, leaving just over 1000 different words. Finally words with a uniform distribution across the categories were removed as clue words.

The weights were then estimated for the remaining clue words by looking at their occurrences within the categories in the training group.[1] The training on group 1 was run; results on group 1 showed that for 209 out of 247 cases, the highest match was the correct category, with the best results for those documents with more clue words. The results were much worse in the test group, with the correct match for only 51% of the 85 documents that had more than one clue word. Whereas these results may have been disappointing, the ideas going into clue word selection were echoed in later work with such as removal of common words, and in the term distribution measures such as IDF (see Section 4.1.4) in the 1970s.

Work also continued on automatic abstracting, such as examining how well automatic abstracts (built similarly to Luhn's "top ranked" extracted sentences) performed in a well-designed user study that involved both question answering and making relevance judgments (Rath *et al.*, 1961). The questions were based on information found in twenty-one documents from Scientific American, Neurology, Science and Time. Four different document "summaries" were produced: the titles, the automatic abstract, the same number of sentences (taken from the full text) as the automatic abstract, and the full text. Fifty participants, each using one of the four different "summaries", were asked to mark documents as relevant if they thought the documents would answer the question and then try to answer the questions. The results indicated that there were no major differences between the groups using complete text and the automatic abstracts to select relevant documents, but that the group with the complete text obtained a significantly higher score on answering the questions.

A survey (Edmundson and Wyllys, 1961) looking at three different automatic abstracting methods (Luhn's and two variations) recommended that more investigation was needed into better measures for the

[1] This was expressed as a probability ratio of the number of occurrences of the ith clue word belonging to documents in the jth category divided by the total number of clue word occurrences in all documents belonging to the jth category.

significance of words, the significance of groups or pairs of words, and the significance of sentences. In particular they suggested that relative measures be used for word frequencies and went on to propose several different measures. Figure 1 in that paper shows an early comparison of various mathematical ways of expressing relative frequency, with these ideas later picked up by automatic indexing researchers.

Researchers in automatic indexing were still leery of using just the frequency measures of words to index documents. This was partially due to the bias inherited from the complex subject headings in use, but also to experiments such as the one by Donald Swanson (Swanson, 1960) showing that adding thesaurus terms to requests based on subject headings retrieved more relevant documents (also more documents). Maron's work on probability had suggested using related words (and related documents) to expand requests and the idea of constructing these association matrices or "thesaurii" was a popular research theme.

One of the early investigations into associations (Stiles, 1961) was run on the existing ASTIA collection of 100,000 documents already indexed by the Uniterm Coordinate Index. Although there were no experiments reported, the paper discussed a way of building expanded requests by adding associated terms that were picked via a two-step associative process. First all the document terms co-occurring with the request terms would be ranked based on a statistical co-occurrence factor, creating an initial set of possible expansion terms; this initial set would then be processed against the full set of document terms to gather even more related terms. This final expanded list would be checked against a threshold and reweighted by the "sum of the association factors for each term, divided by the total number of terms in the expanded list, giving us a weight which will enable us to arrange the terms according to their probable relevance to the request."

Work by Vincent Giuliano and Paul Jones (Giuliano and Jones, 1962) sponsored by the Air Force[2] and NSF was also looking at association matrices, but from the viewpoint of how these could be used in retrieval. Their idea was that a retrieval score could be viewed as a product

[2]Most of the early information retrieval research was sponsored by the U.S. military to complement their own classified work.

of the request and separate linear transformations of three matrices measuring different associations. The first matrix was the index term association matrix based on term occurrence in the collection, the second a document association matrix measuring document similarity, and finally the discriminant matrix which was the document by term matrix. The issue was how to turn these ideas into an operational prototype.

> The linear associative transformations may be developed from at least three equivalent points of view (Giuliano and Jones, 1962):
>
> 1. Reasoning along probabilistic lines, in which term-term association is regarded to be a Markov process.
> 2. Reasoning based upon an electrical network analog.
> 3. Reasoning based upon the imposition of certain mathematical constraints on association and identification transformations, primarily consisting of certain assumptions of linearity and normalizability of transformation matrices.
>
> All three approaches are ultimately equivalent in that they lead to the same set of mathematical formulas. The approaches differ in the interpretations they provide; each gives a different avenue of appeal to intuition.

Since computers at that time could not deal with (3) they built a demonstration based on (2) with 110 documents and 48 requests.

3.1.1 The SMART System at Harvard

At Harvard, and in obvious communication, another researcher was working along the same lines. Gerard Salton also had a contract with the Air Force (and eventually NSF) and from 1961 to 1975 produced a series of semi-annual reports (the ISR reports). These reports consisted of student papers and the initial technical work that usually resulted in the papers published by Salton and the group.

ISR-2, published in September 1962, discussed work on associative techniques using citations, with the resulting 1963 paper (Salton, 1963) giving details of this work. The 1963 paper showed the various association matrices needed, including a term-document incidence matrix whose elements contained the frequencies of the index terms in documents, and a term-term similarity matrix with elements indicating how often terms co-occurred in the same document. Salton went on to suggest that a simple cosine correlation could be used on these matrices to generate the similarities. He further suggested that a "new column representing the request terms could be added to the term-document matrix".

> An estimate of document relevance is then obtained by computing for each document the similarity coefficient between the request column and the respective document column. The documents can be arranged in decreasing order of similarity coefficients, and all documents with a sufficiently large coefficient can be judged to be relevant to the given request.

Figure 3 in that paper contains a flow chart of a "typical document retrieval system using term and document associations". This was the basis of what would become SMART.

The January 1964 ISR-5 report from Salton's group at Harvard (Salton, 1964b; Salton and Lesk, 1965) included a description of the first version of the SMART system. It was built for an IBM 7094 around a supervisory routine (CHIEF) which accepted input instructions as to how the text should be processed. In addition to basic processing, such as the removal of common words and the conflating of word stems, there were multiple processing options, such as the use of a basic alphabetic dictionary, a semantic concept hierarchy, and syntactic procedures using a phrase dictionary and structural matching methods. The table of contents of the June 1964 ISR-7 report (Salton, 1964c) shown below gives some idea of the complexity of the system (and also the large group now at Harvard).

```
The SMART system - An Introduction: Gerard Salton
The SMART system,General Program Description: Michael Lesk
```

```
Dictionary and Hierarchy Construction: Claudine Harris
Dictionary Lookup and Updating Procedures: Mark Cane
Processing of the Concept Hierarchy: George Shapiro
Syntax and Criterion Phrase Procedures: Alan Lemmon
Sentence Matching Program: Edward H. Sussenguth, Jr.
The Criterion Tree File: Tom Evslin, Thomas Lewis
Statistical Phrase Processing: Michael Lesk,Tom Evslin
Statistical Processing/Request Alternation: Michael Lesk
Housekeeping Routines; Michael Lesk, Tom Evslin
Possible Time-Sharing Organization: Joseph J. Rocchio
Clusters by Matrix Analysis: A. Richard LeSchack
```

The big issue was how to test these methods. There were no test collections; indeed there was very little machine readable text. However a small test collection was built with the 405 IRE abstracts (used by Maron earlier). There were 17 requests created by three project staff members with computer background; these same people made full relevance judgments for the 405 abstracts. By ISR-8 (December of 1964) there were some test results (Salton, 1964a; Salton, 1965).

There were nine different text processing methods tested, most of these using a thesaurus built for this collection with 600 concepts corresponding to 3000 word stems in computer literature. It is interesting to note the complexity of these runs compared to later work by Salton and others; this complexity was thought to be necessary in order to get reasonable retrieval effectiveness:

1. thesaurus, titles only (word stems in titles replaced by frequency weighted concepts from thesaurus);

2. thesaurus, use of hierarchy (word stems in full abstracts replaced by frequency weighted concepts from thesaurus and parent added from hierarchy);

3. thesaurus, logical vectors (all word stems replaced by thesaurus concepts but all given weight of 1);

4. word stems, full text (weighted word stems from full abstract);

5. thesaurus, syntactic phrases (weighted thesaurus terms looked up in phrase dictionary, phrase concepts used under certain cases);

6. thesaurus, use of hierarchy (same as 2 except child nodes were used instead of parent nodes);

7. thesaurus, numeric vectors (same as 3 but with frequency weights);

8. thesaurus, statistical phrases for requests only (method 7 but with added phrases for terms occurring in the requests);

9. thesaurus, statistical phrases (same as 7 but with added statistical phrases).

The results showed that "methods one to four tend to produce relatively poorer recall than methods five to nine" and that "these same methods also furnish relatively poor precision."[3] Additionally "the use of the regular thesaurus which provides vocabulary control (method seven) seems much more effective than the use of the original words included in document and search requests (method four)" and finally "the most effective procedures seem to be those which use combinations of concepts (phrases), rather than individual concepts alone."

There were two more ISR reports done by the Harvard group, with ISR-9 (Salton, 1965) including several papers on the updating of the SMART system to handle longer documents (the soon-to-be-arriving Cranfield collection) and one paper by Rocchio discussing evaluation viewpoints. In that paper he discussed the difference between micro-averaging results (using the cumulative number of retrieved documents over all requests) and macro-averaging (where each request is evaluated and then the results are averaged).

> Macro evaluation is a query oriented viewpoint, and as such any performance index formulated, on this basis would be query distributed or averaged on a per-query basis. The justification for considering queries as atomic is simply that this corresponds to the view-point of the system user. . . . The

[3] See the next section for the definition of recall and precision.

> sample mean of an evaluation parameter obtained by averaging over the total number of queries represents an estimate of the worth of the system to the average user.

ISR-10 (Rocchio, 1966) is Joseph Rocchio's thesis on relevance feedback, i.e. the famous Rocchio algorithm! (For an interesting look at the inside operations of SMART lab both at Harvard and Cornell, see Lesk *et al.* (1997).)

3.2 Indexing Wars, Round 2: The Cranfield Tests

Information retrieval was gathering more attention. The IFIP Congress 62 held in Munich in September 1962 had a session on Information Retrieval, including papers by J. C. Gardin on the French SYNTOL system, Roger Needham on two experiments done on the Cambridge EDSAC II system and the previously-mentioned paper on searching by Donald Swanson. The 1963 ADI conference produced its proceedings (under direction of Luhn) in machine-readable format so that it could be used for experimentation. In 1964 the U.S. National Bureau of Standards (now NIST) held a conference on Statistical Association Methods for Mechanized Documentation (Stevens and Giuliano, 1965), including a keynote by Luhn, three days of papers by Maron, Giuliano, Edmundson, Needham, Salton, and others, plus panels including Cyril Cleverdon, Calvin Mooers, and John Tukey. An interesting paper by Sally Dennis of IBM (Dennis, 1964) presented twelve different statistical methods for describing the distribution of unique words in a set of 2649 appealed legal case studies, in particular looking at the skewness of the distribution of these words (a type of language modeling done very early).

The library community had not been sitting still. In 1955 at a Special Libraries meeting in Detroit, Cyril Cleverdon, Librarian of the College of Aeronautics, Cranfield, U.K. presented a paper calling for more testing of indexing. Helen Brownson of the NSF asked Cleverdon for a proposal and two years later (1957) sent $28,000 to fund the first Cranfield study (Cranfield 1). This study (Cleverdon, 1962), running

from 1958 to 1962, was specifically designed to test four manual indexing methods.

Three indexers spent two years producing indexes for 18,000 papers and reports from the field of aerodynamics using the four indexing methods: an alphabetical subject catalogue, a faceted classification scheme, the Universal Decimal Classification and the Uniterm system of co-ordinate indexing. Cleverdon used the "source document" method of evaluation, 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". The searching process consisted of using each index to manually search for the documents, with results showing that 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.

Whereas these 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. Descriptors could consist of multiple terms (phrases) or just be a single term and there could be multiple descriptors per document. More descriptors (exhaustive indexing) leads to better recall (generally), but at the expense of precision (called "relevance" up to 1964). The problem of how to select content descriptors for indexing led Cleverdon to continue his investigations in the Cranfield 2 (1962–1966) project (Cleverdon *et al.*, 1966; Cleverdon and Keen, 1966).

For Cranfield 2 Cleverdon 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 assessments of likely relevance of each of the 1400 documents for each question. Five graduate students spent the summer of 1963 making preliminary (and liberal) judgments for 361 questions against the 1400 documents. These judgments were then conveyed to the authors of the question 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. The test collection consisted of the 1400 documents, the 221 questions, and the list of relevant documents for each question.

Cyril Cleverdon and his team proposed an incredibly ambitious set of variables, especially considering that all of the experimentation was done manually, including computation of the numerical scores on mechanical calculators!

There were two types of indexing: manual indexing of the full documents at 3 levels of exhaustivity (averaging 31, 25, and 13 descriptors per document), and "automatic" indexing using the natural language abstracts and titles. For the manual indexing, there were four types of indexing languages used: single terms, simple concepts, controlled terms, and indexing only the terms contained in the abstracts and titles.

The documents were 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 could then be broken into the single terms, with weights assigned to these terms. The controlled terms were created by translating the simple concepts into the vocabulary of the Thesaurus of Engineering Terms of the Engineers Joint Council.

In addition to the different types of indexing tested, Cleverdon also wanted to try what he called "recall devices", such as the use of synonyms and/or hierarchies from a thesaurus and the use of the (stemmed) word forms. There were also "precision devices" such as weighting and the use of co-ordination (the Boolean "anding" of terms or concepts during the search process).

It soon became apparent that there was too much to do; several subsets of the collection were then used. In particular, the Cranfield 200 collection subset was created using 42 questions on aerodynamics, along with 198 of their relevant documents (but not the source documents for these questions). This was the collection used for some of the experiments.

There was also the issue of how the evaluate the results, in particular which metrics to use. There had been a lot of discussion previously about metrics, centering around the well-known categories shown in Cleverdon's Table 2.1 copied here.

Cleverdon decided to use the "Recall Ratio" defined as $a/(a+c)$, and the "Precision Ratio" $a/(a+b)$. These had been used by Kent and Perry (Kent *et al.*, 1954) and called the Recall Factor and the Pertinency

Table 3.1: Possible categories of documents in searching

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

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 Sanderson (2010) 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 choose these over more complex formulas such as those suggested by Bourne, Farradane, Vickery, etc.

The experiments were done by creating a series of rules that governed each of the many possible combinations of variables. The searchers then manually followed these rules using coded cards stored in what was known as the "Beehive". So for example, one specific type of index was selected, such as the use of the simple concepts and then a series of precision devices were applied using different levels of co-ordination on a per question basis. First all simple concepts in the question were "anded"; then one less concept was used, and so on until only one concept was used for searching.

It should be noted that for each of the 221 questions the recall and precision ratios measured a *single* point. Using the example 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 specific coordinate level.

Figure 3.1 shows the results for one of these experiments (Cleverdon and Keen, 1966) and clearly illustrates the effects of coordinate ranking.

FIGURE 4.100T

Index Language I.1.a. (S.T. Natural language. Coordination)

Exhaustivity of Indexing 3
Search Rule A
Document Relevance 1-4

Number of Documents in Collection 1,400
Number of Questions 221 (Subset 3)
Number of Relevant Documents 1,590
Generality Number 5.1

Coord-ination Level	Documents Retrieved Rel.	Non-rel.	Recall Ratio a/a+c	Precision Ratio a/a+b	Fallout Ratio b/b+d	x	y	z
1	1,510	159,122	95.0%	0.9%	51.696%	221	221	221
2	1,283	58,122	80.7%	2.2%	18.883%	221	221	221
3	946	21,933	59.5%	4.1%	7.125%	215	220	220
4	606	7,359	38.1%	7.6%	2.390%	187	212	212
5	314	2,380	19.7%	11.6%	0.773%	131	197	197
6	154	699	9.7%	18.0%	0.227%	86	164	164
7	74	216	4.7%	25.5%	0.070%	50	140	140
8	22	43	1.4%	33.8%	0.014%	18	105	105
9	8	5	0.5%	61.5%	0.002%	8	78	78
10	1	0	0.1%	100.0%	0.000%	1	52	52
11	0	0				0	32	32
12	0	0				0	15	15
13	0	0				0	8	8
14	0	0				0	4	4
15	0	0				0	3	3

Figure 3.1: Results using natural language and coordination ranking.

So what were the results of this huge set of experiments? Figure 3.2 reproduces the top portion of Figure 8.100T in Cleverdon and Keen (1966) showing an ordered list of the effectiveness of 33 different "index languages" encompassing the types of descriptor terms used. The first seven of these are using the 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.

> 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

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. Quasi-synonyms
7=	63.05	II-11	Simple concepts. Hierarchical and alphabetical selection
9	62.88	II-10	Simple concepts. Alphabetical second stage selection

Figure 3.2: Effectiveness of index languages.

> 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 (Harter, 1971; Rees, 1967; Swanson, 1971). 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 of the Cranfield paradigm (although the general consensus was that the experimental results are valid). In terms of Cyril Cleverdon's reaction, Michael Lesk (Lesk, 2018) reported that:

> Cyril got quite frustrated that all of these different methods had similar results. So he suggested the idea of "cost per relevant document retrieved" which gave a substantial advantage to the simplest indexing method and avoided arguments about whether a small advantage to one or another method in one or another collection justified its use. Everybody else didn't want to mix up economics with performance.

The achievements of the Cranfield work are four-fold. First, Cleverdon established a way of testing that is still in use today. The Cranfield paradigm 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: first, the careful modeling of the task being tested by the test collection and second, his strict separation of the building of the test collection from the experimentation itself.

His second contribution was the firm validation of the use of single terms occurring in the natural text as opposed to a need to manually index. This did not in any way invalidate the importance of manual text analysis for some areas (such as biological science) but it opened the way to the *assumption* that the normal indexing for retrieval would be the text itself.

The third contribution was the work with metrics, in specific the recall/precision graphs and the various intellectual investigations around these. Even if these graphs were generated differently (because of the single-point measures), they illustrated that the various techniques produced different levels of recall and precision and that this was generally an inverse relationship.

The last contribution was that it was the first experiment with a realistic test collection to demonstrate the effectiveness of ranked output (i.e., the coordination of terms), validating the work that had started with Luhn and Maron. And because there was now a realistic test collection, others (such as Salton) could go on to further demonstrate the effectiveness of ranking.

(To read Cleverdon's own account of the Cranfield tests and their background, see his acceptance speech for the 1991 SIGIR award (Cleverdon, 1991).)

3.3 More Focus on Searching

Computers were fast improving: the IBM 7094 had 32 K 36-bit words of magnetic core memory, a clock speed of 0.5 MHz, and cost around 2 million dollars at that time. The competing CDC 1604 had 32 K 48-bit words of magnetic core memory and a clock speed of 0.6 MHz. The IBM

360/65 model (first shipped in 1965) had a clock speed of about 1 MHz and depending on the configuration, there was between 128–1,024K of core memory. It cost about 200,000 dollars a month to rent (1.6 million in 2018 dollars).

These machines could come with removable media, i.e., a disk pack that held up to 29 megabytes. The IBM 360 family soon dominated the market because it worked well for both administrative and scientific applications and was upwards compatible so that companies and universities could buy the lower-end machines and then upgrade *without* having to change their programs. Over a thousand orders were received in the first month after its announcement.

3.3.1 The SMART System at Cornell

Gerard Salton moved to Cornell University in 1965 to join the first computer science department (Lesk *et al.*, 1997). There was initially no SMART staff at Cornell, and also no SMART system since the system built for the IBM 7094 was not portable. Therefore for several years much of the processing stayed at Harvard. However Michael Keen (from the Cranfield project) came to Cornell in 1966–67 to help with analysis of the experimental runs done at Harvard.

There were now three different test collections for experimentation. The IRE-1 collection had been extended using similar abstracts and 17 requests written by one non-staff person to become the IRE-3 collection. The ADI collection was the set of short papers from the 1963 Annual Meeting of the American Documentation Institute; this collection was built to allow experiments with document lengths since there were titles, abstracts, and full texts. The requests for ADI were created by two technical staff not familiar with the system, and once again these people made relevance judgments against the full collection. The final collection was the Cran-1 collection, the 200 document/42 request collection from the Cranfield project. Whereas this collection was only titles and abstracts, there were manual index terms and also a thesaurus.

ISR-13, published in December of 1967 (Salton, 1967), contained seven sections by Michael Keen with extensive analysis of results on these three collections. The first two sections (Keen, 1967a,b) covered

the details of the collections and a major examination of metrics for evaluation. Work done at Harvard earlier on evaluation was extended by Keen to tease out the nitty-gritty details of interpolation and averaging results across full request sets. This section is a goldmine for readers interested in micro-averaging vs. macro-averaging, and also the various interpolation/extrapolation methods considered for recall/precision curves.

The next five sections looked at weighted (numeric) vs. non-weighted (logical) concepts, overlap vs. cosine similarity matching, the effects of suffixing, the effects of document length (and manual indexing), and the uses of a thesaurus, phrases and hierarchies. Whereas these results are what would be expected today, they were very important at this time because they validated the earlier theories and changed the direction of research in the late-1960s. The failure analysis was particularly interesting and showed differences across the three collections (demonstrating why it is important to use multiple collections for testing).

1. Weighted vs. Non-weighted: "The superiority of weighted concepts evidenced by the superiority of numeric as opposed to logical vectors is due to two reasons. The first is that highly weighted matching concepts tend to distinguish between important and trivial occurrences of those concepts in the documents, and thus tend to make better distinctions between relevant and non-relevant. The second reason is that if different concepts in a request receive different weights, such weighting does discriminate between vital and unimportant request notions so that in some cases otherwise similarly matched relevant and non-relevant documents are correctly separated (Keen, 1967c)."

2. Overlap vs. Cosine: "The individual relevant documents display large changes in rank with change in correlation coefficient. Using the Cran-1 Stem results, it is found that of 198 documents relevant to all 42 requests, 95 show rank improvements on cosine over overlap, 62 show the reverse improvement, and 41 show no change in rank. Since the analysis shows that non-relevant documents with strong matches are longer than average, it is now obvious that cosine effectively lowers the ranks of these documents, and

thus provides a better retrieval performance than overlap (Keen, 1967c)."

3. Stemming: There were two different stemming possibilities: an S stemmer (plurals only) and a full suffix removal process using a stem dictionary. "The comparison of the two suffixing dictionaries shows the full stemmer to be superior on the IRE-3 and ADI collections, and S stemmer to be superior on the Cran-1 collection. The aerodynamics terminology appears to offer less opportunity for word conflation than the computer science and documentation terminologies." There were several serious conflating errors in Cranfield; for example, the stem "compress" incorrectly matches the request word "compressor" with the frequently used word "compressible" (Keen, 1967e).

4. Document Length and Manual Indexing: The abstracts perform better than the titles, although there is a small difference for the Cran-1 titles vs. abstracts because the titles are long and very technical. For the ADI collection "the average performance results show that although (full) text is superior to abstracts, the improvement is small". For the Cran-1 (manual) indexing, there is a small superiority of over abstracts possibly because "the indexers chose some terms from the full texts of the documents that the abstractors failed to include, and some of these terms represented subject notions that were asked for in the requests and also by choosing nearly half the number of terms contained in the abstracts, the indexers avoided notions that are not asked for in the requests" (Keen, 1967d).

5. Thesaurus, phrases and hierarchy: "The best thesaurus dictionaries give a performance superior to the stem dictionary on the average; the superiority of thesaurus is least marked in the Cran-1 collection" (see Keen's Figure 5 reproduced here).

   "Phrase dictionaries give a superior performance compared with thesaurus alone by a very small amount only on IRE-3 and ADI, and on Cran-1 the thesaurus alone gives a slightly better result. Looking at the performance of the hierarchy on an individual

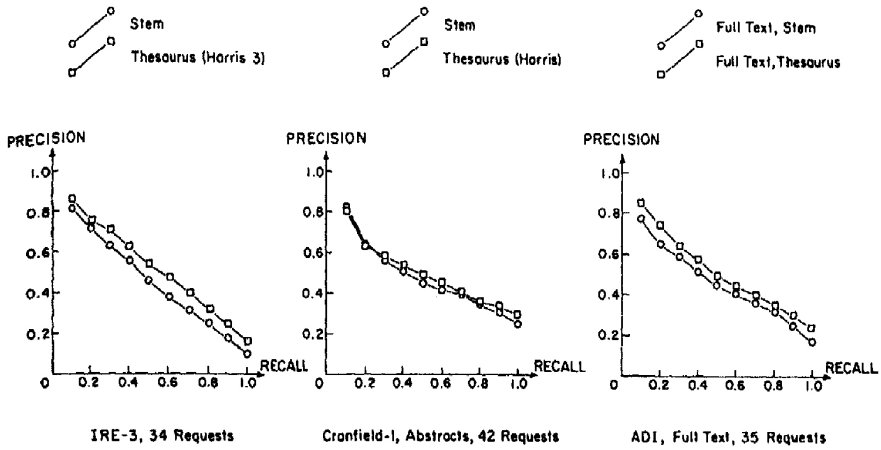

FIG. 5. Comparison of synonym recognition (thesaurus) with word-stem matching process

request basis, the thesaurus is equal to 'parents', and superior to 'all' relations; the hierarchy is thus not to be preferred."

"This analysis and discussion of the phrases and hierarchy has shown that, in their present form, these two types of dictionary do not improve the thesaurus process by an amount that would justify the effort required for construction. Indeed, it might even be questioned whether the effort of constructing a thesaurus itself is worthwhile, since results such as those given in Figures 14, 15, and 16 (not shown here) prove that the improvement of performance in comparison with the stem dictionary is not really large. In situations where economic consideration are all important, or time is very limited, it seems that an automatic stem dictionary will perform quite well, particularly for the high precision user. It is disappointing that the thesauruses tested do not always help the high recall user" (Keen, 1967f).

The paper by Salton and Lesk (Salton and Lesk, 1968) summarizes these results, adding significance testing. They show that the use of abstracts is significantly better than titles for all collections, that the full text for ADI did not perform significantly better than the abstracts, and that the use of stemming did not give significant improvements. For two of the collections (IRE and Cran-1) the use of weighted terms

and the cosine measures were significantly better than no weighting and the overlap measure, however these were not significant improvements for the ADI collection. The use of a thesaurus brought significant improvement for IRE and ADI, but not for the Cran-1 collection. And finally the use of the phrases and hierarchies did not give significant improvement.

It should be noted that these larger experiments validated the work that had been done earlier at Harvard with the first IRE collection. Additionally, however, these were the first experiments that revealed how the different characteristics of test collections could affect results, with the analysis done by Keen providing some clues as to the causes of the differences.

In the summer of 1965 work was started on the SMART system at Cornell (Lesk *et al.*, 1997). The original specifications for this rewrite were as follows:

1. to continue to run the text analysis programs at Harvard, and to rewrite only those programs that dealt with document abstracts and queries[4] after conversion to numeric vectors;

2. to initially concentrate on two major areas at Cornell–the implementation of Rocchio's clustering algorithm and the exploration of various relevance feedback algorithms;

3. to write these programs to handle collections as large as the Cranfield 1400 document/225 query collection (sometimes called the Cran-2 collection at Cornell);

4. and to make the resulting system available for use (including experimental reprogramming) to students on one-semester projects.

By fall of 1966 the CDC 1604 version of SMART was finally implemented; however at that point, Cornell decided to buy a brand-new machine, the IBM 360/65 and work started all over again. The system design was similar, and luckily the modularity of the first system

[4]The use of the term "queries" at Cornell started in the mid 1960s, but this use is synonymous with the term "requests" used earlier. Both words refer to the information request in a test collection.

made many of the higher-level routines less work, but the machine-level routines all had to be redone. By the end of 1967 most of the system was re-implemented, and by mid-1968 the text processing part was transferred to Cornell from Harvard. For details of the SMART system in 1968, see Williamson (1968). That year also saw the publication of the first book on these new methods of retrieval (Salton's *Automatic Information Organization and Retrieval* (Salton, 1968)).

At this point, with the basic retrieval methodology in place, 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 the new ISPRA test collection was built specifically to address this problem (Lesk and Salton, 1968, 1969). The ISPRA documents were 1268 abstracts in the field of documentation and library science from American Documentation and several other journals. 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 by library science students. 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 it was shown that the performance ranking of the various text processing methods (words vs. stems vs. thesaurus) *did not change* depending which of the relevance judgment set that was used. Analysis of the results revealed the likely reason for this:

> The conclusion is then obvious that although there may be a considerable difference in the document sets termed relevant by different judges, there is in fact a considerable amount of agreement for those documents which appear most similar to the queries and which are retrieved early in the search process (assuming retrieval is in decreasing correlation order with the queries). Since it is precisely these documents which largely determine retrieval performance, it is not surprising to find that the evaluation output is

substantially invariant for the different sets of relevant (Lesk and Salton, 1968, 1969).

Cyril Cleverdon also did a test for consistency of relevance judgments for the Cranfield 200 test collection. He asked two additional judges to look at a set of documents for each request; that set consisted of the original relevant documents plus "a number of non-relevant documents retrieved at high coordination levels and also non-relevant documents having a relatively high bibliographic coupling with known relevant documents." Here he was looking at how the ranked list of *index methods* changed depending on which assessments were used. "The figures appear to confirm the original hypothesis, namely that the relevance decisions did not significantly affect the comparative results of the Cranfield II (author's comment: usually called Cranfield 2) project–the (Spearman) rank correlation never falls below 0.92%" (Cleverdon, 1970).

The ISPRA collection was also used for the first test of cross-language retrieval (Salton, 1969a). 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. Four sets of runs were made (E-E, G-G, G-E, and E-G), using the English thesaurus and a German thesaurus that was a translation of the English one. The runs using the German collection were less effective; failure analysis revealed two problems. First whereas the English thesaurus was missing only 6.5 words per English abstract, the German one was missing over 15 words per abstract and many of these words were in the requests. There also looked to be problems with the German relevance assessments. However the runs did show that the same type of basic processing worked for languages other than English, with failure analysis finding similar problems seen today in cross-language retrieval.

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 (Lesk *et al.*, 1997). This involved scalability of the

ranking algorithms, and this critical need for increased efficiency drove many of the clustering theses during those days. (Note that it was assumed that searching clusters was more efficient and effective than sequential searching, later proven incorrect (Voorhees, 1985).)

The first SMART MEDLARS test collection (Salton and Williamson, 1968) was based on the study by F. W. Lancaster on the effectiveness of the MEDLARS system (see Section 3.4.1). Eighteen of the requests that had been used in that study, along with 273 of the 518 abstracts and the full set of relevance judgments were used to compare the SMART system retrieval to the MEDLARS Boolean results.

The problem came in this comparison of the results. Lancaster had measured his recall and precision by using two *separate* collections, a recall base and a 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 Salton (1969b)) show both the creativity and the dedication to proper testing methods possessed by Salton and Keen.

Work was also started both in the use of relevance feedback and clustering (see the next chapter for more on this).

3.3.2 The Comparative Systems Laboratory (CSL) at Case Western University

In about the same timeframe as the Cranfield study and the SMART system, the Comparative Systems Laboratory (CSL) was created by Tefko Saracevic at Case Western University in Cleveland. This project was notable not only for further experimentation in indexing and searching, but also for a "parallel" course (Saracevic, 1968) 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 (Saracevic, 1971). 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 project, 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 volunteers, specialists in their field, who were asked to "provide questions related to their current research interest". The documents were manually indexed using five different indexing languages: keywords selected by indexers, keywords selected by a computer program, a telegraphic abstract (index terms assigned without constraints including syntactic roles and links), a metalanguage and the Tropical Disease Bulletin index terms. The indexing was done on titles or abstracts or full-text for the first three indexing languages, resulting in 5 to 8 terms for titles, 23 to 30 terms for abstracts, and 36 to 40 terms for full-text.

There was 124 questions gathered from the users that were then used for searching with five types of question analysis done on each question (A: unit concepts (just the terms in the original question), B: expansion of A using a thesaurus, C: expansion of A using other tools, D: further expansion of C using thesaurus, and E: a user-verified version of D). Note that these question analysis types were mirroring the types of *natural* steps that a professional searcher (search intermediary) might make, as opposed to the more constrained approach to searching that was used in Cranfield. Additionally all of the searches were done in 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 both 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 non-relevant. 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 non-relevant documents NOT retrieved divided by the total number of non-relevant 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.

This was a massive experiment, with multiple indexing methods working on titles, abstracts, and full text. There were then searches done using the five different types of expansion and narrow vs. broad searching. The paper continues with analysis of all the results: below are some of the more important conclusions.

1. The number of terms assigned per document effects system performance more than does the particular index language used (much the same conclusion as the Cranfield testing).
2. "User questions cannot be searched as stated; as a rule expansion of question terms is necessary. It appears that expansion is best achieved when every available tool (including personal knowledge) is used."
3. "The contact with users in order to validate and expand question terms did not improve system performance: as a matter of fact and quite surprisingly, it slightly decreased the performance."
4. The use of broad searches did retrieve more relevant documents, but with a large cost of non-relevant retrieved. "Thus, it appears that as the retrieval of relevant answers works its way asymptotically to its maximum, the retrieval of non-relevant ones soars almost exponentially."

This was the first set of experiments that looked at manual search methods in addition to the indexing methods. The conclusions reached became part of the methodology that was taught in library science and information science schools to what was to become the new profession of search intermediaries. Further experiments in the 1970s and 1980s

confirmed these results in terms of search methodology, and the finding that the number of non-relevant documents "soars exponentially" was to drive the strict adherence to Boolean by the later online systems.

3.3.3 Other Information Retrieval Research in the Late 1960s

Two more research areas from the late 1960s need mentioning. The first is Julie Beth Lovins stemming algorithm (Lovins, 1968) developed at MIT in 1968. This algorithm was a context-sensitive longest-match stemming algorithm for English that became the stemming algorithm of choice by SMART (and other systems), and later as an alternative to the 1980 Porter stemmer.

The second area of research dealt with evaluation metrics. Whereas the recall/precision metrics of Cleverdon, Rocchio, Keen, etc. were well accepted, there was no consensus on a single number metric to use in comparing systems. John Swets from BBN proposed the use of the E measure and the use of ROC curves derived from statistical decision theory (Swets, 1969). His paper not only presents the reasoning behind the use of this measure but illustrates its usefulness against the experimental data from SMART, Cranfield and testing by the Arthur Little Company. Stephen Robertson's two 1969 papers (Robertson, 1969a,b) discuss in detail the Swet's measure and its variations.

Another single-number measure, the expected search length, was proposed by William Cooper from the University of Chicago (Cooper, 1968). Cooper, like Cleverdon, approached evaluation from the standpoint of users, in particular measuring the effort needed by users to satisfy information needs. He noted that users could be looking for only one "answer" document, could be looking for a small sample of documents, or could want everything that was relevant. The proposed metric (expected search length) could work for all these variations, but has been used heavily in applications such as known item search (where only one item is wanted).[5]

[5]Note that the idea of known item search is similar to the earlier Cranfield source document idea, however in that case the request was based on the source document. Known item search only implies that there is one document that is being used as the evaluation point.

3.4 Operational Systems

There were beginning to be serious operational systems in the 1960s. These systems were basically the "descendents" of the Taube ideology of the Boolean combining of terms, using mostly strict "ands". They matched manually constructed requests to manually indexed documents, returning an unranked set of citations. This was absolutely necessary at this time because of the slowness of the computers (running tape searches) and from the lack of machine-readable text to provide any type of automatic indexing.

A major source of information about the early operational systems is the book by Charles P. Bourne and Trudi Bellardo Hahn (Bourne and Hahn, 2003). This book covers the period between 1963 to 1976, providing extensive details on these early systems. These details cover not just the implementation issues, but interesting discussions of the business deals behind this systems.

3.4.1 The MEDLARS System

By far the largest (and most complex) was the MEDLARS system (Medical Literature Analysis and Retrieval System) started in the early 1960s at the National Library of Medicine (NLM). Index Medicus, a monthly guide to medical articles in thousands of journals, had been published by NLM starting in 1879; by 1961 each monthly edition averaged 450 manually compiled pages referencing more than 10,000 articles. Additionally there was manual indexing of over 120,000 items annually, which was expected to double by 1969 (US Public Health Service *et al.*, 1963).

In 1960 (after a less-than-successful trial of an electronic publishing system using a high-speed camera), the NLM staff put together objectives and specifications for the MEDLARS system. The objectives of MEDLARS may be summarized as follows:

1. Improve the quality of and enlarge IndexMedicus and at the same time reduce the time required to prepare the monthly edition for printing from 22 to 5 working days.

2. Make possible the production of other compilations similar to Index Medicus in form and content.
3. Make possible for IndexMedicus and other compilations, the inclusion of citations derived from other sources, as well as from journal articles.
4. Make possible the prompt (a maximum of two days) and efficient servicing of requests for special bibliographies, on both a demand and a recurring basis, regularly searching up to five years of stored computer files.
5. Increase the average depth of indexing per article by a factor of five, i.e., 10 headings versus 2.
6. Nearly double the number of articles that may be handled annually from 140,000 now to 250,000 in 1969.
7. Reduce the need for duplicative literature screening operations.
8. Keep statistics and perform analyses of its own operations.

In 1961 a request for proposals went out to 72 companies; General Electric was the winner and in 1963 a Minneapolis-Honeywell 800 was delivered to NLM. There was a Graphic Arts Composing Equipment (GRACE) computer-driven phototypesetter for the publications: its first output was the August 1964 Index Medicus (Dee, 2007). The first test searches started in January of 1964, and within the first year there had been over 1,100 searches run for physicians, researchers, etc. (Bourne and Hahn, 2003), making it the first large-scale, computer-based, off-line batch service, open to the general public.

By April 1965 there were 265,000 citations in the database, with over 14,000 records being added monthly. The manual indexing created the unit records for each citation, these were then turned into paper tape for input to the computer (Figure 3.3).

The searching (called a demand bibliography) required the users to complete a search request form, which was converted by trained medical librarians into the search format (Dee, 2007). The request then had to be passed against the entire file of citations (now on magnetic tape), which took about 40 minutes. To improve turnaround, the searches were

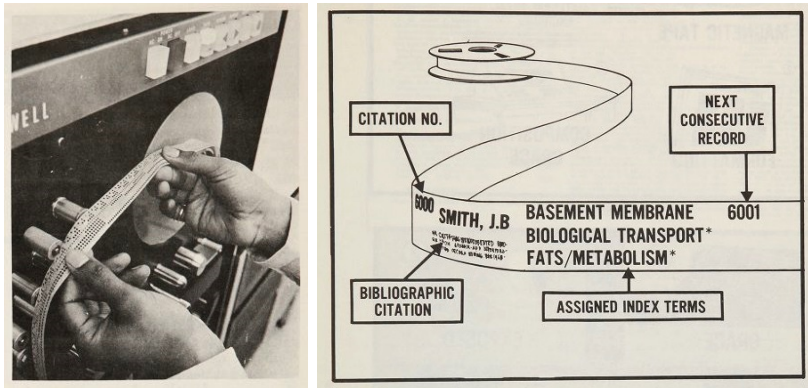

Figure 3.3: Paper tape reader and a schematic of the information on the tape.

run in batch mode (all searches batched for execution during one tape pass) and the results sorted (and printed out) at the end.

During 1966 there were 3035 demand bibliographies produced. Service was getting slower and NLM asked F. W. Lancaster, then working at NLM, to make an evaluation of this search service: to find out how well the service was meeting requirements, to identify factors affecting performance, and to suggest improvements. The evaluation was also required to create a test collection, with documents, requests, indexing, search formulations and relevance assessments, and Cyril 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 by users. Twenty-one groups across the health spectrum were asked to collect requests as they arrived and these users were asked for known relevant documents and also asked to judge a sample of those that were returned. This gave an estimate for precision: recall was calculated based on a complex method using a "recall base" (Harman, 2011; Lancaster, 1969).

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, with huge variation in performance across the requests. The detailed failure analysis showed indexing failures, search failures and also problems with the "user interface". Recommendations

to NLM included asking users to fill out a more detailed search request rather than just talking with the intermediary and to check boxes about things like male/female patients, animal experiments, etc., i.e., things that would match the indexing checkboxes. It was also suggested to have a greatly expanded entry vocabulary available online.

3.4.2 Other Early Operational Systems

Another early online retrieval system was the prototype TIP system at MIT, accessible from 100 teletype terminals on the campus starting in January 1965. It had about 30,000 articles from several years of different physics journals. and could retrieve citations by bibliographic coupling, cited references, and Boolean searching.

Two different online systems were built for NASA's 300,000 bibliographic records; Bunker-Ramo developed the first dial-up service in 1966–67 and Lockheed built an in-house installation and then an online version in 1969. This online version ran on 24 consoles at 9 locations by the end of 1969, searching files with about 700,000 bibliographic records. Data was stored on a combination of magnetic disks and IBM Data Cells. Note that Lockheed's online RECON was based on their early proprietary system called DIALOG developed in 1967 (Bourne, 1961; Bourne and Hahn, 2003).

There were also beginning to be commercial (for profit) systems, developed for (very) specialized areas. The Law Research Service was the first computer-based search service operated on a regular basis, with access to over one million abstracts of New York case law in 1964.

4

Consolidation (1970s)

4.1 Improving Search Effectiveness

The 1970s had two separate but parallel themes, both enabled by major increases in computer power and storage. This allowed researchers to try more computer-demanding experiments, and encouraged more commercial online efforts. The IBM 370 series, introduced in 1970, was backwards compatible with the 360 series (so that programs were portable), and the 370/165 had five times the speed of the 360/65, with core memory now up to 3 MB. These computers were also more application friendly, with multi-tasking, virtual storage, I/O buffers, disk storage, etc.

4.1.1 Request Expansion using Word-Word Associations

Research by the SMART group

In terms of research, work continued on various ways of expanding requests, including word-word associations, document-document associations, and relevance feedback. An experiment by the SMART group (Lesk, 1969) used the same three collections used by Keen (ADI, IRE-3, Cranfield 200) to investigate the use of word-word associations

to expand requests. The association scores were calculated by frequency-weighted cosine correlations of all word pairs in a given collection (28,680 word pairs in Cranfield with 0.6 or higher correlation). Over 18,000 of these involved words with fewer than three occurrences in the collection and were eliminated as being random co-occurrences; similarly 49 pairs containing words occurring in over half the collection were eliminated. An extensive analysis of the remaining word pairs presented an unexpected finding:

> Since it has been suggested that word associations can be used to construct thesauruses, it is important to know whether word-word pairs produced by an association process reflect semantic meanings. . . . Each word pair was examined and judged either as significant or non-significant. Significant pairs are those pairs which seem to be composed of semantically related words. The words are judged to be semantically related if they would normally be used together in discussions of the same topic, considering the most common technical definitions of the words to be their meaning. Overall, only 16.2% of all correlations are judged significant.

More evidence of this comes from the fact that the word-word associations for the IRE collection had only one word pair in common with the hand-built thesaurus. Experiments with the full text ADI collection showed only 19% of the word pairs could be considered semantically related. Further analysis of the word-word associations in Cranfield found the following:

> When the meanings of the word in the collection are considered, it is found that about 73.1% of the pairs are significant, in this "local" sense. For example, consider the associated pair "scheme" and "machine". This was rated non-significant, since the words in their normal technical meanings are not related. However, it is found in examining the ten occurrences of "machine" that all ten imply "digital computing machine"; To determine the fraction of "significant" pairs on the local basis, the list of pairs was rechecked for significance

> and each word looked up in a concordance of the text to determine its local meaning.

Performance on the Cranfield collection using requests expanded by the word-word associations showed improvements, but whereas recall improvement would be expected, the improvement was even more pronounced in precision where the added terms increased the weights of important concepts. Looking at the individual requests, the word-word association improved 25 requests, the thesaurus 24 requests and using the manual index terms improved 24 requests: only 12 requests were improved by all three methods. Basically these methods detect different word relationships. Lesk summarized his results as follows:

- on small collections, associations are not for determining word meanings or relations, since the majority of the associated pairs depend on purely local meanings of the words and do not reflect their general meaning in the technical text;
- associative retrieval is not an effective recall device but is rather a precision device in many cases, operating by increasing the weight of significant terms rather than by introducing new significant terms;
- as a method of improving both precision and recall, a properly made thesaurus is generally preferable to associative procedures.

Research at Cambridge University

There had been investigations ongoing in Cambridge University (Needham and Spärck Jones, 1964) on a type of word-word association to classify keywords into groups or clumps. The advent of the Cranfield collection allowed experiments with these ideas (Spärck Jones and Jackson, 1970; Spärck Jones and Needham, 1968). They used the manually indexed version of the Cranfield 200, with a total of 712 keyterms across the document collection (32 terms per document average). These terms were built into large co-occurrence matrices, and then classes were

formed starting with a given term as a "seed". There were four different ways of building classes: strings, stars, cliques and clumps.

> Thus for strings, we take an element, find an element connected to it, and then a further element connected with the second, and so on; for stars, we take an element, and find elements directly connected with it; for cliques, we find a set of elements which are all connected with one another; while for clumps, we may accept as a class a set of elements with irregular connection pattern if the other external connections of the objects concerned are not so powerful.

As with earlier work, terms with high or low frequencies could be treated differently, in particular there was a restricted set in which high frequency terms were eliminated (leaving 616 terms).

Once the classes were formed, experiments were run using the 42 requests. The base run was a co-ordination run using the terms from the requests against the index terms in the documents. The main experiment was to substitute the class of a term for the term itself; there was also a mixed mode in which the request terms would be supplemented by classes. The results showed that the better runs used the class substitution alone (no request terms), that all four types of classes performed similarly and performed well only if they are kept small. An analysis of the results revealed some of the reasons.

- Both the frequent and infrequent terms are needed, but the former can only be effective if their matching potential is limited as much as possible, as it is if frequent terms are treated as unit classes; the appearance of frequent terms in term groups has unfortunate consequences which more than counterbalance their useful one.
- The fact that classifications of quite different types give similar results is also explained by the fact that they are confined to strongly connected items; for though the patterns of connections which were explicitly used

to form the successful classifications are different, the items brought together are strongly connected, and it is therefore frequently the case that there are other connections between them than those which were actually used.

- If we look at the successful systems as a whole, therefore, we can say that the reason why they are successful seems to be that they expand requests and document descriptions, but only to very closely related terms, and further, only to such related terms as are relatively discriminating.

4.1.2 Request Expansion using Relevance Feedback

Alternative ways of expanding requests include the use of relevance feedback; this assumes that there is user input (or simulated input) that provides indications of which documents are relevant. Rocchio's relevance feedback algorithm (Rocchio, 1965, 1971) developed in the mid 1960s, became the basis of a series of experiments at Cornell in the late 1960s. The Rocchio algorithm used the terms in the relevant documents to improve the requests; terms could be added from these documents or terms could be deleted if they occurred in non-relevant documents.

Early results of work using the abstracts and titles from the Cranfield 200 collection had been very promising, however it was soon shown (Hall and Weiderman, 1967) that some of the improvements came from the reranking of the already-retrieved documents, which would not reflect how a system would operate with real users. New methods of evaluation that avoided the "ranking effect" were developed (Chang *et al.*, 1969, 1971) and further experiments showed that this mode of request modification was helpful and preferable to the term-term association techniques, assuming there was some way of identifying relevant documents.

Eleanor Ide's investigations at Cornell (Ide, 1968, 1969, 1971) using Rocchio's algorithm drew the following conclusions.

1. In the experimental collection (the Cranfield 200) the Rocchio strategy is superior on 36% and equal on 32% of the queries (requests) that retrieve some but not all relevant documents on the first iteration.
2. Adding extra weighting to the original query terms made little difference in performance.
3. The first relevant document retrieved generally does not contain enough information for retrieval and documents retrieved soon after the first ($N = 5$) can add useful information ... a recommendation that N be set to some value that most users would consider reasonable, but that for some queries N should be raised until at least one relevant document is retrieved.
4. Because of the variability in negative feedback performance, feedback of non-relevant documents cannot be recommended as a general strategy.

In 1971 Salton published the SMART book, covering the majority of the SMART experiments at both Harvard and Cornell (Salton, 1971). There were sections on evaluation, language analysis, relevance feedback, and clustering. The book is particularly interesting since readers can trace the development of ideas over these years; most of the chapters come directly from the ISR reports. There are ten chapters in the general area of relevance feedback, and four for clustering, reflecting the then-current emphasis on these areas.

4.1.3 Document-Document Association, Clustering and the Cluster Hypothesis

Clustering was of particular interest to Salton because of its efficiency aspect; theoretically one could use clustering to scale up for searching huge numbers of documents by only searching selected clusters. By 1974 there were three different clustering algorithms implemented in SMART: CLUSTR (a modified Rocchio algorithm), DCLUSTR (based on Doyle's algorithms), and GROTRE. Of the five doctoral theses done by 1974 at Cornell, four had major clustering components.

However clustering (and the subsequent search) have a complex, interlocking set of pieces: first how to generate the clusters (size, density requirements, etc.); then how to characterize each cluster (the centroid); and finally how to search the clusters (hierarchically, best match, etc.). This would be ongoing work in the SMART project for another 10 years!

The group at Cambridge was also looking at clustering, both for efficiency and effectiveness (Jardine and van Rijsbergen, 1971).

> It is intuitively plausible that associations between documents convey information about the relevance of documents to requests. This hypothesis, which we call the *cluster hypothesis* can easily be checked on a particular document collection.

This was the document-document association factor noted in earlier work such as Maron and Kuhns, but never tested. The Jardine and van Rijsbergen paper also defined an effectiveness measure E (not the Swets version) that combines recall and precision but with a weighting factor that allows the relative importance of each to affect the final metric. This measure was used both to create benchmarks showing the maximum effectiveness that could be obtained for various clustering and searching scenarios, but also the actual performance obtained using a single-link clustering method on the Cranfield 200 collection (manual indexed version only).

The resulting clusters were hierarchical and the proposed search method was a simple downward search, i.e., take the best matching cluster starting at the top and continue downward. Three different methods of cluster representation were tried, with the best method weighting each term in the cluster representation by the number of documents in the cluster containing that given term. Further work (van Rijsbergen, 1973) used the same techniques on the ISILT/Keen and INSPEC collections, but found clustering was less effective on these collections (even using a broader type of search).

The cluster hypothesis could also predict when using term-term associations would be successful; term-term association experiments run on ISILT/Keen and INSPEC found no improvement (Spärck Jones, 1973a). In an effort to understand the reasons for this, Spärck Jones compared

many characteristics of the three collections (Cranfield 200, ISILT/Keen and INSPEC), and the paper is a tour-de-force in understanding some of the factors that differ across collections.

Spärck Jones found no obvious problems, but observed that the cluster hypothesis held for Cranfield but not ISILT/Keen and INSPEC.

> The associations are based on the initial term descriptions of the documents. Clearly if relevant and non-relevant documents are not well-separated by their descriptions, our chances of increasing the separation, using these descriptions, is poor.

Note that although the three collections were all built using the Cranfield method, the tasks being modeled were subtly different, and the collections are likely less homogeneous. The Cranfield 200 collection was the already-discussed manual indexed version. The ISILT/Keen collection was a new collection built by Michael Keen (Keen, 1973) to test five indexing languages when used with manual searching. There were 797 documents (mostly abstracts) from library science; and 63 real user requests with complete relevance judgements covering a very wide variety of library topics, such as mobile library routing, surveys on children's books, etc. The INSPEC collection (Aitchison and Tracy, 1970) was built by an operational SDI service from the Institute of Electrical Engineers to test several indexing languages: there were 542 documents/abstracts from electrical engineering and 97 requests based on users' SDI profiles (with partial relevance judgments).

4.1.4 Inverted Document Frequency Weighting (IDF)

In 1972 Spärck Jones published her paper on the IDF (inverted document frequency) weighting (Spärck Jones, 1972); its effect was immediate and profound. Her 2004 discussion (Spärck Jones, 2004) of what prompted this proposal illustrates some of the frustration with the then-current methods.

> My previous research had concentrated on automatic methods for constructing term classifications intended, by analogy

> with manual thesauri, as recall-promoting devices. ... Trying to understand what was happening in detail showed that terms that occurred in many documents dominated the classes. Thus anything that increased their matching potential, as term substitution did, would inevitably retrieve non-relevant rather more than relevant documents. However, these frequent terms were also common in requests, and simply removing them ... could have a damaging effect on performance. The natural implication was therefore that less frequent terms should be grouped but more frequent ones should be confined to singleton classes. This could give better performance than terms alone, but not for all test collections. What all this suggested was that it might be more profitable to concentrate on the frequency behaviour of terms, and forget about classes. More specifically, it led to the idea that all terms should be allowed to match but the value of matches on frequent terms should be lower than that for non-frequent terms.

The 1972 IDF proposal was based on the idea of specificity (and exhaustivity) as taken from a manual indexing point of view. Indexers could chose many descriptors for a given document (exhaustivity) but the nature of those descriptors (their specificity) in terms of the level of detail (e.g., beverage vs. tea) determined their discriminating power. Spärck Jones redefined these ideas for automatic indexing.

> We can thus re-define exhaustivity and specificity for simple term systems: the exhaustivity of a document description is the number of terms it contains, and the specificity of a term is the number of documents to which it pertains.

Heretofore term coordination matching had treated both frequent and non-frequent words the same; one way of changing this would be to have a weighting scheme that assigned more weight to non-frequent terms and less to frequent ones.

> The natural solution is to correlate a term's matching value with its collection frequency.[1] At this stage the division of terms into frequent and non-frequent is arbitrary and probably not optimal: the elegant and almost certainly better approach is to relate matching value more closely to relative frequency. The appropriate way of doing this is suggested by the term distribution curve for the vocabulary, which has the familiar Zipf shape. Let $f(n) = m$ such that $2^{m-1} < n \leq 2^m$. Then where there are N documents in the collection, the weight of a term which occurs n times is $f(N) - f(n) + 1$. For the Cranfield collection with 200 documents, for example, this means that a term occurring ninety times has weight 2, while one occurring three times has weight 7.

The paper continued with experiments on Cranfield 200, ISILT/Keen and INSPEC showing major improvements.

> The performance improvement obtained here nevertheless represents as good an improvement over simple unweighted keyword stem matching as has been obtained by any other means, including carefully constructed thesaurii.

4.1.5 Early Experiments using IDF and other Collection Frequency Measures

The publication of the IDF weighting triggered several papers investigating the issue of collection term frequency, in addition to new methods using this weighting scheme.

The 1973 SMART group paper (Salton and Yang, 1973) started with an in-depth analysis of term frequency distribution in three new collections. The Cranfield 424 collection was an automatically-indexed subset of the Cranfield 1400 collection, with 155 queries (requests). The Medlars 450 collection was an expanded version of the earlier Medlars collection, this time with 1033 abstracts and 30 queries. The TIME collection was a set of 425 full-text articles from TIME magazine, with

[1] The number of documents containing the term, not the total term frequency in the collection.

83 queries gathered from user questions. (For more on these collections, see (Harman, 2011).) Salton and Yang made a series of conjectures as to how the frequency of terms (within a collection) would affect retrieval.

1. Terms with very high frequency would not be useful.
2. Terms with medium total frequency and reasonably skewed distributions can help retrieve an adequate number of relevant documents, and also provide a high matching coefficient, and therefore good retrieval performance, for those items in which the term appears many times; it is likely that these may be powerful terms.
3. Terms with very skewed distributions, occurring in only a few documents, will produce matches for few documents, but matching items will exhibit a high query-document correlation, and stand a good chance of being judged relevant; such terms are likely to be useful, but not as important as terms of category (2).
4. Very rare terms should be considered to be of some importance, because a match between a query and a document, even though rare, will isolate a few documents from the bulk of the remaining ones; there is some indication that the elimination of rare terms leads to decreased retrieval effectiveness.
5. Terms with medium or low total frequency and flat distributions cannot be used to retrieve documents precisely; however, they do differentiate the small class of items in which they occur from the remainder.

They also introduced a new collection frequency[2] measure:

> The *discrimination value* of a term is a measure of the variation in the average pairwise document similarity which

[2]Here collection frequency is the total term frequency, not just the number of documents containing that term.

> occurs when the given term is assigned to a collection of documents. A good discriminator is one which when assigned as an index term will render the documents less similar to each other; that is, its assignment will decrease the average document pair similarity. Contrariwise, a poor discriminator increases the interdocument similarity.

(See the paper and its appendix for the mathematical description of the measure.) Both this measure and the IDF measure were used as "collection frequency" measures in the experiments.

Their experiments with the three collections explored different methods using these weights and the document frequency weights.

1. Terms with low values (low term frequencies, or high document frequencies, or low discrimination values) can be eliminated . . . as index terms for the collection . . . the corresponding process may be called CUT.
2. The calculated term values may be used as weights for example by multiplying any existing term weights by the new calculated values; this process may be termed MULT.
3. Finally, methods (a) and (b) can be combined by removing low value terms, and using the calculated values as term multipliers (CUT+MULT).

The results showed some variation across the three collections, but also fairly consistent performance (the discrimination value produced similar results to the IDF).

1. The use of term frequency weights as opposed to binary weights is nearly always justified; in particular, TF weighting is better almost everywhere for TIME with its reasonably large medium frequency vocabulary; it is also generally better for MED, and for CRAN in the low and medium recall range; only for the CRAN collection at very high recall are the binary weights superior.

2. The use of inverse document frequency weights is also justified almost everywhere; only for CRAN at very high recall are the standard binary weights better than the IDF runs.
3. Both term frequency as well as inverse document frequency weights are important, the results concerning the best procedure to be followed differ from collection to collection. In particular, the pure term deletion system (IDF CUT) is clearly best for MED; for CRAN and TIME on the other hand, the combined deletion and weighting system (IDF CUT+MULT) is preferred except at very high recall.

Spärck Jones also investigated various term weighting schemes, but with a different approach (Spärck Jones, 1973b). She looked at the three basic factors going into weighting: term occurrence in a collection (IDF), term occurrence in a document (TF), and the number of terms in a document (document length). These were combined into specific weighting types: Type 1 was document term frequency, Type 2 was document term frequency modified by document length and Type 3 was term collection frequency (IDF). A set of alternative mathematical formulas to express these factors is given in the paper.

The three weighting types were combined using what she termed "notational co-ordination", that is each term in the document was assigned a weight (as opposed to the normal binary weights of co-ordination matching), and these weights get added during the co-ordination match. "This means, for example, that a match on a single term with weight 8 is deemed a match on level 8." Experiments were run on the Keen abstract collection, and on three manually-indexed collections (Keen, INSPEC, Cranfield) and it was shown that the IDF weighting worked well on all three collections, but that the document term frequency and document length factors had less impact, even for the Keen Abstracts.

The surprisingly simple IDF measure developed by Spärck Jones has continued to dominate the term weighting metrics used in information retrieval, despite several efforts to develop more complex measures of term distribution. It has been incorporated in (probably) all information

retrieval systems, and used in languages other than English (even in languages without word boundaries). It has been used in summarization to locate the best sentences for extraction, and even in music and image retrieval.

4.2 Operational Systems Expand

Technology was moving ahead: some events in the late 1960s and 1970s that changed the retrieval world.

- The MARC record (MAchine-Readable Cataloging record) was developed at the Library of Congress in 1966.
- Starting in 1966 NSF awarded over 25 million dollars to various professional societies to develop computerized information retrieval systems in various disciplines.
- In October 1967 Lawrence G. Roberts published the first paper on the design of the ARPANET (Advanced Research Projects Agency Network); two years later BBN was funded to build a network between UCLA, Stanford, UCSB and U. of Utah.
- In 1969 Kenneth Thompson and Dennis Ritchie developed UNIX at Bell Labs, the first operating system designed to run on computers of all sizes, making open systems possible.
- Edgar F. Codd of IBM San Jose Research Lab, wrote his landmark paper in 1970 on his basic ideas for a relational database system (Codd, 1970).
- 1971 saw the first regular use of an 8 inch floppy disk for magnetic storage by Alan Shugart at IBM.
- Ray Tomlinson of BBN invented an email program to send messages across a distributed network in 1971; the @ symbol was first used in March, 1972 and by July Larry Roberts had developed a process for managing e-mail.
- *The Art of Computer Programming: Volume 3: Sorting and Searching*, was published in 1973 (Knuth, 1973).

- In 1973 Bob Metcalfe and David Boggs, researchers at Xerox Palo Alto Research Center, developed Ethernet.
- Vint Cerf and Bob Kahn published "A Protocol for Packet Network Interconnection" in 1974 (Cerf and Kahn, 1974) specifying in detail a Transmission Control Program (TCP).
- BBN opened Telenet, the first public packet data service (a commercial version of ARPANET) in 1974.
- Reporters and editors at the NY Times began using video terminals attached to electric typewriter keyboards and used facsimile transmission for printing simultaneously in London and Paris.
- The first Apple Computer (Apple I) (Figure 4.1) was demonstrated at the Homebrew Computer Club in Menlo Park, California in July 1976. There were approximately 200 of these sold at the price of $666. The Apple II became available in 1977 and was the first highly successful personal computer.

Figure 4.1: Apple I on display at the Smithsonian. Picture by Ed Uthman.

A critical component driving the expansion of the online systems was the growing amount of machine readable text. Governments and professional organizations were rapidly computerizing their bibliographic printing operations to speed the updating process (similar to the earlier work at NLM), with machine-readable databases on magnetic tapes as an important byproduct. The Chemical Abstract Service started in

1961 with a title, author and bibliographic reference database (CT) of 750 journals (68,400 articles). This was followed in 1968 with CA Condensates (CAC) which contained index terms (key-word phrases) in addition to the CT information. The COMPENDEX database from Engineering Index began in 1969, along with the Institute of Electrical Engineers's INSPEC in 1970 (Williams, 1985a).

Government organizations also had large bibliographic databases. The National Technical Information Service (NTIS), originally called the Publications Board, was created after World War II as the U.S. government repository for scientific research and information and had over 140,000 references in the early 1970s. There was also an agricultural bibliographic database CAIN, and the Department of Education's ERIC (Educational Resources Information Center) bibliographic database, which served as a decentralized subject-speciality clearinghouse for multiple journals, with 122,000 citations by the end of 1972.

4.2.1 Development of Current-Awareness (SDI) Systems

Several of these organizations were considering offering current-awareness services (SDIs) to their subscribers. This involved building "profiles" of a subscriber's research interests and running these profiles periodically against the updated magnetic tape databases. Matches would then be sent to the subscriber so that they could obtain the articles of interest to them.

The United Kingdom Chemical Information Service (UKCIS) was established by the U.K. Chemical Society to investigate the usefulness of providing an SDI service based on Chemical Abstracts. This group ran a series of experiments in the early 1970s to explore various issues for this type of service, such as how best to build these profiles, how well did the profiles work, etc. (Barker *et al.*, 1972b). Some details of these experiments are provided here to illustrate how these early SDI systems operated. Equally important, these details show the tight interconnection between commercial research and its business operation.

There were 250 participants in the UKCIS experiment, drawn from industry, government and universities. These participants submitted a free-text description of their needs and these were converted into

profiles by UKCIS staff. Profiles contained terms and phrases (with full truncation available), along with ANDs, ORs, and NOTs; they averaged 20 to 30 terms per profile. These profiles were searched biweekly against two database tapes, Chemical Titles (CT) and the Chemical-Biological Activities (CBAC) which contained full digests, together with molecular formulae and CAS registry numbers of specific compounds mentioned in the text. This experiment ran for 10 months, with many different aspects examined; here are some of the result highlights:

- A detailed analysis of 55 profiles showed that 22% were less specific than the user's statement and 27% were more specific.
- user assessment of the returned documents showed an average precision of 40% for CT and 29% for CBAC, with wide variations across profiles.
- A sample of 47 profiles for CBAC was used to calculate relative recall; manual searching found 80% of the documents, whereas machine searching of the profiles found 65%.
- Failure analysis showed that over 50% of the missing documents were caused by inadequate concept expansion or too narrow a profile.

They also looked at how the different types of data provided by the four major magnetic tape databases produced by Chemical Abstracts affected profile performance (Barker *et al.*, 1972a). In addition to the CT database and the CBAC database, there was a Chemical Abstracts Condensates (CAC), and Polymer Science and Technology (POST) similar to the CBAC database. The CT database had the least expensive subscription and was the most current, covering 130,000 citations annually; the CAC was double the price, and was published about 8 weeks after the CT version, but it contained 300,000 citations annually. The last two databases covered specialized areas, with only 13,000 and 21,000 citations respectively. The goal here was understand how the various levels of available content of these databases affected the matching performance of the profiles.

There were a total of 193 SDI profiles processed, with 48 of these used for the full testing of the two specialized databases. There were five variations of each profile written to target specific databases, but with two variations using titles only to mimic use of the CT database.

```
a CAC-1 or CAC-2, titles-only
b CAC-1 or CAC-2, keywords-only
c CAC-1 or CAC-2, titles-plus-keywords
d CBAC or POST, titles-only
e CBAC or POST, titles-plus-digests.
```

The participants were then asked to judge relevancy, with precision and relative recall being used as metrics. The results showed that using a digest in addition to the titles on the CBAC or POST got 68% better recall than using the titles alone, with a loss of 23% precision. For the CAC, the use of keywords in addition to titles increased recall by 35%, with a 10% fall in precision. So clearly the additional content helped, but came at a higher cost both in fees and in lack of currency. Note that this experiment is the source of the UKCIS test collection, including the 193 profiles (and 48 profile subset), along with 27,361 titles from the various databases.

A similar experiment for SDI profiles (with similar results) was run by the Institute of Electrical Engineers (INSPEC) to test several indexing languages in terms of performance in matching profiles (Aitchison and Tracy, 1970). There were 97 profiles used for testing against the abstracts and titles of articles in physics and electrical engineering. Four different indexing methods were used: titles only, titles and abstracts, terms from the printed subject index to Science Abstracts, and a controlled language using a thesaurus (this is the experiment that led to the INSPEC collection mentioned in the last section).

Vaswani and Cameron of the National Physical Laboratory ran experiments in word association with the goal of improving recall (Vaswani and Cameron, 1970). Although the original report (like the INSPEC one) is difficult to find, a report on the available test collections (Spärck Jones and van Rijsbergen, 1976) provides details of the creation of the NPL collection. It contained 93 queries (user based), and 11,571

abstracts in electronics, computers, physics and geophysics. There was a basic dictionary of 1000 index terms (stems) used to index the collection and various word association methods were tried using co-occurance of these words in the abstracts. The abstract of the original report summarized their results:

> The best strategy depends on the user's requirements. For a single strategy key-words are simplest but the quantities of output are erratic and may usefully be controlled according to word associations. If two strategies can be used key-words alone may be followed by associations, yielding in a similar output quantity 30% more relevant documents. The corresponding use of clusters is marginally better but unlikely to justify its extra cost.

4.2.2 Development of Online Retrospective Search

Note that these experiments were for SDI services rather than providing the ability to do retrospective searching within the databases. NLM's MEDLARS initiated an experimental service in June of 1970 for online access to their databases (as opposed to the batch processing traditionally used) (Bourne and Hahn, 2003; McCarn, 1971). The service was called AIM-TWX, with AIM standing for Abridged Index Medicus and TWX for the AT&T TeletypeWriter Exchange Network, consisting of extremely slow teletype terminals connected via telephone lines. The abridged part meant there were only 130,000 citations from the 100 most important journals from a 5-year period.

Users (over 50 medical schools, hospitals, etc.) called in over dedicated phone lines, keyed in a password and then typed in a search request. The computer fed back the number of documents associated with each input term and allowed the user at that point to combine terms with ORs, ANDs, or NOTs. The search was then run and the list of matching documents could be printed out. The indexed documents were stored on a time-sharing computer (IBM 360/67) run by System Development Corporation, who used a version of their ORBIT system (called ELHILL) for the searching. The service operated for four hours, five days a week; users paid for the terminals and for the TWX toll

calls (20 to 70 cents per minute, with the average search taking 15 to 20 minutes).

F. W. Lancaster was asked to do an evaluation of the pilot system (Lancaster *et al.*, 1972). He looked at 47 searches completed by physicians on 8000 epilepsy abstracts, with 16 users at six separate centers.

> On the whole I believe the results to be surprisingly good. Although a few users went badly astray, the majority were able to conduct productive searches. The precision achieved in most cases was high and the cost in time appears to be well within tolerable limits. It is not to be expected that the requester himself will perform as well as would a trained analyst. . . . Nevertheless, acceptable results were obtained in most cases by a simple and straightforward approach. It is noteworthy that many of these users were seeking only a few relevant references.

Note that this was probably the first ever evaluation of an online search service and Lancaster went on to make several recommendations including the following.

> Ultimately, however, on-line retrieval systems for use by people who are not information specialists should be designed to: (1) allow input of natural-language requests, and (2) avoid the necessity for inputting requests as formal search statements with Boolean operators. The advantages of the natural-language approach include the obvious one that the user does not need to learn the terms of some restricted vocabulary and the less obvious one that natural-language statements tend to represent true information needs better than statements that have been influenced by the logical and linguistic constraints of a system. . . . The use of Boolean algebra for querying computer-based retrieval systems may have been a mistake. The mistake arose through the way that early computer-based systems developed. That is, they were viewed as more mechanized versions of semi-manual

systems such as those employing edge-notched cards or the optical coincidence principle.

NLM discontinued the AIM-TWX service in April of 1972, expanding into its own MEDLINE service, with a user group of ninety-two institutions and a database of 450,000 references. The communications were over TYMNET as opposed to the slow TWX terminals (and there were no toll charges).

LEADERMART, part of an ongoing project at Lehigh University sponsored by NSF, became operational in September of 1971 with eight user groups; by 1972 it was available to 18 campus groups, along with other universities in nearby Pennsylvania and a telecommunications link connection to the University of Georgia (Kasarda and Hillman, 1972). The half-million records maintained online included COMPENDEX (Engineering Index Abstracts), CAS Condensates (Chemical Abstracts Service), ASCE Journal Abstracts (Civil Engineering), Tall Structures Abstracts (Civil Engineering), and CIS Documents (Information Science, fulltext). Note that this service was also available commercially at a charge of $55 per connect hour (after the NSF funding ran out), however it was soon suspended by Lehigh University as being too costly to operate (Bourne and Hahn, 2003).

Battelle's BASIS system was also available in 1971, and by 1973 had over a million records online for searching. The records included the full CAC file (from Chem Abstracts), and the complete NTIS database with 140,000 records. Using TYMNET they were able to service sixty users simultaneously and by March of 1974 there were over 900 users, mostly including government subscribers such as EPA, NBS, and DoD. However there were also commercial subscribers such as Calspan and Exxon. By the end of 1974 the service was mostly shut down, with subscribers transferred to other commercial online systems (Bourne and Hahn, 2003). Note that Battelle had the common problem of the non-profits (and government) of being seen to compete with commercial services and this issue continued to haunt the research/government systems throughout the 1980s and 1990s.

When NLM brought up their own MEDLINE service in 1972, SDS decided to commercialize their ORBIT online service using the ERIC

database. The Chemical Abstracts Condensates was added in early 1973 and by the end of the year over 50 organizations were using the service, with five databases (MEDLINE, CHEMCON, ABI/INFORM (business), ERIC, and CAIN (agriculture) including 2 million records. There were over 20 significant databases available in 1976 (Bourne and Hahn, 2003).

The Lockheed commercial online service started in 1972 using their DIALOG search system (the online service itself was renamed DIALOG in 1976). They offered the ERIC database and the NTIS database, and had 37 customers by summer of 1973, with eight databases. By the end of 1976 there was access to 50 databases and more than 12 million records. Additionally the DIALOG management (Roger Summit was the driving force) wanted to expand to serve users in libraries and in 1974 they were awarded a 2-year contract from NSF for experimental systems to be set up in four public libraries (Bourne and Hahn, 2003).

McCarn's 1978 survey of the online systems (McCarn, 1978) shows the explosive growth in both bibliographic databases and searching; in 1976 there were 33 million bibliographic references available, with over 1 million online searches. Just one year later, there were 50 million references and over 2 million searches!

These databases (and search services) were mostly bibliographic data; there was little full text available for searching. This was largely an outgrowth of the limited interests of the abstracting industry, but it was also a huge undertaking in terms of creating large datasets of full text. However the legal community had requirements for using full text, and the financial resources to create it. In 1973 the OBAR system (which had started inputting (manual keypunching) legal records for the Ohio Bar Association) was renamed LEXIS and launched with the goal of creating a database of all state laws and U.S. Federal law. By the end of 1976 there were full text legal records for six states and some of the Federal government. The system was set up for end users (lawyers or clerks) rather than librarians and by 1977 they could brag about being the first online service to make a profit.

4.3 Research Re-enters a Theory-Building Phase

Stephen Robertson in his 1977 survey of theories and models in information retrieval (Robertson, 1977b) noted that since the Maron and Kuhns paper (Maron and Kuhns, 1960) there had been few theory papers.

> It seems that after the very empirical approach to information retrieval initiated by the Cranfield experiments in the sixties, we are re-entering a theory-building phase. But the theories we are building today look very different from those in the earlier tradition. Whether or not one actually conducts a retrieval-performance experiment, the very *idea* of doing so has had a profound influence on the theorists.

Indeed, most of the core information retrieval papers after the mid-1970s were either investigating theories for older empirical approaches, or basing new approaches on more theoretical grounds. It isn't clear why this was happening; certainly a major effect was the 1972 Spärck Jones paper on IDF, which spurred interest in how term frequencies could be explored. Additionally, however, researchers had become convinced that their attempts to expand requests had less impact than finding better ways of using the naturally-occurring terms in the documents, either in the index stage or the search stage.

4.3.1 Investigation of Term Frequency Properties

Bookstein and Swanson (1974) proposed the use of a 2-Poisson model for characterizing term frequencies, based on their appearances in "useful" vs. "nonuseful" documents. Stephen Harter's follow-up work (Harter, 1975a,b) examined the distribution of what he called "specialty" words in 650 abstracts of the works of Sigmund Freud. He hypothesized that non-specialty words would have a random distribution, described by a Poisson density function. Speciality words, however, would have two different frequency distributions because they are central subjects in some documents but only appear peripherally in others, leading to a 2-Poisson distribution. This model could be used for term selection in automatic indexing. Harter went on to develop methods for doing this

and found reasonable success in predicting which words would appear in a manually-created index.

Salton *et al.* (1974, 1975) also investigated term frequency properties using their term discrimination value described earlier (Salton and Yang, 1973). For each of their three collections (Cranfield, MEDLARS, and TIME), they ranked the terms by their term discrimination value and plotted these ranks against the frequency of that term in the collection, resulting in a U-shape with the low frequency terms having high ranks, along with the high frequency terms. The middle frequency terms have low ranks, i.e., high term discrimination values.

The issue was how to effectively use the low and high frequency terms, while still maintaining good performance. Their indexing strategy proposal was then as follows:

1. ‘‘terms whose document frequency lies between n/100 and n/10 (where n is the number of documents in the collection) should be used for indexing purposes directly without any transformation’’;
2. high frequency terms should be turned into indexing phrases;
3. low frequency terms should be expanded into groups via a thesaurus.

Experiments were done using this idea. The phrases were based on phrases found in the query texts, i.e. candidate phrases were selected this way and then verified in the collection, and the manual thesaurii for the three collections were used for the low frequency terms. The results for the three collections were very impressive, however there was no document collection frequency weighting (IDF or term discrimination weight) used so it is difficult to compare this with the results of their 1973 paper.

4.3.2 Investigation of Term Frequency Properties based on Relevancy

In an effort to find better weights for the request terms, Spärck Jones experimented with finding "optimal" performance for a test collection (Spärck Jones, 1975). The concept was to use term information from the known relevant documents, i.e., weight the request terms based on their frequency in the set of relevant documents for that request. Something similar had been done by Miller in 1971 when he was investigating a probabilistic search strategy for MEDLARS (Miller, 1971). Miller had weighted the request terms based on the user's estimate of the frequency of that term in relevant references. Spärck Jones instead used the actual frequencies of terms in the relevant documents, combined with the IDF type of frequency weighting

$$w_i = \log N - \log n_i + \log r_i - \log R$$

where N is the number of documents in the collection, n_i is the frequency of term i, R is the number of relevant documents for the request and r_i is the frequency of the term i in them.

This indeed worked very well, with the recall/precision graphs (not shown) measuring a large improvement over using IDF alone. However when this was tried in a predictive manner, i.e. use half the collection to find the term weights which were then used on the other half, it showed little improvement over IDF.

This idea was expanded into a theoretical (and experimental) investigation of the various methods of using relevance information (Robertson and Spärck Jones, 1976). The general notion was to examine the interaction between request term occurrence in a collection and request term occurrence in relevant documents. Table 4.1 shows that interaction, with N being the number of documents in the collection, R the number of relevant documents for a given request, n the number of documents in the collection indexed by term t, and r the number of relevant documents for a given request that are indexed by t.

Table 4.1: Request term occurrence in relevant and non-relevant documents

	Relevant	Non-relevant	
Indexed	r	$n - r$	n
Not indexed	$R - r$	$N - n - R + r$	$N - n$
	R	$N - R$	N

This leads to four possible formulas for weighting:

$$w^1 = \log \frac{\frac{r}{R}}{\frac{n}{N}} \tag{F1}$$

$$w^2 = \log \frac{\frac{r}{R}}{\frac{n-r}{N-R}} \tag{F2}$$

$$w^3 = \log \frac{\frac{r}{R-r}}{\frac{n}{N-n}} \tag{F3}$$

$$w^4 = \log \frac{\frac{r}{R-r}}{\frac{n-r}{N-n-R+r}}. \tag{F4}$$

These four formula were informally defined as follows for a given request term t.

> Function F1 represents the ratio of the proportion of relevant documents in which t occurs to the proportion of the entire collection in which it occurs, while F2 represents the ratio of the proportion of relevant documents to that of non-relevant documents. F3 represents the ratio between the "relevance-odds" for term t (i.e. the ratio between the number of relevant documents in which it does occur and the number in which it does not occur) and the "collection odds" for t, which F4 represents the ratio between the term's relevance odds and its "non-relevance odds".

Tests of all four formulas were tried out on the manually indexed Cranfield 200 collection. The graphs in the paper show the order of performance, with F4 somewhat the best (at higher precision), and F1 the "worst", but all considerably better than simple IDF. However when

used predictively (trained on half the collection and tested on the other half), the results had the same order but were much poorer.

4.3.3 The Probabilistic Theory of Relevance Weighting

The appendix of this same paper (Robertson and Spärck Jones, 1976) presents the first formal version of the probabilistic theory of relevance weighting. This was elaborated in a survey paper the following year (Robertson, 1977b) which provides an intuitive version of this theory, with a follow-on formal version later that year (Robertson, 1977a).

The 1977 survey paper covered many of the retrieval models that had been proposed up to that time. However it also looked at the retrieval problem from the user's point of view, i.e. what model would best describe the optimal performance for a user. Several hypotheses were made based on this viewpoint.

> Document ordering hypothesis: For optimum performance, the systems should order the documents and allow the searcher to search down the ordered list as far as s/he wants to go... The hypothesis arises from the essentially probabilistic nature of information retrieval. Given that the system cannot predict with certainty which documents the user will find relevant or useful, it also cannot predict how many documents the user will need to see in order to satisfy his problem or information need.

This hypothesis then leads to an informal statement of the probability ranking hypothesis.

> Probability ranking hypothesis: For optimum performance, the system should rank the documents according to their probability of being judged relevant or useful to the user's problem or information need. There is an immediate and significant corollary. If the system uses an explicit (numerical) match function to rank the documents, then both the formulation of the match function itself, and the derivation

> of any of its components combine to specify the order. It follows that the design and choice of a particular match function is not independent of the design and choice of components such as index term weights, request term weights, etc. All the parts are subordinate to the probability ranking hypothesis, and the appropriate combination of parts has to be chosen in accordance with the hypothesis. The common approach of considering term weighting separately from the match function is not valid (or at least suboptimal).

One of the assumptions for all of the term weighting methods is that the occurrence of the terms is independent and weights can therefore be assigned independently. van Rijsbergen (1977) presented a theoretical basis for using index terms without the independence assumption and developed methodology to estimate term dependency and incorporate it within a weighting scheme. This was closely followed by a paper by Harper and van Rijsbergen (1978) using these ideas in both weighting and relevance feedback. Harper used the Cranfield 1400 (manually indexed) and showed some improvement using a slightly expanded query based on co-occurrence data, with more improvement when this was also used for relevance feedback.

4.3.4 Further Experiments with Relevance Weighting

Spärck Jones continued her work with the relevance weighting formulas developed in 1976, this time working with the UKCIS collection to see if the same results were found in a different collection, and importantly an operational service collection (Spärck Jones, 1979a). She looked at the scale effect (the results for F4 did indeed scale), and tried using a smaller part of the collection (one-quarter) for training, finding that it worked almost as well as using half. This led to a second experiment investigating just how little relevance information was needed (Spärck Jones, 1979b). She found that using only two relevant documents (retrieved by a simple term search) worked almost as well as using half the collection for the Cranfield 1400 collection (manually-indexed version); but that the recall was significantly lower for the UKCIS collection (showing again the importance of using multiple collections).

4.3.5 The Information Retrieval Community Expands

The late 1970s also saw efforts to expand the information retrieval research community and demonstrate its relevance to the commercial search world. In 1975 Keith van Rijsbergen published his textbook *Information Retrieval* (van Rijsbergen, 1975). Whereas Salton's 1968 book was very much a "how-to" book on the various aspects of building an information retrieval system (based on his SMART project), this book covered much of the earlier theory, updating it to current work (and theory) in information retrieval. The various chapters included automatic text analysis, automatic classification, file structures, search strategies and evaluation and is a well-balanced view of the theory and methodology in retrieval as of 1975, along with a very complete bibliography.

There was a proposal in 1975 by Spärck Jones and van Rijsbergen (Spärck Jones, 1975; Spärck Jones and van Rijsbergen, 1976) for the creation of large test collection(s). The proposal had two overarching criteria: first that this would allow for a commonality of testing across retrieval researchers, and second that it would be adequate for many various projects. The proposal called for a set of 30,000 documents broadly representative of "service" data bases in size and subject composition, including all bibliographic data, abstracts, citations, natural language indexing, and controlled language indexing. There were also to be 700 to 1000 requests, each including a verbal need statement, lists of free and controlled terms, and a Boolean specification. And finally the relevance judgments, with two relevance grades and a novelty indication, including judgments by different people. Unfortunately no funding was available and nothing was built.

The first annual[3] ACM SIGIR (Special Interest Group for Information Retrieval) conference was held in May of 1978 in Rochester, N.Y. and chaired by Robert Dattola, with 14 papers. The second SIGIR was chaired by Bob Korfhage in Dallas, Texas, again with 14 papers and one panel. In 1980 the conference moved to Cambridge, U.K., was chaired by Keith van Rijsbergen and had expanded to 23 papers.

[3]There was one earlier meeting in 1971.

4.3.6 Early Prototypes based on Information Retrieval Research

One of the papers presented in 1978 was on the SIRE system from Syracuse University (McGill *et al.*, 1976). This system allowed both traditional Boolean retrieval and retrieval using a system based on SMART. This system was one of five prototypes presented at the first SIGIR conference.

As researchers became more confident in their retrieval methods, other prototypes were being built. R. N. Oddy's THOMAS system (Oddy, 1977) focussed on browsing, i.e., users having a general interest in a subject and wanting to explore related literature.

> Interaction with an on-line system has two effects upon a user which should be distinguished. Firstly, the user learns how to express his need more effectively to the system... Secondly, the user's notion of what he requires becomes clear to himself, and may shift in emphasis.

The paper described the THOMAS system, built to explore MEDLARS data at the University of Aston in Birmingham, and based on having a dialog with a user. The user experiment to compare THOMAS to MEDUSA (the traditional Boolean access) found that "THOMAS showed the user about as many references as MEDUSA but the demand on user effort was much less than that demanded by MEDUSA".

> The art of information system design (which, I am certain, has a long future) is to find the form and timing of information presentation which will best aid the system user in whatever task he has in hand (Oddy, 1977).

Tamas Doszkocs also built a prototype based on MEDLARS (Doszkocs, 1982; Doszkocs and Rapp, 1979). His CITE system started with a user's natural language query and provided ranked output, relevance feedback and other query expansion methods. It was deployed in a serious user preference test within the operational online catalog system (see the next chapter for details).

5

Now What (1980s)?

5.1 Research Builds on the 1970s

Computers and related technologies were now exploding in terms of capability, and prices were coming down steeply.

- Hard disk prices went from $200 per MB in 1980 to $45 per MB in 1985 to $5 per MB in 1990.
- The IBM 3380 was the world's first gigabyte-capacity disk drive. The unit had nine 14 inch platters and weighed more than 50 lbs. Depending on the version, it cost around $50,000 in 1980.
- In 1980 Seagate released the first 5.25-inch hard disk holding 5 MB.
- The first successfully marketed IBM personal computer, the IBM PC was available in 1981.
- The CD-ROM format was developed by the Japanese company Denon in 1982. It was an extension of Compact Disc Digital Audio, and adapted the format to hold any form of digital data, with a

storage capacity of 553 MB. CD-ROM was then introduced by Denon and Sony at a Japanese computer show in 1984.

The early 1980s saw the publication of two more books on information retrieval. Karen Spärck Jones (along with other authors) published *Information Retrieval Experiment* (Spärck Jones, 1981), a book explaining how to design experiments, including chapters by Stephen Robertson on methodology, Jean Tague on the pragmatics of experimentation, and F. W. Lancaster on evaluation in an operational environment. Part 3 of the book is a summary of the results of experiments, going back to the late 1950s. Spärck Jones wrote an over 70-page historical review in this part that could be considered the first "history of information retrieval"!

The book by Salton and McGill (Salton and McGill, 1983) presented detailed chapters on various retrieval processes, such as text analysis and indexing, evaluation, feedback, etc. Unlike earlier Salton books, however, there was discussion of areas outside of the SMART environment, such as commercial inverted file systems (like DIALOG and LEXIS) and also natural language processing and data management systems.

5.1.1 Continuing Research using the Probabilistic Models

Much of the information retrieval research in the 1980s was a continuation/elaboration of work done in the 1970s, such as combining earlier theories and methodologies to explore new possibilities, or working in smaller specialized areas to better understand the issues. There was consideration of users, both theoretically and in user studies.

The 1980 paper out of the Cambridge project (Robertson *et al.*, 1980) is an example of the first type of research. The goal here was to explore ways of incorporating the probabilistic work on searching by Robertson and Spärck Jones with the work done by Harter on the 2-Poisson method for selecting index terms. "In particular, we hope to develop and test a model, within the framework of the probabilistic theory of document retrieval, which makes optimum use of within-document frequencies in searching." Additionally they wanted to explore the term dependency investigations by van Rijsbergen and Harper including ideas on query expansion and relevance feedback. The 1980 paper started

with a summary, both of the theoretical work and the results of these earlier experiments. There was then a deeper look at the Harter ideas in terms of how they could be converted to a weighting scheme (the mathematical derivation of specific formulas are in the paper). The NPL collection (see Section 4.2.1) was used for a series of experiments with various trial weighting schemes exploring the Harter ideas, but results were disappointing.

Another example of the first type of work is Robertson, Maron and Cooper's 1982 paper (Robertson, 2003; Robertson *et al.*, 1982). Here the aim was to unify the much older Maron/Kuhns probabilistic model (called Model 1) with the newer Robertson and Spärck Jones probabilistic one (Model 2).

> The two earlier models (Models 1 and 2 respectively) dealt with situations in which the system possesses data about the individual document in relation to a class of queries, or about the individual query in relation to a class of documents. (If there is data about the individual query in relation to the individual document, then no retrieval system is necessary.) Suppose, therefore, that we have both kinds of information. What kind of model do we need to take account of all the information we have in assessing probability of relevance?

W. Bruce Croft (Croft, 1981) also looked at how to combine within-document frequencies with collection frequencies within a probabilistic framework. The first part of his paper discussed various term weighting schemes based on term distributions within documents and noted that "what these weights do have in common is that ... they include a measure of the significance of a term in a particular document based on its frequency of occurrence. We shall therefore refer to these weights as *term significance* weights." From the viewpoint of indexing, "a document representative should be regarded as a set of term significance weights which can be used to decide whether the associated terms should be assigned", or that "significance weights are interpreted as being estimates of the probability" that a term is indexed.

A second paper (Croft, 1983) gave details on a large set of experiments using both the Cranfield 1400 abstracts and the NPL collection.

Note that these are very different collections, both in document size (Cranfield has over 150 words per document versus less than 50 for NPL), and in the range of term frequencies within a given document. Table 1 in the paper shows the wider range of frequencies in Cranfield and, more importantly, the much higher percentage of term frequencies that are greater than one.

Croft provided three different baselines: coordination match only; cosine correlation using within-document frequency only; and using IDF for weighting query terms. Even the baseline performances across the two collections were different. For Cranfield the cosine correlation performed as well as the IDF query term weighting, but for NPL the cosine correlation did not even beat the coordination match.

The next stage was trying a combination match using a constant times the number of matches plus term weighting using IDF. Here there was little difference between the collections and little difference between the combination match and using IDF weighting only.

Croft then introduced the term significance weighting, multiplying the combination match by the probability that term i is assigned to document d ($P(x_i|d)$). This probability was estimated from the within-document frequencies, but with a normalization factor. Specifically

> If w_{di} is the within-document frequency of term i in document d, then n_{di}, which is the normalized within-document frequency, is calculated as $w_{di}/\max(w_{d1}, w_{d2}, \ldots)$.
>
> $P(x_i|d)$ is then estimated as $K + (1 - K)n_{di}$, where K is a constant between 0 and 1.

The constant K was added here to adjust the normalization based on the collection (similar to the constant C in the combination match). Various levels of K were tried with both collections; for a K of 0 (no normalization) performance was significantly higher for Cranfield than just the combination match alone, but significantly worse for the NPL collection. The optimal value for K was 0.3 for Cranfield and 0.5 for NPL. Similar results carried over into the feedback experiments in the paper.

These experiments not only provided a way of incorporating within-document frequency in the probabilistic model, but showed once again the importance of using multiple test collections, and understanding (and accounting for) their different characteristics.

5.1.2 Research in Specialized Areas

Work was still ongoing at the Cornell SMART project (Lesk *et al.*, 1997). Ed Fox re-implemented the original Cornell version of SMART for operation using the UNIX system in the early 1980s; Chris Buckley added efficiency improvements (Buckley and Lewit, 1985), and there were other students working in mostly specialized areas.

Fox worked on extended Boolean retrieval (Salton *et al.*, 1983) and built the CACM collection as part of his project (Fox, 1983). It contained metadata from all articles in issues of the CACM (Communications of the ACM) from 1958 to 1979, including titles and some abstracts for use in automatic indexing, the authors, the computing reviews categories, the citations, the co-citations, and the links (references to or citations between articles). The queries for this collection were true user queries, gathered from faculty, staff and students from various U.S. computer science departments, who made the relevance judgments.

Ellen Voorhees re-examined the cluster hypothesis, as well as re-visiting the effectiveness of cluster searching (Voorhees, 1985). She investigated a new method for determining if the cluster hypothesis held for a given collection, with results from four collections showing some differences based on the new test. Using single-link clustering, she ran a series of retrieval experiments using clustering on these collections and concluded that a straight sequential search was more effective, even when the cluster hypothesis applied.

Joel Fagan looked at multiple ways of defining and using non-syntactic phrases as concepts. The use of these phrases did not significantly improve performance over a basic tf*idf run and Fagan's detailed analysis provided a good insight into some of the reasons. For example "the meaning of the source text phrase descriptor in a query may differ significantly from the meaning of the source text of a phrase descriptor in a document."

At Cambridge, Martin Porter developed a new stemming algorithm (Porter, 1980). This algorithm was much simpler than the earlier Lovins stemmer (Lovins, 1968), which used auxiliary files containing a list of over 260 possible suffixes, a large exception list, and a cleanup rule. The Porter algorithm looked for about 60 suffixes, producing word variant conflation intermediate between a simple singular-plural technique (an S stemmer) and Lovins algorithm. It should be noted that the Porter stemmer is still heavily used for English and stemmers for many other languages have been developed (many by Porter) using the same simple techniques (Willett, 2006).

5.1.3 Research with Users

None of this effort took much account of actual users; the emphasis was on the effectiveness of system methods as measured by test collections. Nicholas Belkin's paper on "Anomalous States of Knowledge as a Basis for Information Retrieval" (Belkin, 1980) emphasized thinking about real users and what they might actually be trying to do (as opposed to just document retrieval). The idea behind ASK (Anomalous States of Knowledge) was that a user is trying to fill in a gap in their knowledge, but this gap might be difficult to specify exactly. This has to do both with the type of knowledge that is being sought (i.e., simple fact versus complex background knowledge) and with the ease of expressing that need to a retrieval system. "We may at least speculate that ASKs will fall into classes which require different sorts of answers and therefore different retrieval strategies, each designed to retrieve texts appropriate to the class of anomaly."

He and Bob Oddy did a design study (Belkin *et al.*, 1982) to test this model by tape recording interviews with 35 users at 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 the paper listed five

different ASK types based on these interviews, and the categories 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.

Note that there were real users now; the operational systems (see next section) had created a new type of users called search intermediaries, whose job was to learn to effectively search online services. Whereas there had been work much earlier in how to best index large bibliographic databases, there had been little work in how to best search them. The group from Syracuse University, School of Information Studies, investigated searching the various types of available document representations for two commercial datasets.

Katzer *et al.* (1982) used 12,000 documents from INSPEC, along with 84 real user queries (the INSPEC test collection later used by the SMART group) to compare performance using seven different document representations: title, abstract, descriptors (index terms from a controlled vocabulary) and identifiers (freely chosen index terms), plus their combinations. Trained professional search intermediaries did the searching and found some significant differences in recall (free-text did better than controlled vocabulary) but none in precision. More importantly, the various document representations found different relevant documents; the overlap in the retrieved relevant sets was small.

This experiment was repeated (Das-Gupta and Katzer, 1983) on the PsychInfo 1980 database, with 12,000 abstracts (and indexing information). There was relatively little difference in performance and again little pairwise overlap in the relevant documents (between 23% and 27%).

5.2 Operational Systems: Online Services Making Big Bucks

The research prototypes, such as the SIRE system (McGill *et al.*, 1976) were now being implemented in small-scale commercial operations. An example of this was the MASQUERADE in-house system (Brzozowski, 1983) at Marathon Oil based on the SIRE system. This system was capable of doing Boolean or ranked output (cosine with within-document frequency and IDF) and was able to handle 20 years of internal technical reports along with research library collections. The probabilistic retrieval systems were represented by the prototype CUPID (Cambridge University Probabilistic Independence Datamodel) (Porter, 1982) that could index and search effectively over 10,000 documents.

However the major commercial systems were continuations of the mostly Boolean systems from the 1970s, only now they were massively expanded and profitable. The number of available electronic databases went from 500 in 1980, to 3000 by 1985 and to over 6800 by 1990.

There were four main types of electronic databases: online catalogues of library holdings, online equivalents of the print abstracting/indexing services such as COMPENDEX and INFORM, full-text databases and numeric databases.

The online catalogues were huge unions of the catalogs of many libraries, with several large vendors such as OCLC (now called Online Computer Library Center) and Research Libraries Information Network (RLIN). Libraries tended to subscribe to only one and that one was picked based on the type of searches allowed such as truncation, subject search and year of publication.

By far the largest number of electronic databases were online equivalents of abstracting/indexing services. To give some idea of the size and scope of these databases, the BioSciences Information Service's BIOSIS Previews database (from printed Biological Abstracts and BioResearch Index) had 4,264,000 references by end of 1984. Engineering Information's COMPENDEX had 1,685,669 references by 1984. There were multidisciplinary databases such as SCISearch from the Institute of Scientific Information that included references from over 4000 journals (Williams, 1985a).

In terms of full text databases, the legal area had LEXIS and WESTLAW and there were several news sources, such as NEXIS which included not only newspapers (New York Times, Washington Post, Christian Science Monitor, etc.) but also magazines such as Business Week, Byte, and Dun's Business Month. The newspapers were updated within 48 hours of publication, and there were wire service stories (Tenopir, 1984).

Some of the vendors also offered full text journals and reference books to compliment their bibliographic files. Note that the availability of full text required large amounts of additional storage and also a reasonable baud rate for access from subscribers, so the decision to offer full text was based on marketability.

The numeric databases were often a part of larger datasets such as those from Chem Abstracts or the Toxicology database from the National Library of Medicine (NLM). However the usage of these was small compared to the business numeric databases, including commodities quotation systems, with less than 30,000 connect hours versus tens of millions of connect hours (Williams, 1985a).

Databases were mostly provided through third-party vendors. Whereas LEXIS and NEXIS were both built by and provided by Mead-DataCentral, most of the other databases were built either as part of a bibliographic/abstracting operation, or as an independent database, and then marketed via a vendor such as Dialog or BRS (Bibliographic Retrieval Services). These vendors paid royalties and use fees (aggressively negotiated) to the database producers and the competition between vendors was intense.

Martha Williams (Williams, 1985b) annually tracked the usage of databases using a sampled survey of 500 organizations. Here are two items from her 1983 survey and one from two years later.

- In 1982, the size of the US Information Center/Library market for online usage of word-oriented databases was 1.25 million hours and 1.59 million (extrapolated) hours in 1983. 1.59 million versus 1.25 million hours is an increase for the year of 27%. That is the growth in usage for this market. The market growth in the more usual terms, i.e. dollars, increased from $125 million

to $159 million ($509 million in 2018 dollars). Another fact to consider is that one producer alone, MDC (Mead Data Central), accounted for one third of the revenue. MDC and Dialog together accounted for 69% of the use and 82% of the revenues in 1982.

- Dialog clearly had the greatest breadth of user clientele. Dialog achieved first place in revenue among seven out of the eight user classes. It also achieved five firsts for usage. The other first place rankings were MDC with firsts in the legal class for both use and revenue. NLM was first for use in the medical class as might be expected, but it was only second for revenue being displaced by Dialog which generated the most revenue from the medical community. In 1982, and 1983, BRS was first for usage and second for dollars in the Academic community.
- In the first quarter of 1985 the average expenditure per hour for the total US online information center/library market was $112.50; in 1983 it was $99.34. The average expenditure per hour, or average revenue generated per hour, takes into account all items for which charges appear on bills sent by vendors to user accounts. This includes charges not only for online connect time but also for prints, displays, communications, SDI's, documentation, terminal rental, training sessions or anything else for which there is a charge (Williams, 1986).

There were 363 vendors of online and time-sharing search services worldwide in 1985, however only a few ruled the market. An in-depth look at Dialog shows some of the reasons for its success. First they had over 200 databases with 100 million items (250 gigabytes of storage) in 1984. But they also went after a broad set of clients, attracted by their databases and by a "full-service" system.

> One of the most successful operators, in terms of databases offered, client base and turnover is Dialog. . . . Dialog Version 2 offers the client the capability to create a tailor-made

> package with data provided with tagged output, online editing of saved searches and SDIs, and a "Report" feature which formats numeric data. Dialmail offers current awareness searches and offline prints; all can be delivered via Dialnet, a dedicated wire. In addition to such convergent services, Dialog offers Dialorder (complete texts may be ordered online), and Dialindex (Davenport and Cronin, 1987).

In contrast, Mead Data Central dominated the legal sector and the business sector, offering full-text with almost 20 million documents (and a staff of 1300) in 1984. The emphasis was both on free-text, end-user searching and on extensive training in law schools and business schools.

> Services now include Exchange (a financial data processing system which qualifies marketing leads) aimed at Merrill/Lynch, Paine Webber and similar corporations, and Eclipse (an off-peak electronic clipping service which creates and sustains a customised file) intended to give Nexis additional competitive edge (Davenport and Cronin, 1987).

The other two leading vendors were Medline, with a set of 21 medically-oriented databases in 1984, and BRS, which provided the largest full-text scientific journal collection, including major medical journals.

Large organizations, especially in business, legal services, or medicine, were able to select multiple vendors if necessary. However smaller academic or public libraries usually had to pick only one. A major consideration was the databases offered and the price, but there were many other factors (Rice, 1985), such as how easy was the system to use and how much training was needed.

5.2.1 Research within the Online Community

Research into better understanding of the search intermediary process was done within the various Information Science programs such as those at Syracuse, Rutgers, and the University of Maryland. One example

was the set of papers by Raya Fidel based on case studies of search intermediaries.

In 1984 she observed five professional search intermediaries performing 10 to 13 searches in the course of their regular job in the health services domain, and found two distinct patterns of searching called "operationalist" and "conceptualist" (Fidel, 1984). The operationalist style tended to use various system features to "modify a retrieved set without changing the conceptual meaning it represents". The conceptualist, on the other hand, "modified the retrieved set by changing the meaning of the concept", possibly by broadening or narrowing the descriptors used or looking into different facets. The paper discussed how these different styles affected search tactics during the whole process from preparing the search, selecting a database, choosing search terms, etc.

The second paper looked at how search terms were selected (Fidel, 1988), in particular noting when controlled vocabulary was used versus free-text searching. This time there were 281 searches from 47 searchers in many subject areas. Here she found that the subject area had the biggest impact, with "searchers of the scientific literature" using a higher-proportion of free-text terms, except in medicine where there was a reliable source of controlled terms. Additionally searches involving multiple databases used free-text terms since there was no common thesaurus. In 1991 she did a much larger study of both searching styles and search term selection (Fidel, 1991).

There were also studies on the effectiveness of these searches. Whereas the vendors did not report much on this aspect, there was a major study in 1985 looking at effectiveness of retrieval using the IBM STAIRS system (Storage And Information Retrieval System) to search 40,000 documents (350,000 pages) involved in a corporate law suit (Blair and Maron, 1985). The STAIRS system, first developed in 1969 (supposedly for defense in IBM's antitrust suit), was used by large corporations and government agencies for in-house searching of their text bases. The system was an advanced Boolean system, including field searching and the possibility of ordering retrieved sets of 200 or less by date, author, or frequency of query terms.

The experimental setup involved two lawyers, the principal defense lawyers in the suit, who generated 51 different information requests,

which were then translated into search queries by experienced paralegal searchers. The retrieved sets were judged by the lawyers as "vital", "satisfactory", "marginally relevant", and "irrelevant". Further modification could then be made to the queries and this continued until the lawyer stated that "he or she was satisfied with the search results for that particular query (i.e., ... more the 75% of the relevant had been retrieved)".

Precision was calculated over the full set of queries for each information request, and was found to range from 100% to 19.6%, with an average of around 75%. The recall was more difficult to define and was estimated by doing additional sample searches (and relevance judgements). "To find the *unretrieved* relevant documents we developed sample frames consisting of subsets of the unretrieved database that we believed to be rich in relevant documents." The values of recall ranged from 78.6% to 2.8%, with an average of only 20%.

Further analysis was done, particularly into the fact that whereas the lawyers thought they had 75% of the relevant, in actuality they only had 20%. The following were some of the conclusions.

- The low values of Recall occurred because full-text retrieval is difficult to use to retrieve documents by subject because its design is based on the assumption that it is a simple matter for users to foresee the exact words and phrases that will be used in the documents they will find useful, and *only* in those documents.
- The belief in the predictability of the words and phrases that may be used to discuss a particular subject is a difficult prejudice to overcome.
- the amount of search effort required to obtain the same Recall level increases as the database (size) increases, often at a faster rate than the increase in database size. On the database we studied, there were many search terms that, used by themselves, would retrieve over 10,000 documents. Such an *output overload* is a frequent problem of full-text retrieval systems.

- This study is a clear demonstration of just how sophisticated search skills must be to use STAIRS, or any other full-text retrieval vendor.
- We have shown that the system did not work well in the environment in which it was tested and that there are theoretical reasons why full-text systems applied to large databases are unlikely to perform well in any retrieval environment.

This gloomy analysis did not have much effect on the booming online vendor community. However it enraged information retrieval researchers, such as Salton, since the broad-sweeping conclusions assumed that the only valid search mechanism was Boolean retrieval!

As more PCs became available and there were many more potential end-users, the large vendors such as Dialog and BRS offered simpler systems like BRS/After Dark to expand their cliental (and increase use after business hours). These were not generally very successful, and Roger Summit, founder and long-time head of Dialog noted that only 12% of their usage was from end-users in a paper "In Search of the Elusive End User" (Summit, 1989). In this paper he commented that "End users are costly to serve. Their practice of occasional (as opposed to frequent) searching requires simpler interfaces than are offered today, but these interfaces must produce satisfactory results and be understandable."

In a Letter to the Editor shortly thereafter, Cyril Cleverdon reacted (Cleverdon, 1990):

> I do not argue with the points he raises, but I find it astonishing that he makes no mention of the problem which far outweighs all the others. I refer to the insistence on Boolean searching. For the occasional user of an online system, the preparation of a Boolean search is a serious problem. Having little or no practical experience, there is a high probability that the search will be so broad that it will result in a large unmanageable output, or will be so restricted that nothing is retrieved.

5.3 Research in the Second Half of the 1980s

5.3.1 Research with Online Card Catalogs

End users were 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 (Siegel *et al.*, 1984) 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 (Doszkocs, 1982; Doszkocs and Rapp, 1979) incorporated natural language input and ranking similar to the SMART system, with the other being a more conventional Boolean system (the ILS system). 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. The CITE system was overwhelmingly preferred!

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, evaluated within an OPAC operational setting (the Okapi system), took place starting in 1984 (Mitev *et al.*, 1985), examining the effects of system changes on user search performance (Robertson and Hancock-Beaulieu, 1992). 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 plus transaction logging. In general users liked the new system, but it was found that only 38% of the searches were for subjects and that the average number of terms per subject search was just over two.

The Okapi 1987 experiments (Walker and Jones, 1987) investigated stemming and spelling correction, with two versions of the system installed on alternative days on two terminals at the library. The control system had weak stemming, with the experimental system having a

strong stemmer (Porter), spelling correction and a lookup table for phrases and equivalence classes of related terms. The weak stemming retrieved more records in almost half of the initial searches repeated from the transaction logs, but 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 (Walker and de Vere, 1989) examined the effects of relevance feedback. The 1987 system (with weak stemming and no spelling correction) was the base system, with a second system offering query expansion (look for books similar to), and a third system including a "shelf browsing" system based on the Dewey system. A laboratory experiment with over 50 subjects was done and based on the interviews, the query expansion system was considered highly acceptable, with the shelf browsing much less so.

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, this time within an operational setting (Hancock-Beaulieu and Walker, 1992; Walker and Hancock-Beaulieu, 1991). Query expansion was offered only when searchers had chosen at least two items as relevant. For over half of the time it was used, no additional items were selected by the users and 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.

Note that these user tests were the first time that experimental search methodologies had been evaluated in an operational setting (with the exception of the single CITE evaluation by NLM). As such they 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.

5.3.2 Research with Search Intermediaries

An extensive user study was led by Tefko Saracevic, now at the School of Communication, Information and Library Studies at Rutgers (Saracevic and Kantor, 1988a,b; Saracevic *et al.*, 1988). The stated goal of the study was "to contribute to the formal characterization of the elements involved in information seeking and retrieving, particularly in relation to the cognitive context and human decisions and interactions involved in the processes."

Note that this study had a clear separation between the users (the people with the questions or information need) and the searchers, who were assumed to be professional searchers (search intermediaries). Therefore the general model for information seeking was broken down into areas concerning the user (the problem underlying question, the intended use of information, their previous knowledge and expectations of results), and areas concerning the searcher (their language and logic abilities, learning style and online experience). Various tests and metrics were discussed, including not only recall and precision but also some utility metrics such as time, cost, user satisfaction, etc.

Forty volunteer users participated, including providing a question (and context), taking part in an interview and reviewing results. There were 40 information professionals who were paid for their time and searched five or six of the questions (each question was searched by five different searchers).

Saracevic's papers contained numerous small studies comparing user characteristics, searcher characteristics, and the results of the searches. Here is a small sample of what was found:

> - The utility of results (or user satisfaction) may be associated with high precision, while recall does not play a role that is even closely as significant. . . . In a way this points out to the elusive nature of recall; this measure is based on the assumption that something may be missing. Users cannot tell what is missing any more than searcher or systems can.

- More (search) cycles, which allow for feedback, seem to produce better searches as to relevance odds, and more search terms . . . seem to result in worse searches.
- In general, (professional) searches based on problem and intent statements by users out-performed on the average all other types of searches, including searches based on the written questions.
- The mean overlap of search terms was 27%, but the distribution was skewed towards the low end: in 44% of the comparisons between searches the overlap was between 0% and 20%.
- The mean overlap in retrieved items was 17% and the distribution was even more skewed: in 59% of comparisons between retrieved sets the overlap was between 0% and 5%. It seems that different searchers for the same question more or less look for and retrieve a different portion of the file. They seem to see different things in a question and/or interpret them in a different way and as a result retrieve different items.

It is interesting to compare this study to Saracevic's earlier one (Saracevic, 1971) before there were professional searchers. Although that study was mostly comparing the different ways of indexing documents, it was also found then that using just the question as stated by the user was not sufficient and that the search needed to be broadened. The Rutgers study can also be compared with the overlap study from Syracuse University where there was little overlap in retrieved documents based on which parts of an indexed document were used for searching.

5.3.3 Rethinking Information Retrieval Systems Research

Karen Spärck Jones's acceptance speech for the Salton award (Spärck Jones, 1988) characterized many of the problems in the late 1980s for the field. Because work in the 1960s and 1970s had concentrated on retrieval process, there had been two "unfortunate consequences".

> One was that, in spite of references to environmental parameters and so forth, it tested information systems in an abstract, reductionist way which was not only felt to be disagreeably arid but was judged to neglect not only important operational matters but, more importantly, much of the vital business of establishing the user's need. The second unfortunate consequence of the way we worked in the sixties and seventies was that while the research community was struggling to satisfy itself in the laboratory, the operational world could not wait, and passed it by. The research experiments were so small, the theory was so impenetrable, and the results it gave were at best so marginal in degree and locus, that they all seemed irrelevant.

She also worried about the increasingly complex "information management" systems or highly-touted expert systems.

> It is difficult, therefore, to see how any credibility can be attached to work in the intrinsically much more uncertain and miscellaneous area represented by the multifaceted systems of the future unless the tests done to justify their design are well-founded. This is not going to be easy: as the techniques become more complex, so the numbers of variables and parameters increase, the testing requirements go up. ... But the most important problem we have to face in test, and especially experimental, design is that as we increase the emphasis on interaction with the individual user, we get less repeatability. This is not a new problem, but it is exacerbated by the complexity of the system, the power of the interface, and the need for tests on a large scale.

Whereas the user focussed research in the late 1980s had been partially driven by the growth in online systems, the more system-oriented research in the late 1980s was partially driven by the rise of expert systems and "knowledge bases".

One example of this was the 1986 paper "A New Theoretical Framework for Information Retrieval" (van Rijsbergen, 1986). This paper

asserted that using only statistical "word-matching" had a limited performance and that some type of logic was required to move further. It was suggested that there needed to be a model of the content of a document using appropriate semantics and that the use of semantics "comes via an appropriate logic". William Cooper had suggested the use of logic as a way of defining relevance earlier (Cooper, 1971), but the techniques coming from the expert system community made it possible to now envision systems based on logic.

> Such a logic would be based on a formal semantics for text. The semantics would provide a limited representation of the meaning of any text but it would not be the meaning. A logic would then be interpretable in that semantics (van Rijsbergen, 1986).

This type of thinking also reflected the growing interest in natural language processing and knowledge bases, such as demonstrated by the expert system SCISOR (Rau, 1987). SCISOR retrieved information in the domain of corporate takeovers based on an elaborate knowledge base that handled inexact retrieval. Later that system was used for extracting information from online news (Jacobs and Rau, 1990).

5.3.4 New Models at the University of Massachusetts, Amherst

The group led by Bruce Croft at the University of Massachusetts, Amherst, developed an expert system for information retrieval and went on to define methods to use inference networks to allow the combining of different "evidence" to improve retrieval. Their I^3R system (Croft and Thompson, 1987) used a blackboard-type architecture that could employ various system "expert" components as needed, with users influencing the system actions by stating goals they wished to achieve, by evaluating system output, and by choosing particular facilities directly. The next step was the defining of a "plausible inference" model (Croft *et al.*, 1988).

> The task of a retrieval system can be described as determining if there is a plausible relationship between a document

> and the query and assessing the credibility of this relationship. Documents are retrieved in an order defined by the credibility of their relationship with the query.

The final step was the design of inference networks (Turtle and Croft, 1989, 1991a) that would enable the principled combining of evidence. The evidence could be different (multiple) representations of documents, addition of alternative terms from knowledge bases, the use of multiple queries, etc. The framework of choice here was a Bayesian inference network, "a directed, acyclic dependency graph (DAG) in which nodes represent propositional variables or constants and edges represent dependence relations between propositions".

The document network consisted of document nodes, text representation nodes, and concept representation nodes. The document nodes were the roots, one node for each document, with the first layer of children being the text nodes, containing the text for the given document. Note that other types of document children could be included, such as figures, tables, video, etc. Below the text nodes were the concept nodes, "each corresponding to a single representation technique that has been applied to the document texts". So this could include the single terms, the phrases, the manual index terms, etc. The query network was composed of (possible) multiple query representations, with these broken down into multiple concepts, with concepts defined similarly to the document network.

The various estimates for probabilities within the network were encoded via a link matrix. Four different link matrix forms were discussed in the papers: three for Boolean operators and one for a simple probabilistic model. A third paper (Turtle and Croft, 1991b) discussed efficient implementations of the inference model.

5.3.5 Some Other Advanced Models and Systems

The continued advancement in computer design, memory and processor speed enabled other advancements in the late 1980s. Additionally the growth of the commercial online systems was beginning to draw new interest from outside the "traditional" information retrieval community.

William Cooper (now at the University of California, Berkeley) proposed the use of the maximum entropy principle as a type of probabilistic approach to weighting of terms (Cooper, 1983). His paper, a serious critique of then current Boolean systems, broadly discussed various ways of helping users, including different ways of implementing maximum entropy within a Boolean system.

The Connection Machine by Thinking Systems had a parallel architecture, with one to four modules of 16,834 processing elements each, with each element containing 4096 bits of memory (Stanfill and Kahle, 1986). These could be used in what was called memory-based reasoning, where large amounts of data could be analyzed for patterns that were input to similarity-based induction. This early type of machine learning was demonstrated by learning pronunciation of novel words (Stanfill and Waltz, 1986), but could have been used for classification and other similar tasks.

Another proponent of machine learning was Norbert Fuhr from TH Darmstadt. His probabilistic weighting of terms was based on learning weights from past "use" of these terms. Here "use" was abstractly defined and could relate to use in manual indexing via a controlled vocabulary (Fuhr, 1989) or use within a relevance feedback operation (Fuhr and Buckley, 1991). He also proposed that this learning could be continuous as more data became available.

Richard Belew at the University of California, San Diego, designed and tested a connectionist representation called AIR (Adaptive Information Retrieval) (Belew, 1989). This was built, along with a user interface, for a small dataset representing 1800 papers in artificial intelligence. The initial network contained document representations, including authors and initial keywords. Users would input a simple query using these features and spreading activation was used to return appropriate network nodes. Users could then mark relevant nodes on the interface, modifying the network.

A team from Bellcore, the University of Chicago, and the University of Western Ontario (Deerwester *et al.*, 1990; Furnas *et al.*, 1988) investigated a new method of automatic indexing (and retrieval) specifically aimed at solving the "vocabulary mismatch" between queries and documents. Note that this was a long-standing problem in retrieval, or

any type of information access and had driven the use of thesaurus, and term association research. They used a type of factor analysis, called Singular-Value Decomposition (SVD) in which a large term by document matrix is "compressed" into a set of factors that can represent both documents and queries.

A special issue of the IEEE Computer Society on Data Engineering (Lee, 1990) detailed eight different advanced system approaches, including articles by Salton, Croft, Stanfill, Hollaar, and Faloutsos.

5.3.6 Rethinking User Interfaces

There was increased thinking about the "end-users" as opposed to the search intermediaries and new models for end user searching were being proposed, such as the one by Marcia Bates from the University of California, Los Angeles (Bates, 1989).

> In real-life searches in manual sources, end users may begin with just one feature of a broader topic, or just one relevant reference, and move through a variety of sources. Each new piece of information they encounter gives them new ideas and directions to follow and, consequently, a new conception of the query. . . . In other words, the query is satisfied not by a single final retrieved set, but by a series of selections of individual references and bits of information at each stage of the ever-modifying search. A bit-at-a-time retrieval of this sort is here called berrypicking.

There was also the very innovative prototype interface (the Bookhouse) designed especially for end users looking for fiction books in a library by Annelise Mark Pejtersen from FCI Informatics Research-Center (Pejtersen, 1989). The plans and eventual icon-based interface (four delightful figures are in the paper) were based on several years of field studies of user-librarian conversations. The prototype was tested in public library for six months and was well received by users of all ages.

Helping end users improve queries was explored by Donna Harman (Harman, 1988) from the National Institute of Standards and

Technology. The motivational vision was a screen allowing users to pick new terms, and experiments were run on the Cranfield 1400 collection to discover the optimal terms to display. Three term finding methods were tried: relevance feedback, term variants (stemming), and nearest neighbor terms. Results showed significant improvement in general, and very large improvements theoretically possible if users were able to pick the best terms (no actual user studies were done).

A final example of end user focus is the sense-making method from Brenda Devlin from Ohio State University (Devlin, 1983)

> Sense-Making studies and applications, thus, have all incorporated two or more of the following:
>
> SITUATIONS: The time-space contexts at which sense is constructed.
>
> GAPS: The gaps seen as needing bridging, translated in most studies as "information needs" or the questions people have as construct sense and move through time-space.
>
> USES: The uses to which the individual puts newly created sense, translated in most studies as information helps and hurts.

6

Explosion (1990s)

6.1 Pre-Web and the Arrival of Search Engines

The computer world in the early 1990s had completely changed from the world ten years earlier. Not only had computer power and storage continued to have exponential growth (and lower prices), but there were now individual workstations, such as Sun's SPARC 2, with CPU speed ranging from 25 to 40 MHz and having 48 to 128 MB of RAM. These workstations used the UNIX operating system, including built-in networking software.

But the most important change was in the telecommunication networks. The ARPANET's 118 nodes had been mainly for government agencies and the military, but in 1981 NSF funded CSNET (Computer Science Network) to allow computer science departments to hook into ARPANET; eventually there were over 180 institutions using CSNET.

In 1985 NSF started funding five supercomputing centers at universities, which were shortly thereafter interconnected by NSFNET. The six-site backbone ran on leased 56-kbit/s links, using TCP/IP protocols. This backbone also supported regional networks, which in turn were linked into many mostly university sites (Figure 6.1). In 1988 the backbone (T-1) was expanded to 13 nodes, running at 1.5-Mbits/s.

By the end of 1991, the T-1 backbone had been upgraded to 45 Mbit/s (T-3) and the total traffic exceeded 1 trillion bytes or 10 billion packets per month. By 1992 there were over 7,500 networks in the system and over 1 million computers connected internationally.

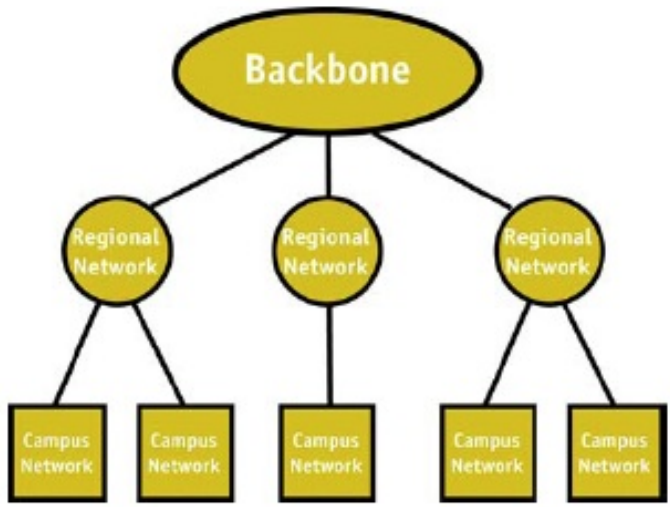

Figure 6.1: NSF's three tiered network architecture.

So what could the network be used for? There was email (provided a person had a network connection) and there was the ability to transfer files via FTP. There were commercial online services such as AOL and CompuServe where users with accounts could access the networks via phone lines and modems. But very quickly other trial software was built for the network, mainly by students at universities.

- In 1990 Archie was released by Peter Deutsch, Alan Emtage, and Bill Heelan at McGill University. Archie was a "search engine" that allowed users to log into a specific site (an Archie server) and using command lines, search for data, such as files from public FTP sites that had been previously collected for that server (or other connected Archie servers).

- In early 1991 Tim Berners-Lee designed the HyperText Transfer Protocol (HTTP), the HyperText Markup Language (HTML) and

the first Web browser for the NeXt environment. He called this program "World Wide Web" and built the first website at CERN.

- In 1991 the WAIS (Wide Area Information Server) system was developed by Brewster Kahle and Harry Morris, both of Thinking Machines, Cambridge, Massachusetts, in collaboration with Apple Computer, Dow Jones, and KPMG Peat Marwick. WAIS was a client-server full text searching system that used the ANSI Standard Z39.50 to index and search databases on WAIS servers. The code was open-sourced and various groups used the software to build databases for their information.

- Also in 1991, Gopher was released by Paul Lindner and Mark P. McCahill from the University of Minnesota. Gopher was a protocol that enabled storage of documents, indexing them via directories or menus and connections to various Gopher servers (named for the mascot of the University of Minnesota). There were also search engines such as Veronica (Very easy rodent-oriented netwide index to computerized archives) built for these Gopher servers.

- In July 1992 the WWW client software was made publicly available by CERN; by November there were 23 servers on the Web. It was not heavily used however because it was only developed for the NeXt operating system at that point.

In January 1993 Marc Andreessen from the National Center for Supercomputing Applications (NCSA) at the University of Illinois in Urbana/Champlain released the Mosaic web browser, based on the Berners-Lee proposal, but built for the UNIX operating systems. It came with very liberal (free to non-commercial groups) licensing agreements, with a version for Windows in August. Mosaic was quickly picked up by many groups: by the end of 1993 there were more than 1000 downloads per day of the code. It also created intense interest in the WWW because of its reliability, its ease of installation and use, and its interface. "In January 1993, there were 50 known Web servers. By October, there were more than 500. By June 1994, there were 1,500" (Wolfe, 1994).

> The secret of Mosaic's success is no mystery. When you browse with Mosaic, you see a series of well-proportioned "pages", with neat headlines and full-color images. You can fiddle with the screen to suit your own preferences. (I like grayish-purple text, with links in blue.) You can mark your progress forward and back in the Web, and make a "hotlist" of places you visit often. On the Macintosh version, which I use, you move up and down the page in the conventional fashion, using a scroll bar on your right (Wolfe, 1994).

The following timeline shows the incredible growth of the web and the rapid emergence of new search engines. (Note: this information was compiled from many sources and is only a reasonable estimate of the dates and web size; one source was Seymour *et al.* (2011).)

1993–Mosaic released in January; 130 websites in June grew to 623 by December.

1994–Yahoo! started in April; first WWW meeting held in May; Lycos went public in July; 10,222 websites by December.

1995–Infoseek started in February; Excite started in October; AltaVista launched by DEC in December with 300,000 hits on its first day.

1996–In January there were 100,000 websites, doubling by June with over half being ".com" sites.

1997–In April The Search Engine meeting was held in Bath, U.K. Invited presenters from most major (and minor) search engines (Excite, Lycos, Verity, Personal Library Software, InfoMarket, Autonomy, Muscat, Excalibur, Claritech, etc.) mostly talked during breaks about what was the correct business model for these engines, such as subscription based (like journals), or ad-based (like TV).

1998–Google Search started; Microsoft started a search portal called MSN Search, using search results from Inktomi. It did not have in-house searching until 2005 and changed its name to Bing in 2009.

6.2 IR Research Expands in All Directions in the 1990s

The beginning of the 1990s saw continuation of the traditional IR research, still working with the small test collections. But as the computers got faster, and more machine-readable text became available for development and testing, research expanded in all directions. Not only did the systems of the 1980s double their performance, but many new systems were started by diverse groups. There was expansion into new applications of IR technology, some of which were driven by the growth of the web, but other areas of long-term interest became possible because of the increased technology and data. The experiments discussed in this section are only a small sample of the work that was done, particularly in the late 1990s when information retrieval research "exploded". Some of the papers mentioned won the SIGIR Test-of-Time award;[1] others were added to continue some of the themes discussed in earlier chapters.

6.2.1 Early 1990s Research

The established groups still continued traditional research. The SMART group investigated a set of 12 different relevance feedback techniques with the small collections, using both the vector space model and probabilistic models (Salton and Buckley, 1990). The group at the University of Massachusetts, Amherst, worked with the old CACM collection on methods for handling structured queries in their inference network (Croft *et al.*, 1991).

At this point in time there were no publicly available larger test collections, although several groups were able to work with larger sets of documents obtained privately. The National Institute of Standards and Technology (NIST) built and tested a working prototype using ranking techniques for the U.S. Internal Revenue Service using 806 MB of manuals, legal code and court cases (Harman and Candela, 1990). Xerox Parc (Cutting *et al.*, 1992) designed a special browsing interface for 5000 articles from the New York Times News Service. The browsing

[1] This is an award given annually to papers that have been shown to have a major influence on the field ten years later.

technique called "Scatter/Gather" first clustered the documents into major news areas and users could select one or more clusters.

In 1992 a new book *Information Retrieval: Data Structures and Algorithms* (Frakes and Baeza-Yates, 1992) was published that contained not only algorithms for the basic parts of an information retrieval system (suffixing, making inverted files, ranking, feedback, etc.), but even some code on a CD-ROM and a ftp site.

6.2.2 TIPSTER and TREC

In late 1990 DARPA announced a new program called TIPSTER, with the stated goal of "significant advances in the state-of-the-art in information extraction and in document detection". This was the first major funding for information retrieval research since the late 1970s. For the document detection part, there was to be a large test collection built by NIST with two gigabytes of documents.

This test collection was built using an extension of the Cranfield paradigm. The two gigabytes of documents were from multiple domains, with about half being full text from newspapers and newswires. There were fifty queries (called topics in TREC) built by people not familiar with information retrieval research and relevance assessments were made for each query, with an average of over 1500 documents judged per query. The selection of documents to judge was made by the pooling method, taking the top 100 documents returned by each system in TREC-1.

This test corpus was used in 1992 at the first TREC (Text REtrieval Conference) (Voorhees and Harman, 2005) and was made freely available for research. Twenty-five systems and three contractors took part in TREC-1, using a wide variety of retrieval models that were tested for the first time on such a large amount of text. Researchers were able to improve performance by TREC-2 by using the TREC-1 collection for training, and systems had basically doubled their performance within three years of testing (Harman, 2005).

During the 1990's TREC continued to build test collections in various areas of interest, including cross-language retrieval, retrieval from speech

(as opposed to written text), filtering tasks, and question-answering (Figure 6.2).

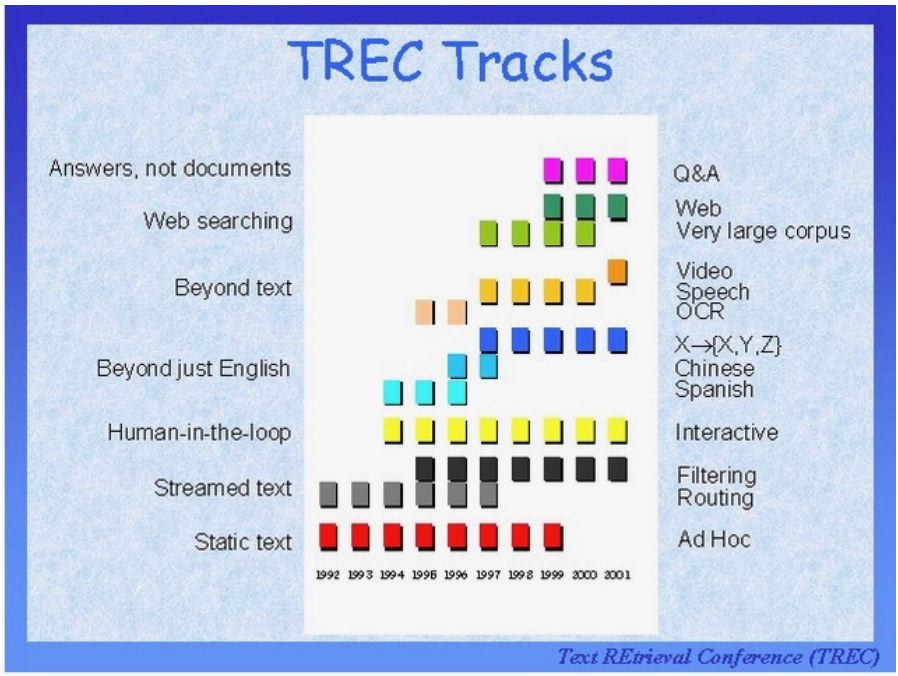

Figure 6.2: The first 10 TRECs.

For more on TREC, including the full set of online proceedings and availability of test data, see the NIST website (https://trec.nist.gov/). There is also a book (Voorhees and Harman, 2005) and discussion papers such as "Further Reflections on TREC" (Spärck Jones, 2000).

6.2.3 Research on Basic Retrieval Algorithms

The research done by all of the systems to handle TREC's longer documents (and topics), and to find new ways of expanding the queries in TREC-3 make for compelling reading. Systems were based on different models and therefore had to take different approaches to solve these common problems. The following papers give some idea of that research.

Stephen Robertson and Steve Walker's 1994 paper documented their initial experiments resulting in the BM25 algorithm (Robertson and

Walker, 1994). The paper explored the incorporation of within-document term frequencies, how to normalize them for document length, and also the use of query term frequencies.

The SMART system already incorporated within-document term frequency and query term frequency, however the document length normalization was only minimally handled by the use of the cosine similarity measure. A "pivoting" formula was developed by Amit Singhal (Singhal *et al.*, 1996) to modify the cosine normalization.

The removal of the concepts field in TREC-3 caused many systems to investigate query term expansion. An example of this is a paper from Jinxi Xu and the UMass/INQUERY group (Xu and Croft, 1996) comparing expansion using pseudo relevance feedback and expansion using their PhraseFinder (based on co-occurrence of terms).

There was work with neural nets, such as the research by K. L. Kwok (Kwok, 1995) of Queens University or Josiane Mothe (Mothe, 1994) of the Universite P. Sabatier, France. There was further work with Latent Semantic Indexing (LSI) and a similar concept used by one of the TIPSTER contractors (HNC) (Caid *et al.*, 1995).

6.2.4 Comparison to Boolean Systems

One of the concerns in the 1990s was the comparison of these new systems to the traditional Boolean systems, particularly in domains where there was considerable experience with Boolean searching. William Hersh ran an experiment using "relatively inexperienced physician end users" to test a CD-ROM retrieval system (Knowledge Finder) (Hersh *et al.*, 1994) and demonstrated that end-users could effectively use these systems.

In the same year, Howard Turtle (at Westlaw) ran a large-scale comparison of natural language vs. Boolean retrieval using a large collection of full-text legal cases (Turtle, 1994). The results showed that natural language queries had significantly higher average precision except at the lowest recall level, retrieved more relevant in the top twenty documents (57% vs. 42.3%) and outperformed Boolean on 36 queries out of the 44 queries.

6.2.5 Other User Studies

The 1990s also saw more diverse user studies. The information science community continued studies such as the one by Carol Barry from Louisiana State University on relevance criteria (Barry, 1994; Barry and Schamber, 1998). However the ranking systems had grown effective enough to allow user testing, both as to how well the ranking worked for real users (as opposed to test collections), and as a basis for trying new interfaces.

For example, the group from Rutgers University (Koenemann and Belkin, 1996) compared the use of relevance feedback by search intermediaries and university students. Marti Hearst from Xerox Corporation (Hearst *et al.*, 1996) tested three different interfaces (standard ranked retrieval, Scatter/Gather document clustering and TileBars) and James Allan from the University of Massachusetts, Amherst (Allan *et al.*, 1998a) tried a new 3-D visualization interface using 8 librarians and 12 general users.

6.2.6 Text Categorization and Filtering

Text categorization and filtering had been of interest since the early 1960s, however technology limitations and the lack of machine-readable data was a major barrier to experimentation. Manually creating categories/classifications, such as for newswires, is expensive and in 1990 Reuters and the Carnegie Group experimented with an automatic method (Hayes and Weinstein, 1990). The training and test collection used for this work was cleaned up for research distribution by David Lewis at the University of Massachusetts, and became heavily used (and replaced by a better one!). The research in classification was both for accuracy and for efficiency (see Lewis and Gale (1994) for an efficiency example).

Traditional filtering had also started early with the SDI systems, where manually-built profiles were used. The TREC filtering task (Robertson and Callan, 2005) attracted new experimentation, however one of the issues in the TREC task was how to effectively evaluate filtering, particularly when the specific application was not defined. But as the 1990s moved on, real applications developed.

One was the DARPA program (Topic Detection and Tracking (TDT)) (Allan *et al.*, 1998b) where the goal was to detect new events and then track them via newsfeeds. The technology started here can be seen in the "breaking news" and selective newsfeeds seen today on the Internet.

The second of these filtering papers by Jonathan Herlocker from Oregon State University looked at methods to perform collaborative filtering (Herlocker *et al.*, 1999), in particular training and testing with the MovieLens recommendation site with 122,176 ratings from 1173 users.

6.2.7 New Retrieval Models for Ranking

The rapidly increasing number and size of collections inspired research by Jamie Callan into handling distributed collections, such as collections that could belong to different organizations, or collections that could be split into smaller sections for efficiency considerations (Callan *et al.*, 1995).

The general availability of large amounts of text (not just test collections) led to more general language modeling approaches. The basic idea of language modeling came from the speech recognition world, where the probability of recognizing a given word could be based on recognizing its surrounding words using training data.

The Ponte and Croft language modeling paper (Ponte and Croft, 1998) looked at the probability distribution of terms occurring in all documents (not just relevant ones) using several probability measures. A second language modeling paper (Berger and Lafferty, 1999) was based on the statistical machine translation type of language modeling, with three more papers presented at SIGIR 2001 (Lafferty and Zhai, 2001; Lavrenko and Croft, 2001; Zhai and Lafferty, 2001).

6.2.8 Research to Improve Web Performance

The web was now getting huge; searching was getting more difficult and people were using the web for many different types of information needs. Searching could involve very specific questions (which may or may not find anything on the web), but also could involve very broad

topics, for which far too many web pages would be returned. Since most people only input a couple of words in the query, there was little that could be done by traditional ranking methods for the broad topics.

Kleinberg (1998, 1999) from Cornell investigated ways of using the links in the web to help with these broad topics. He reasoned that links that had been included in a web page by its author would point to some type of "authority" on that topic, much as citations in a paper.

The 1998 paper "The Anatomy of a Large-scale Hypertextual Web Search Engine" (Brin and Page, 1998) by two Stanford graduate students detailed the Google system, including many of the engineering challenges. The design goals make an interesting read now.

A new technology in Google was the PageRank algorithm (Page *et al.*, 1998) developed at the Stanford Labs. Like the earlier Kleinberg work, it could be compared to citation linking, however with a major difference in the way it is calculated.

6.2.9 Evaluation

The heavy use of test collections meant that researchers were concerned about the validity of the results based on these collections. Two papers were presented at SIGIR 1998, the first by Ellen Voorhees looking at the effects of variation in relevance judgments (Voorhees, 1998) and the second by Justin Zobel (Zobel, 1998) examining the reliability of results using the collections, i.e., how complete were the relevance judgments. A later paper by Chris Buckley (Buckley and Voorhees, 2000) investigated the stability of traditional IR metrics.

There was also interest in new ways of measuring relevancy. Jaime Carbonell and Jade Goldstein from Carnegie Mellon University suggested incorporating "novelty" (Carbonell and Goldstein, 1998) into the relevancy scoring. Kalervo Järvelin and Jaana Kekäläinen of the University of Tampere (Järvelin and Kekäläinen, 2000) proposed a completely new metric using graded relevance judgments/scores (0–3) and then accumulating scores while moving down the ranked list. This discounted cumulative gain metric and its successor, the normalized Discounted Cumulative Gain (nDCG) have been heavily used by both the web community and the IR community.

Two more evaluation "competitions" started. The first NTCIR (NACSIS Test Collection for Information Retrieval) workshop was organized by Noriko Kando (Kando, 1999) at the end of August 1999 to work with Asian languages. The CLEF conference (Cross-Language Evaluation Forum), a spinoff from the TREC cross-language track, was coordinated by Carol Peters (Peters, 2003) and took place in late September of 2000, specifically concentrating on European languages.

6.2.10 The 1990s: the End of One Era and the Beginning of Another

In 1990 there was very little publicly available machine-readable text, and only the older small test collections. The operational systems were strictly Boolean, with large amounts of manually-produced metadata. Members of IR research community were a small minority believing in ranking, or even full reliance on automatic indexing.

By the end of the 1990s, there was large amounts of publicly available machine-readable text, there were multiple large test collections, and ranking and automatic indexing were the assumed standard by almost everyone. Additionally search engines had created a compelling application for the IR algorithms and the interest in IR research was booming, as can be seen by the growth of the SIGIR conferences.

The SIGIR conference held in Chicago in 1991 had 76 submissions (32 were accepted). By 1996, the conference (in Zurich) had 139 submissions. and the Melbourne SIGIR conference in 1998 had 161 papers submitted (83 selected). These submitted papers came from diverse research areas: the SIGIR 1999 (Berkeley, California) submissions had 52 papers in "core" IR, 38 in NLP, categorization and clustering, 20 in multi-media and 25 in HCI (Human Computer Interaction) and IR. These papers were also from more diverse research groups: SIGIR 2000 in Athens had 41 papers submitted from Asia/Pacific, 42 from Europe and 57 from the Americas.

Information retrieval finally came into its own in the 1990s; it was heady times for the community!!

7

And it Continues

It is now the fall of 2018. There is "infinite" computer power, "infinite" communication bandwidth, and "infinite" storage. There are over 2 billion "smart" phones that use wireless communication for phoning plus sending text messages and email and accessing the Internet, along with thousands of other applications online.

SIGIR 2018, held in Ann Arbor, Michigan, had 409 papers submitted for the full paper track and 1247 authors from 33 countries. SIGIR also co-sponsors five conferences, three of which are new since 2000, all with papers involving information retrieval. And there is a continually growing set of regional conferences in information retrieval, such as ECIR, AIRS, and ADCS.

Whereas the early growth of information retrieval technology had been driven by increased computer power, memory, storage, and faster networks, it is now driven mostly by the massive increase in the amount and type of information available. The size of the Internet is huge, with an estimate of over 4.5 billion indexed pages in Google in October 2018.

Some of this information is just an extension of what could have been found in the past. Most professional societies, newspapers, etc. have online versions of their text plus scanned versions in all the past archives.

Google and others have scanned in many types of "orphan" writing, such as old theatre magazines, social "blue books", town records, out-of-copyright books, etc., making almost anything that was ever printed available on the web. Museums and libraries have uploaded image collections for sharing.

But this is dwarfed by the amount (and type) of information being produced by individuals or companies. Because it is so easy to find "stuff", people, groups of people and companies create "stuff", including online encyclopedias (like Wikipedia), user forums (like MacForum), reviews (like TripAdvisor), shopping sites (like Amazon) and billions of hours of video on YouTube (5 billion hours watched each day, 300 hours of video uploaded each minute as of July 2018).

The tools built within the science of information retrieval are critical to the new tools being used for accessing all this content, both directly and indirectly. Search engines such as Google, Bing, Baidu, and Yandex incorporate some of the basic IR algorithms, but additionally have hundreds of other factors used to rank results. Video searching includes text search on metadata and the output of speech recognition systems. New tools, such as the recommender systems that are vital to advertising and online shopping, have search components.

Most of the trends from the 1990s have continued. The use of language modeling has become a powerful tool for many applications of information retrieval. Cross-language retrieval, filtering, recommenders, and question answering have become major areas of research. And work with users, including modeling behavior, modifying interfaces, providing new tools, and increased user testing (daily by search engines) has become ever more important.

The major search engines of the late 1990s have taken over the market, with the earlier ones either bought up or moved to more niche applications. These behemoths have adopted the advertising business model (along with Facebook, Twitter, etc.) and are huge moneymaking machines.

The evaluation forums such as TREC (26th year), CLEF (19th year) and NTCIR (14th running) have all continued and have been joined by two more evaluation forums, Forum for Information Retrieval Evaluation (FIRE in its 9th year in India) and the Russian Information

Retrieval Evaluation Seminar (ROMIP). All these evaluations have featured diverse tasks over the years that have tracked the academic and commercial interests. Some of the tasks this year involved medical record retrieval, incident recognition (for emergency response), image search, search for complex tasks, work with SMS messaging systems, entity recognition in English (TREC) and Indian languages (FIRE), and lifelogging. The INEX evaluation effort (Fuhr *et al.*, 2008) ran for 6 years looking at information retrieval with a concentration on the use of XML. TRECvid (Smeaton *et al.*, 2006), an evaluation for video retrieval, spun off from TREC. Test collections from these evaluations are still being used, along with test collections built by various research groups. The Cranfield paradigm for building collections is also still used, even as the type of material to be searched is more varied, such as blogs, tweets, etc. And there are still valid complaints about the use of test collections, in particular the problems with *relevance*.

Tefko Saracevic's 2007 articles on relevance (Saracevic, 2007a,b) concluded with:

> Information technology, information systems, and information retrieval will change in ways that we cannot even imagine, not only in the long run, but even in the short term. They are changing at an accelerated pace. But no matter what, relevance is here to stay. Relevance is timeless. Concerns about relevance will always be timely.

Relevance is one way of measuring user satisfaction and users have always been at the heart of information access. In the early libraries, the notion of classification schemes like Dewey were aimed at people being able to find books quickly and hopefully grouped together on a shelf. The issues in indexing looking at specificity and exhaustivity attempted to deliver documents that answered the user's information need without overwhelming them. These ideas were incorporated in the notion of relevance in the early test collections and the ability to retrieve relevant documents at high ranks and to retrieve all of them has driven information retrieval research right from the beginning.

The information world today is so broad that new ways of measuring relevance have been needed. The commercial search engines may still

use test collections and measures like the nDCG (normalized discounted cumulative gain), but place more emphasis on analysis of user logs to understand user satisfaction, including user click modeling (Chuklin *et al.*, 2015) and statistics like abandonment rate (Diriye *et al.*, 2012).

Information retrieval, however, encompasses more than just web searching. Just as there were two parallel search worlds in the 1980s, the web community and the traditional information retrieval research community serve different goals (although with much more cooperation today). The retrieval technologies developed over the years are used in such diverse applications as legal discovery, medical record searching for clinical trials, and information "exploration" where the issue of user satisfaction is complex. Search serves as critical component of question answering systems, decision support systems, and "event" recognition systems such as for disaster coverage, etc. Here success is based on how well the output of the retrieval system serves the goal of the larger application system.

Looking ahead the next twenty years, it is hard to envision where information retrieval technology will be used. How will search cope with the "internet of things", how will today's search engines evolve, and how will new applications beyond imagination "reuse" the technology that started over 70 years ago. It is still heady times for the IR community!

Acknowledgments

I was very fortunate in being able to get comments on each chapter (decade) from researchers who had been heavily involved during that decade. So many many thanks to:

- Chapter 2: Georgette Dorn, Library of Congress
- Chapter 3: Michael Lesk, Rutgers University
- Chapter 4: Stephen Robertson, University College London
- Chapter 5: Ellen Voorhees, National Institute of Standards and Technology
- Chapter 6: Jamie Callan, Carnegie-Mellon University

Also many thanks to the anonymous reviewers for excellent suggestions (and to Mark Sanderson who asked me to write this monograph).

Appendices

A

Early Test Collections

The following tables detail the early test collections that were used by more than one group and were generally available. Table A.1 shows the general information about each collection: the number of documents, the number of requests/queries, and the original reason for building the collection. Table A.2 shows some statistics on the average document length, the average request/query length, and the average number of relevant documents per query. These are only approximations as different methods of text processing would give different counts.

The test collections fall into two major groups. One group (marked with the *) consists of those built by the SMART group at Harvard and Cornell. These collections used automatic indexing only and the statistics in Table A.2 generally reflect removal of common words, and stemming.

The second group of test collections was built in various places. The statistics in Table A.2 were taken from the sources shown in the table.

Table A.1: General information about each collection

Collection	Year	# Docs	# Questions	Why built
Cran-2	1964	1398	225	Ccompare indexing methods
Cran-1	1964	200	42	Cranfield subset of 42 questions
IRE-3*	1965	780	34	Indexing/dictionary experiments
ADI*	1965	82	35	Document length experiments
ISPRA*	1967	1268	48	Multiple relevance judgements
ISPRA*	1968	1095/468	48	CLIR English/German
MED273*	1967	273	18	Comparison to Lancaster study
Cran424*	1970	424	155	Cornell “reduced” Cranfield subset
MED450*	1970	450	30	“Corrected” Medlars
MEDLARS*	1970	1033	30	Larger document collection
OPHTH.*	1970	853	30	Specific medical domain
TIME*	1970	425	83	Full text articles
INSPEC	1970	542	97	Aitchison indexing experiments
NPL	1970	11,429	93	Clustering experiments
ISILT/Keen	1972	800	63	Keen’s indexing experiments
UKCIS	1974	27,361	75	Chem Abstracts searching experiment
INSPEC	1982	12,684	77	Indexing, Boolean
CACM*	1982	3204	52	Additional metadata
ISI/CISI*	1982	1460	76	Co-citations

Table A.2: General statistics

Collection	Doc.length	Quest.length	# Relevant	Comments
Cran-2	53.6/29.9	8.9/7.9	7.2	Unique automatic terms/manual index terms
Cran-1	74/33	17	4.7	Subset of Cran-2, no source docs (Keen, 1967a)
IRE-3*	49	20	17.4	
ADI*	710/35/7	8	4.9	Text, abstracts, title lengths
ISPRA*	Abstracts	Longer	17.8	Relevant by author
ISPRA*	Aabstracts	Longer	14.2/13.6	Relevant English/German
MED273*	60?	9.3	4.8/11.1	Precision/recall bases
Cran424*	83.4	8	6.4	Fixed # of docs, no source docs
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	Full text documents
INSPEC	12.2	5.6	5.2	Manual index terms (Spärck Jones, 1973b)
NPL	20.0	7.2	22.4	Titles, abstracts (Croft, 1983)
ISILT/Keen	7.2	5.3	14.9	Manually indexed version (Spärck Jones, 1973b)
UKCIS	6.7	18.3	49.9	Titles, manual profiles, only partial judgments (Spärck Jones, 1979b)
INSPEC	36	17.9	33.0	Titles and abstracts (Voorhees, 1985)
CACM*	24.5	10.8	15.3	
ISI/CISI*	46.5	28.3	49.8	

References

Aitchison, T. M. and J. M. Tracy (1970), 'Comparative Evaluation of Index Languages: Part II, Results'. Institution of Electrical Engineers.

Allan, J., J. Callan, W. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu (1998a), 'INQUERY Does Battle With TREC-6'. In: *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*. pp. 169–206.

Allan, J., R. Papka, and V. Lavrenko (1998b), 'On-line New Event Detection and Tracking'. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 37–45.

Bagg, T. C. and M. E. Stevens (1961), *Information Selection Systems Retrieval Replica Copies: a State-of-the-Art Report*, Vol. 157. US Dept. of Commerce, National Bureau of Standards.

Barker, F. H., D. C. Veal, and B. K. Wyatt (1972a), 'Comparative Efficiency of Searching Titles, Abstracts, and Index Terms in a Free-Text Data Base'. *Journal of Documentation* **28**(1), 22–36.

Barker, F. H., B. K. Wyatt, and D. C. Veal (1972b), 'Report on the Evaluation of an Experimental Computer-Based Current-Awareness Service for Chemists'. *Journal of the Association for Information Science and Technology* **23**(2), 85–99.

Barry, C. L. (1994), 'User-defined Relevance Criteria: an Exploratory Study'. *Journal of the American Society for Information Science* **45**(3), 149.

Barry, C. L. and L. Schamber (1998), 'Users' Criteria for Relevance Evaluation: A Cross-Situational Comparison'. *Information Processing and Management* **34**(2/3), 219–236.

Bates, M. J. (1989), 'The Design of Browsing and Berrypicking Techniques for the Online Search Interface'. *Online Review* **13**(5), 407–424.

Belew, R. K. (1989), 'Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents'. In: *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 11–20.

Belkin, N. (1980), 'Anomalous States of Knowledge as a Basis for Information Retrieval'. *The Canadian Journal of Information Science* **5**, 133–143.

Belkin, N., R. Oddy, and H. Brooks (1982), 'ASK for Information Retrieval: Part II. Results of a Design Study'. *Journal of Documentation* **38**, 145–164.

Berger, A. and J. Lafferty (1999), 'Information Retrieval as Statistical Translation'. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Blair, D. C. and M. E. Maron (1985), 'An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System'. *Communications of the ACM* **28**(3), 289–299.

Bohnert, L. M. (1955), 'Two Methods of Organizing Technical Information for Search'. *Journal of the Association for Information Science and Technology* **6**(3), 134–151.

Bookstein, A. and D. R. Swanson (1974), 'Probabilistic Models for Automatic Indexing'. *Journal of the Association for Information Science and Technology* **25**(5), 312–316.

Bourne, C. P. (1961), 'The Historical Development and Present State-of-the-Art of Mechanized Information Retrieval Systems'. *Journal of the Association for Information Science and Technology* **12**(2), 108–110.

Bourne, C. P. (1962), 'The World's Technical Journal Literature: An Estimate of Volume, Origin, Language, Field, Indexing, and Abstracting'. *Journal of the Association for Information Science and Technology* **13**(2), 159–168.

Bourne, C. P. and T. B. Hahn (2003), *A History of Online Information Services, 1963–1976*. MIT press.

Bracken, R. H. and B. G. Oldfield (1956), 'A General System for Handling Alphameric Information on the IBM 701 Computer'. *Journal of the ACM (JACM)* **3**(3), 175–180.

Bracken, R. H. and H. Tillitt (1957), 'Information Searching with the 701 Calculator'. *Journal of the ACM (JACM)* **4**(2), 131–136.

Brin, S. and L. Page (1998), 'The Anatomy of a Large-scale Hypertextual Web Search Engine'. In: *Proceedings of the Seventh International Conference on World Wide Web 7*. pp. 107–117.

Brzozowski, J. (1983), 'MASQUERADE: Searching the Full Text of Abstracts using Automatic Indexing'. *Journal of Information Science* **6**(2-3), 67–73.

Buckland, M. K. (1992), 'Emanuel Goldberg, Electronic Document Retrieval, and Vannevar Bush's Memex'. *Journal of the American Society for Information Science (1986–1998)* **43**(4), 284.

Buckley, C. and A. F. Lewit (1985), 'Optimization of Inverted Vector Searches'. In: *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 97–110.

Buckley, C. and E. M. Voorhees (2000), 'Evaluating Evaluation Measure Stability'. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 33–40.

Burke, C. (1992), 'The Other Memex: The Tangled Career of Vannevar Bush's Information Machine, the Rapid Selector'. *Journal of the American Society for Information Science (1986–1998)* **43**(10), 648.

Bush, V. (1945), 'As We May Think'. *The Atlantic*. pp. 1–26.

Caid, W. R., S. T. Dumais, and S. I. Gallant (1995), 'Learned Vector-Space Models for Document Retrieval'. *Information Processing and Management* **31**(3), 419–429.

Callan, J. P., Z. Lu, and W. B. Croft (1995), 'Searching Distributed Collections with Inference Networks'. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 21–28.

Carbonell, J. and J. Goldstein (1998), 'The Use of MMR, Dversity-Based Reranking for Reordering Documents and Producing Summaries'. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 335–336.

Cerf, V. and R. Kahn (1974), 'A Protocol for Packet Network Intercommunication'. *IEEE Transactions on Communications* **22**(5), 637–648.

Chang, Y., C. Cirillo, and J. Razon (1969), 'Evaluation of Feedback Retrieval using Modified Freezing, Residual Collection and Control Groups'. In: *Scientific Report ISR-16 to NSF*. Cornell University, Ithaca, N.Y., Chapt. X.

Chang, Y., C. Cirillo, and J. Razon (1971), 'Evaluation of Feedback Retrieval using Modified Freezing, Residual Collection and Control Groups'. In: *The SMART Retrieval System*. Prentice-Hall, Englewood Cliffs, New Jersey.

Chuklin, A., I. Markov, and M. d. Rijke (2015), 'Click Models for Web Search'. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **7**(3), 1–115.

Cleverdon, C. (1962), *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Aslib Cranfield Research Project, Cranfield, U.K.

Cleverdon, C. (1970), *The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages*. Cranfield Library Report No. 3, Cranfield, U.K.

Cleverdon, C. (1990), 'Letter to the Editor'. *Online Review* **14**(1), 35.

Cleverdon, C. (1991), '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*. pp. 3–12.

Cleverdon, C. and E. Keen (1966), *Factors Determining the Performance of Indexing Systems, Vol. 2: Test Results*. Aslib Cranfield Research Project, Cranfield, U.K.

Cleverdon, C., J. Mills, and E. Keen (1966), *Factors Determining the Performance of Indexing Systems, Vol. 1: Design*. Aslib Cranfield Research Project, Cranfield, U.K.

Codd, E. F. (1970), 'A Relational Model of Data for Large Shared Data Banks'. **13**(6), 377–387.

Cooper, W. (1968), 'Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems'. *American Documentation* **January** pp. 30–41.

Cooper, W. (1971), 'A Definition of Relevance for Information Retrieval'. *Information Storage and Retrieval* **7**, 19–37.

Cooper, W. S. (1983), 'Exploiting the Maximum Entropy Principle to Increase Retrieval Effectiveness'. *Journal of the Association for Information Science and Technology* **34**(1), 31–39.

Croft, W. B. (1981), 'Document Representation in Probabilistic Models of Information Retrieval'. *Journal of the Association for Information Science and Technology* **32**(6), 451–457.

Croft, W. B. (1983), 'Experiments with Representation in a Document-Retrieval System'. *Information Technology-Research Development Applications* **2**(1), 1–21.

Croft, W. B., T. Lucia, and P. R. Cohen (1988), 'Retrieving Documents by Plausible Inference: a Preliminary Study'. In: *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 481–494.

Croft, W. B. and R. H. Thompson (1987), 'I^{3}R: A New Approach to the Design of Document Retrieval Systems'. *Journal of the American Society for Information Science* **38**(6), 389.

Croft, W. B., H. R. Turtle, and D. D. Lewis (1991), 'The Use of Phrases and Structured Queries in Information Retrieval'. In: *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 32–45.

Cutting, D. R., D. R. Karger, J. O. Pedersen, and J. W. Tukey (1992), 'Scatter/gather: A Cluster-based Approach to Browsing Large Document Collections'. In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 318–329.

Das-Gupta, P. and J. Katzer (1983), 'A Study of the Overlap Among Document Representations'. In: *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 106–114.

Davenport, L. and B. Cronin (1987), 'Marketing Electronic Information'. *Online Review* **11**(1), 39–47.

Dee, C. R. (2007), 'The Development of the Medical Literature Analysis and Retrieval system (MEDLARS)'. *Journal of the Medical Library Association: JMLA* **95**(4), 416.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990), 'Indexing by Latent Semantic Analysis'. *Journal of the American Society for Information Science* **41**(6), 391–407.

Dennis, S. F. (1964), 'The Construction of a Thesaurus Automatically from a Sample of Text'. In: *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation*. pp. 61–148.

Devlin, B. (1983), 'An Overview of Sense-Making Research: Concepts, Methods and Results'. In: *Proceedings of the Annual Meeting of the International Communication Association*. pp. 1–72.

Diriye, A., R. White, G. Buscher, and S. Dumais (2012), 'Leaving So Soon?: Understanding and Predicting Web Search Abandonment Rationales'. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. pp. 1025–1034.

Doszkocs, T. E. (1982), 'From Research to Application: the CITE Natural Language System'. In: *Research and Development in Information Retrieval*. pp. 251–262.

Doszkocs, T. E. and B. A. Rapp (1979), '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*. pp. 131–139.

Edmundson, H. P. and R. E. Wyllys (1961), 'Automatic Abstracting and Indexing Survey and Recommendations'. *Communications of the ACM* **4**(5), 226–234.

Eliot, S. and J. Rose (2009), *A Companion to the History of the Book*, Vol. 98. John Wiley & Sons.

Fidel, R. (1984), 'Online Searching Styles: A Case-Study-Based Model of Searching Behavior'. *Journal of the Association for Information Science and Technology* **35**(4), 211–221.

Fidel, R. (1988), 'Factors Affecting the Selection of Search Keys'. In: *Proceedings of the 51st Annual Meeting of the American Society for Information Science*, Vol. 25, pp. 76–79.

Fidel, R. (1991), 'Searchers' Selection of Search Keys: I. The Selection Routine, II. Controlled Vocabulary or Free-Text Searching, III. Searching Styles'. *Journal of the American Society for Information Science* **42**(7), 490–527.

Fox, E. (1983), *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.

Frakes, W. B. and R. Baeza-Yates (1992), *Information Retrieval: Data Structures and Algorithms*. Prentice Hall PTR.

Fuhr, N. (1989), 'Models for Retrieval with Probabilistic Indexing'. *Information Processing and Management* **25**(1), 55–72.

Fuhr, N. and C. Buckley (1991), 'A Probabilistic Learning Approach for Document Indexing'. *ACM Transactions on Information Systems (TOIS)* **9**(3), 223–248.

Fuhr, N., J. Kamps, M. Lalmas, and A. Trotman (2008), *Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Revised and Selected Papers*, Vol. 4862. Springer.

Furnas, G. W., S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum (1988), 'Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure'. In: *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 465–480.

Gilreath, J. and D. L. Wilson (1989), *Thomas Jefferson's Library: a Catalog with the Entries in his Own Order*. The Lawbook Exchange, Ltd.

Giuliano, V. E. and P. E. Jones (1962), 'Linear Associative Information Retrieval'. Technical report, Arthur D. Little, Inc., Cambridge, Massachusetts.

Gull, C. D. (1987), 'Historical Note: Information Science and Technology: From Coordinate Indexing to the Global Brain'. *Journal of the American Society for Information Science* **38**(5), 338–366.

Gull, D. (1956), 'Seven Years of Work on the Organisation of Materials in a Special Library'. *American Documentation* **7**, 320–329.

Hall, H. and N. Weiderman (1967), 'The Evaluation Problem in Rrelevance Feedback'. In: *Scientific Report ISR-12 to NSF*. Cornell University, Ithaca, N.Y., Chapt. XII.

Hancock-Beaulieu, M. and S. Walker (1992), 'An Evaluation of Automatic Query Expansion in an Online Library Catalogue'. *Journal of Documentation* **48**(4), 406–421.

Harman, D. (1988), 'Towards Interactive Query Expansion'. In: *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 321–331.

Harman, D. (2005), 'The TREC Ad Hoc Experiments'. In: *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, Chapt. 4.

Harman, D. (2011), 'Information Retrieval Evaluation'. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **3**(2), 1–119.

Harman, D. and G. Candela (1990), 'Retrieving Records from a Gigabyte of Text on a Minicomputer using Statistical Ranking'. *Journal of the American Society for Information Science* **41**(8), 581.

Harper, D. J. and C. J. van Rijsbergen (1978), 'An Evaluation of Feedback in Document Retrieval using Co-occurrence Data'. *Journal of Documentation* **34**(3), 189–216.

Harter, S. (1971), 'The Cranfield II Relevance Assessments: a Critical Evaluation'. *Library Quarterly* **41**, 229–243.

Harter, S. P. (1975a), 'A Probabilistic Approach to Automatic Keyword Indexing. Part I. On the Distribution of Specialty Words in a Technical Literature'. *Journal of the American Society for Information Science* **26**(4), 197–206.

Harter, S. P. (1975b), 'A Probabilistic Approach to Automatic Keyword Indexing. Part II. On the Distribution of Specialty Words in a Technical Literature'. *Journal of the American Society for Information Science* **26**(5), 280–289.

Hayes, P. J. and S. P. Weinstein (1990), 'CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories'. In: *IAAI*, Vol. 90. pp. 49–64.

Hearst, M., J. Pedersen, P. Pirolli, H. Schutze, G. Grefenstette, and D. Hull (1996), 'Xerox Site Report: Four TREC Tracks'. In: *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*.

Herlocker, J. L., J. A. Konstan, A. Borchers, and J. Riedl (1999), 'An Algorithmic Framework for Performing Collaborative Filtering'. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 230–237.

Hersh, W., C. Buckley, T. Leone, and D. Hickam (1994), '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.

Holmstrom, J. (1948), 'Section III. Opening Plenary Session'. In: *The Royal Society Scientific Information Conference, 21 June–2 July 1948: Report and papers submitted*.

Ide, E. (1968), 'New Experiments in Relevance Feedback'. In: *Scientific Report ISR-14 to NSF*. Cornell University, Ithaca, N.Y., Chapt. VIII.

Ide, E. (1969), *Relevance Feedback in an Automatic Document Retrieval System*. Scientific Report ISR-15 to NSF.

Ide, E. (1971), 'New Experiments in Relevance Feedback'. In: G. Salton (ed.): *The SMART Retrieval System*. Prentice-Hall, Englewood Cliffs, New Jersey.

Jacobs, P. S. and L. F. Rau (1990), 'SCISOR: Extracting Information from On-Line News'. *Communications of the ACM* **33**(11), 88–97.

Jardine, N. and C. J. van Rijsbergen (1971), 'The Use of Hierarchic Clustering in Information Retrieval'. *Information Storage and Retrieval* **7**(5), 217–240.

Järvelin, K. and J. Kekäläinen (2000), '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.

Kando, N. (1999), 'Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition'. In: *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*.

Kasarda, A. J. and D. J. Hillman (1972), 'The LEADERMART System and Service'. In: *Proceedings of the ACM Annual Conference–Volume 1*. pp. 469–477.

Katzer, J., M. McGill, J. Tessier, W. Frakes, and P. DasGupta (1982), 'Study of the Overlap among Document Representations'. *Information Technology: Research and Development* **1**(2), 261–274.

Keen, E. (1967a), 'Test Environment'. In: *Scientific Report ISR-13 to NSF*. Cornell University, Ithaca, N.Y., Chapt. I.

Keen, E. (1967b), 'Evaluation Parameters'. In: *Scientific Report ISR-13 to NSF*. Cornell University, Ithaca, N.Y., Chapt. II.

Keen, E. (1967c), 'Search Matching Functions'. In: *Scientific Report ISR-13 to NSF*. Cornell University, Ithaca, N.Y., Chapt. III.

Keen, E. (1967d), 'Document Length'. In: *Scientific Report ISR-13 to NSF*. Cornell University, Ithaca, N.Y., Chapt. V.

Keen, E. (1967e), 'Suffix Dictionaries'. In: *Scientific Report ISR-13 to NSF*. Cornell University, Ithaca, N.Y., Chapt. VI.

Keen, E. (1967f), 'Thesaurus, Phrase and Hierarchy Dictionaries'. In: *Scientific Report ISR-13 to NSF*. Cornell University, Ithaca, N.Y., Chapt. VII.

Keen, E. M. (1973), 'The Aberystwyth Index Languages Test'. *Journal of Documentation* **29**(41), 1–35.

Kent, A., M. M. Berry, F. U. Luehrs, and J. W. Perry (1955), 'Machine Literature Searching VIII. Operational Criteria for Designing Information Retrieval Systems'. *Journal of the Association for Information Science and Technology* **6**(2), 93–101.

Kent, A., J. W. Perry, and M. M. Berry (1954), 'Machine Literature Searching V. Definition and Systematization of Terminology for Code Development'. *Journal of the Association for Information Science and Technology* **5**(3), 166–173.

Kilgour, F. G. (1997), 'Origins of Coordinate Searching'. *Journal of the Association for Information Science and Technology* **48**(4), 340–348.

Kleinberg, J. M. (1998), 'Authoritative Sources in a Hyperlinked Environment'. In: *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.

Kleinberg, J. M. (1999), 'Authoritative Sources in a Hyperlinked Environment'. *Journal of the ACM (JACM)* **46**(5), 604–632.

Knuth, D. (1973), *The Art of Computer Programming, Vol. 3: Sorting and Searching*. Addison-Wesley, Reading, MA.

Koenemann, J. and N. J. Belkin (1996), 'A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 205–212.

Kwok, K. L. (1995), 'A Network Approach to Probabilistic Information Retrieval'. *ACM Trans. Inf. Syst.* **13**(3), 324–353.

Lafferty, J. and C. Zhai (2001), 'Document Language Models, Query models, and Risk Minimization for Information Retrieval'. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 403–410.

Lancaster, F. (1969), *Evaluation of the MEDLARS Demand Search Service*. National Library of Medicine, Washington, D.C.

Lancaster, F., R. L. Rapport, and J. Penry (1972), 'Evaluating the Effectiveness of an On-line, Natural Language Retrieval System'. *Information Storage and Retrieval* **8**(5), 223–245.

Lavrenko, V. and W. B. Croft (2001), 'Relevance Based Language Models'. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 120–127.

Lee, D. L. (1990), 'Special Issue on Document Retrieval'. *Quarterly Bulletin of the IEEE Computer Society on Data Engineering* **13**(1), 1–63.

Lesk, M. (1995), 'The Seven Ages of Information Retrieval'. UDT Occassional Paper 5, IFLA.

Lesk, M. (2018), Personal communication.

Lesk, M., D. Harman, E. Fox, and C. Buckley (1997), 'The SMART Lab Report'. *SIGIR Forum* pp. 2–22.

Lesk, M. and G. Salton (1968), 'Relevance Assessments and Retrieval System Evaluation'. In: *Scientific Report ISR-14 to NSF*. Cornell University, Ithaca, N.Y., Chapt. III.

Lesk, M. E. (1969), 'Word-Word Associations in Document Retrieval Systems'. *Journal of the American Society for Information Science* **20**(1), 27–38.

Lesk, M. E. and G. Salton (1969), 'Relevance Assessments and Retrieval System Evaluation'. *Information Storage and Retrieval* **4**, 343–359.

Lewis, D. D. and W. A. Gale (1994), 'A Sequential Algorithm for Training Text Classifiers'. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 3–12.

Lovins, J. B. (1968), 'Development of a Stemming Algorithm'. *Mechanical Translation and Computational Linguistics* **11**(1 and 2), 22–31.

Luhn, H. (1957), 'A Statistical Approach to Mechanized Encoding and Searching of Literary Information'. *IBM Journal* **October**, 309–317.

Luhn, H. P. (1953), 'A New Method of Recording and Searching Information'. *American Documentation* **4**(1), 14–16.

Luhn, H. P. (1958a), 'The Automatic Creation of Literature Abstracts'. *IBM Journal of Research and Development* **2**(2), 159–165.

Luhn, H. P. (1958b), 'A Business Intelligence System'. *IBM Journal of Research and Development* **2**(4), 314–319.

Luhn, H. P. (1960), 'Key Word-in-Context Index for Technical Literature (KWIC Index)'. *American Documentation* **11**(4), 288–295.

Luhn, H. P. (1961), 'Selective Dissemination of New Scientific Information with the Aid of Electronic Processing Equipment'. *Journal of the Association for Information Science and Technology* **12**(2), 131–138.

Maron, M. and J. Kuhns (1960), 'On Relevance, Probabilistic Indexing and Information Retrieval'. *Journal of the ACM* **7**, 216–244.

Maron, M., J. Kuhns, and L. Ray (1959), 'Probabilistic Indexing. a Statistical Technique for Document Identification and Retrieval'. Technical report, Thomson Ramo Woolridge, Inc., Los Angeles, CA.

Maron, M. E. (1961), 'Automatic Indexing: an Experimental Inquiry'. *Journal of the ACM (JACM)* **8**(3), 404–417.

Maron, M. E. (2008), 'An Historical Note on the Origins of Probabilistic Indexing'. *Information Processing & Management* **44**(2), 971–972.

McCarn, D. B. (1971), 'Networks with Emphasis on Planning an On-Line Bibliographic Access System'. *Information Storage and Retrieval* **7**(6), 271–279.

McCarn, D. B. (1978), 'Online Systems-Techniques and Services'. *Annual Review of Information Science and Technology* **13**, 85–124.

McGill, M. J., L. C. Smith, S. Davidson, and T. Noreault (1976), 'Syracuse Information Retrieval Experiment (SIRE): Design of an On-line Bibliographic Retrieval System'. *SIGIR Forum* **10**(4), 37–44.

Miller, W. L. (1971), 'A Probabilistic Search Strategy for MEDLARS'. *Journal of Documentation* **27**(4), 254–266.

Mitev, N., G. Venner, and S. Walker (1985), *Designing an Online Public Acess Catalogue: Okapi, a Catalogue on a Local Area Network*. Library and Information Research Report 39, London: British Library.

Mooers, C. N. (1950), 'Information Retrieval Viewed as Temporal Signaling'. In: *Proceedings of the International Congress of Mathematicians*, Vol. 1. pp. 572–573.

Mooers, C. N. (1951), 'Zatocoding applied to mechanical organization of knowledge'. *Journal of the Association for Information Science and Technology* **2**(1), 20–32.

Mooers, C. N. (1960), 'The Next Twenty Years in Information Retrieval; Some Goals and Predictions'. *Journal of the Association for Information Science and Technology* **11**(3), 229–236.

Mothe, J. (1994), 'Search Mechanisms Using a Neural Network Model: Comparison with the Vector Space Model'. In: *Intelligent Multimedia Information Retrieval Systems and Management – Volume 1*. pp. 275–294.

Needham, E. and K. Spärck Jones (1964), 'KEYWORDS AND CLUMPS: Recent work on Information Retrieval at the Cambridge Language Research Unit'. *Journal of Documentation* **20**(1), 5–15.

Oddy, R. (1977), 'Information Retrieval through Man and Machine Dialog'. *Journal of Documentation* **33**(1), 1–14.

Page, L., S. Brin, R. Motwani, and T. Winograd (1998), *The PageRank Citation Ranking: Bringing Order to the Web.*

Pejtersen, A. M. (1989), 'A Library System for Information Retrieval based on a Cognitive Task Analysis and Supported by an Icon-Based Interface'. In: *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 40–47.

Peters, C. (2003), *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21–22, 2000, Revised Papers*, Vol. 2069.

Ponte, J. M. and W. B. Croft (1998), 'A Language Modeling Approach to Information Retrieval'. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 275–281.

Porter, M. F. (1980), 'An Algorithm for Suffix Stripping'. *Program* **14**(3), 130–137.

Porter, M. F. (1982), 'Implementing a Probabilistic Information Retrieval System'. *Information Technology: Research and Development* **1**(2), 131–156.

Rath, G., A. Resnick, and T. Savage (1961), 'Comparisons of Four Types of Lexical Indicators of Content'. *Journal of the Association for Information Science and Technology* **12**(2), 126–130.

Rau, L. F. (1987), 'Knowledge Organization and Access in a Conceptual Information System'. *Information Processing and Management* **23**(4), 269 – 283.

Rees, A. (1967), 'Evaluation of Information Systems and Services'. In: *Annual Review of Information Science and Technology*. Interscience, Chapt. 3.

Rees, J. and A. Kent (1958), 'Mechanized Searching Experiments using the WRU Searching Selector'. *Journal of the Association for Information Science and Technology* **9**(4), 277–303.

Rice, B. A. (1985), 'Evaluation of Online Databases and Their Uses in Collections'. *Library Trends* **33**(3), 297–325.

Robertson, S. (1969a), 'The Parametric Description of Retrieval Tests: Part I: The Basic Parameters'. *Journal of Documentation* **25**(1), 1–27.

Robertson, S. (1969b), 'The Parametric Description of Retrieval Tests: Part II: Overall Measures'. *Journal of Documentation* **25**(2), 93–103.

Robertson, S. (2003), 'The Unified Model Revisited'. In: *SIGIR 2003 Workshop on Mathematical/Formal Models in Information Retrieval*.

Robertson, S. and J. Callan (2005), 'Routing and Filtering'. In: *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, Chapt. 5.

Robertson, S. and K. Spärck Jones (1976), 'Relevance Weighting of Search Terms'. *Journal of the American Society for Information Science* **27**(3), 129–146.

Robertson, S. E. (1977a), 'The Probability Ranking Principle in IR'. *Journal of Documentation* **33**(4), 294–304.

Robertson, S. E. (1977b), 'Theories and Models in Information Retrieval'. *Journal of Documentation* **33**(2), 126–148.

Robertson, S. E. and M. M. Hancock-Beaulieu (1992), 'On the Evaluation of IR Systems'. *Information Processing and Management* **28**(4), 457–466.

Robertson, S. E., M. Maron, and W. S. Cooper (1982), 'The Unified Probabilistic Model for IR'. In: *Research and Development in Information Retrieval*. pp. 108–117.

Robertson, S. E., C. J. van Rijsbergen, and M. F. Porter (1980), 'Probabilistic Models of Indexing and Searching'. In: *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*. pp. 35–56.

Robertson, S. E. and S. Walker (1994), 'Some Simple Effective Approximations to the 2-poisson Model for Probabilistic Weighted Retrieval'. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 232–241.

Rocchio, J. (1965), 'Relevance Feedback in Information Retrieval'. In: *Scientific Report ISR-9 to NSF*. Harvard University, Cambridge, Massachusetts, Chapt. XXIII.

Rocchio, J. J. (1966), *Document Retrieval Systems – Optimization and Evaluation*. Scientific Report ISR-10 to NSF, Cambridge, Massachusetts.

Rocchio, J. J. (1971), 'Relevance Feedback in Information Retrieval'. In: G. Salton (ed.): *The SMART Retrieval System*. Prentice-Hall, Englewood Cliffs, New Jersey.

Salton, G. (1963), 'Associative Document Retrieval Techniques Using Bibliographic Information'. *Journal of ACM* **10**(4), 440–457.

Salton, G. (1964a), 'The Evaluation of Automatic Retrieval Procedures– selected test results using the SMART system'. In: *Scientific Report ISR-8 to NSF*. Harvard University, Cambridge, Massachusetts, Chapt. IV.

Salton, G. (1964b), 'A Flexible Automatic System for the Organization, Storage and Retrieval of Language Data'. In: *Scientific Report ISR-5 to NSF*. Harvard University, Cambridge, Massachusetts, Chapt. I.

Salton, G. (1964c), *Scientific Report ISR-7 to NSF*. Harvard University, Cambridge, Massachusetts.

Salton, G. (1965), 'The Evaluation of Automatic Retrieval Procedures– Selected Test Results Using the SMART System'. *American Documentation* **16**(3), 209–222.

Salton, G. (1965), *Scientific Report ISR-9 to NSF*. Harvard University, Cambridge, Massachusetts.

Salton, G. (1967), *Scientific Report ISR-9 to NSF*. Harvard University, Cambridge, Massachusetts.

Salton, G. (ed.) (1968), *Automatic Information Organization and Retrieval*. McGraw-Hill Book Co., New York, N.Y.

Salton, G. (1969a), 'Automatic Processing of Foreign Language Documents'. In: *Proceedings of the 1969 Conference on Computational Linguistics*. pp. 1–28, Association for Computational Linguistics.

Salton, G. (1969b), 'A Comparison between Manual and Automatic Indexing Methods'. *Journal of the Association for Information Science and Technology* **20**(1), 61–71.

Salton, G. (ed.) (1971), *The SMART Retrieval System*. Prentice-Hall, Englewood Cliffs, New Jersey.

Salton, G. and C. Buckley (1990), 'Improving Retrieval Performance by Relevance Feedback'. *Journal of the American Society for Information Science* **41**(4), 288–297.

Salton, G., E. A. Fox, and H. Wu (1983), 'Extended Boolean Information Retrieval'. *Communications of the ACM* **26**(11), 1022–1036.

Salton, G. and M. E. Lesk (1965), 'The SMART Automatic Document Retrieval Systems: an Illustration'. *Communications of the ACM* **8**(6), 391–398.

Salton, G. and M. E. Lesk (1968), 'Computer Evaluation of Indexing and Text Processing'. *Journal of the ACM (JACM)* **15**(1), 8–36.

Salton, G. and M. McGill (eds.) (1983), *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York, NY.

Salton, G. and D. Williamson (1968), 'A Comparison Between Manual and Automatic Indexing Methods'. In: *Scientific Report ISR-14 to NSF*. Cornell University, Ithaca, N.Y., Chapt. VI.

Salton, G. and C. Yang (1973), 'On the Specification of Term Values in Automatic Indexing'. *Journal of Documentation* **29**(4), 351–372.

Salton, G., C.-S. Yang, and C. T. Yu (1974), 'A Theory of Term Importance in Automatic Text Analysis'. In: *Scientific Report ISR-22 to NSF*. Cornell University, Ithaca, N.Y., Chapt. III.

Salton, G., C.-S. Yang, and C. T. Yu (1975), 'A Theory of Term Importance in Automatic Text Analysis'. *Journal of the Association for Information Science and Technology* **26**(1), 33–44.

Sanderson, M. (2010), 'Test Collection Based Evaluation of Information Retrieval Systems'. *Foundations and Trends in Information Retrieval* **4**, 247–375.

Sanderson, M. and W. B. Croft (2012), 'The History of Information Retrieval Research'. *Proceedings of the IEEE* **100**(Special Centennial Issue), 1444–1451.

Saracevic, T. (1968), 'Linking Research and Teaching'. *American Documentation* **October**, pp. 398–403.

Saracevic, T. (1971), 'Selected Results from an Inquiry into Testing of Information Retrieval Systems'. *Journal of the American Society for Information Science* **March–April**, pp. 126–139.

Saracevic, T. (2007a), '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.

Saracevic, T. (2007b), '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.

Saracevic, T. and P. Kantor (1988a), 'A Study of Information Seeking and Retrieving. II: Searchers, Searches, and Overlap'. *Journal of the American Society for Information Science* **39**(2), 197–216.

Saracevic, T. and P. Kantor (1988b), 'A Study of Information Seeking and Retrieving. II: Users, Questions, and Effectiveness'. *Journal of the American Society for Information Science* **39**(2), 177–196.

Saracevic, T., P. Kantor, A. Y. Chamis, and D. Trivison (1988), 'A Study of Information Seeking and Retrieving. I: Background and Methodology'. *Journal of the American Society for Information Science* **39**(2), 161–176.

Segesta, J. and K. Reid-Green (2002), 'Harley Tillitt and Computerized Library Searching'. *IEEE Annals of the History of Computing* **24**(3), 23–34.

Seymour, T., D. Frantsvog, and S. Kumar (2011), 'History of Search Engines'. *International Journal of Management & Information Systems* **15**(4), 47–58.

Shera, J. H. (1955), 'The Truth, the Whole Truth...'. *American Documentation* **6**, 56.

Siegel, E., K. Kameen, S. Sinn, and F. O. Weise (1984), 'Research Strategy and Methods used to Conduct a Comparative Evaluation of Two Prototype Online Catalog Systems'. In: *Proceedings of the National Online Meeting.* pp. 503–511.

Singhal, A., C. Buckley, and M. Mitra (1996), 'Pivoted Document Length Normalization'. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* pp. 21–29.

Smeaton, A. F., P. Over, and W. Kraaij (2006), 'Evaluation Campaigns and TRECvid'. In: *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. pp. 321–330.

Smith, E. S. (1993), 'On the Shoulders of Giants: From Boole to Shannon to Taube; The Origins and Development of Computerized Information from the Mid-19th Century to the Present'. *Information Technology and Libraries* **12**(2), 217.

Spärck Jones, K. (1972), 'A Statistical Interpretation of Term Specificity and its Application in Retrieval'. *Journal of Documentation* **60**(5), 493–502.

Spärck Jones, K. (1973a), 'Collection Properties Influencing Automatic Term Classifications Performance'. *Information Storage and Retrieval* **9**, 499–513.

Spärck Jones, K. (1973b), 'Index Term Weighting'. *Information Storage and Retrieval* **9**(11), 619–633.

Spärck Jones, K. (1975), 'A Performance Yardstick for Test Collections'. *Journal of Documentation* **31**(4), 266–272.

Spärck Jones, K. (1979a), 'Experiments in Relevance Weighting of Search Terms'. *Information Processing and Management* **15**(3), 133–144.

Spärck Jones, K. (1979b), 'Search Term Relevance Weighting Given Little Relevance Information'. *Journal of Documentation* **35**(1), 30–48.

Spärck Jones, K. (ed.) (1981), *Information Retrieval Experiment*. Butterworths.

Spärck Jones, K. (1988), 'A Look Back and a Look Forward'. In: *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 13–29.

Spärck Jones, K. (2000), 'Further Reflections on TREC'. *Information Processing and Management* **36**(1), 37–86.

Spärck Jones, K. (2004), 'IDF Term Weighting and IR Research Lessons'. *Journal of documentation* **60**(5), 521–523.

Spärck Jones, K. and D. Jackson (1970), 'The Use of Automatically-Obtained Keyword Classifications for Information Retrieval'. *Information Storage and Retrieval* **5**, 175–201.

Spärck Jones, K. and R. M. Needham (1968), 'Automatic Term Classifications and Retrieval'. *Information Storage and Retrieval* **4**(2), 91–100.

Spärck Jones, K. and C. van Rijsbergen (1976), 'Information Retrieval Test Collections'. *Journal of Documentation* **32**(1), 59–75.

Stanfill, C. and B. Kahle (1986), 'Parallel Free-Text Search on the Connection Machine System'. *Communications of the ACM* **29**(12), 1229–1239.

Stanfill, C. and D. Waltz (1986), 'Toward Memory-based Reasoning'. *Communications of the ACM* **29**(12), 1213–1228.

Stevens, M. E. (1970), *Automatic Indexing: A State-of-the-Art Report.*

Stevens, M. E. and V. E. Giuliano (1965), *Statistical Association Methods for Mechanized Documentation: Symposium Proceedings, Washington, 1964*, Vol. 269. US Government Printing Office.

Stiles, H. E. (1961), 'The Association Factor in Information Retrieval'. *Journal of the ACM (JACM)* **8**(2), 271–279.

Summit, R. K. (1989), 'In Search of the Elusive End User'. *Online Review* **13**(6), 485–491.

Swanson, D. (1971), 'Some Unexplained Aspects of the Cranfield Tests of Indexing Language Performance'. *Library Quarterly* **41**, 223–228.

Swanson, D. R. (1960), 'Searching Natural Language Text by Computer'. *Science* **132**(3434), 1099–1104.

Swets, J. A. (1969), 'Effectiveness of Information Retrieval Methods'. *American Documentation* **January**, pp. 72–89.

Taube, M. and Associates (1955), 'Storage and Retrieval of Information by Means of the Association of Ideas'. *Journal of the Association for Information Science and Technology* **6**(1), 1–18.

Taube, M., C. Gull, and I. S. Wachtel (1952), 'Unit terms in coordinate indexing'. *Journal of the Association for Information Science and Technology* **3**(4), 213–218.

Tenopir, C. (1984), 'Full-text Databases'. *Annual Review of Information Science and Technology* **19**, 215–246.

Turtle, H. (1994), 'Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance'. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 212–220.

Turtle, H. and W. B. Croft (1989), 'Inference Networks for Document Retrieval'. In: *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 1–24.

Turtle, H. and W. B. Croft (1991a), 'Evaluation of an Inference Network-Based Retrieval Model'. *ACM Transactions on Information Systems (TOIS)* **9**(3), 187–222.

Turtle, H. R. and W. B. Croft (1991b), 'Efficient Probabilistic Inference for Text Retrieval'. In: *Intelligent Text and Image Handling – Volume 2*. pp. 644–661.

US Public Health Service *et al.* (1963), 'The MEDLARS Story at the National Library of Medicine'. *Washington, D.C., Government Printing Office.*

van Rijsbergen, C. J. (1973), 'Further Experiments with Hierarchic Clustering in Document Retrieval'. *Information Storage and Retrieval* **10**(1), 1–14.

van Rijsbergen, C. J. (1975), *Information Retrieval.* Butterworths.

van Rijsbergen, C. J. (1977), 'A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval'. *Journal of Documentation* **33**(2), 106–119.

van Rijsbergen, C. J. (1986), 'A New Theoretical Framework for Information Retrieval'. In: *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* pp. 194–200.

Varlejs, J. (1999), 'Ralph Shaw and the Rapid Selector'. In: *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems.* pp. 148–155.

Vaswani, P. and J. Cameron (1970), *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.

Voorhees, E. and D. Harman (eds.) (2005), *TREC: Experiment and Evaluation in Information Retrieval.* The MIT Press.

Voorhees, E. M. (1985), 'The Cluster Hypothesis Revisited'. In: *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* pp. 188–196.

Voorhees, E. M. (1998), '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.

Walker, S. and R. de Vere (1989), *Improving Subject Retrieval in Online Catalogues: 2. Relevance Feedback and Query Expansion.* British Library Research Paper 72, London: British Library.

Walker, S. and M. Hancock-Beaulieu (1991), *Okapi at City, an Evaluation Facility for Interactive IR.* British Library Research Report 6056, London: British Library.

Walker, S. and R. M. Jones (1987), *Improving Subject Retrieval in Online Catalogues: 1. Stemming, Automatic Spelling Correction and Cross-Reference Tables.* British Library Research Paper 24, London: British Library.

Willett, P. (2006), 'The Porter Stemming Algorithm: Then and Now'. *Program* **40**(3), 219–223.

Williams, M. E. (1985a), 'Electronic Databases'. *Science* **228**(4698), 445–450.

Williams, M. E. (1985b), 'Usage and Revenue Data for the Online Database Industry'. *Online Review* **9**(3), 205–210.

Williams, M. E. (1986), 'Online Government Databases–An Analysis'. *Online Review* **10**(4), 227–36.

Williamson, D. (1968), 'A Cornell Implementation of the SMART System'. In: *Scientific Report ISR-14 to NSF*. Cornell University, Ithaca, N.Y., Chapt. VII.

Wolfe, G. (1994), 'The Second Phase of the Revolution has Begun'. *Wired* **2**(10), 116–121.

Xu, J. and W. B. Croft (1996), 'Query Expansion Using Local and Global Document Analysis'. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 4–11.

Zhai, C. and J. Lafferty (2001), 'A Study of Smoothing Methods for Language Models Applied to Information Retrieval'. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 334–342.

Zobel, J. (1998), '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.