Foundations and Trends® in Information Retrieval
Vol. XX, No. XX (2017) 1–60

DOI: 10.1561/XXXXXXXXXX

now
the essence of knowledge

Offline Evaluation for Information Retrieval

Jin Young Kim
Microsoft
jink@microsoft.com

Emine Yilmaz
University College London
emine.yilmaz@ucl.ac.uk

Paul Thomas
Microsoft
pathom@microsoft.com

Contents

Abstract

Offline evaluation characterizes an information retrieval (IR) system without relying on actual users in a real-world environment. Offline evaluation, notably test collection based evaluation, has been the dominant approach in IR evaluation and it is no exaggeration to say that shared evaluation efforts such as the TREC conferences have defined IR research over the years. The reason for this success lies in the ability to compare retrieval systems in a reusable manner.

Several recent trends however necessitate a change in the role and methods of offline evaluation. First and foremost, online search engines with large-scale user base has become commonplace, enabling online evaluation based on user behavior There are new endpoints for search, such as mobile phones and conversational agents, and the types of search results has diversified beyond a list of web documents to include other result types. Finally, crowdsourcing has provided ways for human judgments of any kind to be collected at a large scale. However, online evaluation based on user behavior has its own challenges due to repeatability as well the extensive amount of time needed to get online evaluation signals from the users. Furthermore, most smaller companies and academic researchers do not have access to such large scale user base. Hence, recent research in IR evaluation has focused on the advent of new offline evaluation paradigms which are more user-centric, diverse and agile.

⋆⋆ This survey aims to provide an overview of recent research in IR evaluation pertaining to the trends above, covering latest developments since the last comprehensive survey Sanderson [2010]. We first introduce offline evaluation for IR, focusing on how it relates to other evaluation paradigms such as online evaluation. We also overview traditional offline evaluation for IR, and how recent trends have shaped the research so far. We then review research in offline evaluation on three levels: human judgments, evaluation metrics and experiment design. This organization will allow readers to follow recent developments in research from micro-level (human judgment) to macro-level (experiment). Finally, we discuss evaluation practice in industry, which has been a major driving force in research and development in IR.

Maarten: Explain how your survey relates to Mark SandersonâĂŹs survey and why we need your survey now.
Jin: done

now Publishers Inc.. *Offline Evaluation for Information Retrieval*. Foundations and Trends® in Information Retrieval, vol. XX, no. XX, pp. 1–60, 2017.

DOI: 10.1561/XXXXXXXXXX.

1

Introduction

In this chapter, we survey the area and lay conceptual foundations for the rest of the paper. We first provide an overview of different approaches to IR evaluation. We then focus on offline evaluation, explaining traditional approaches and recent trends. Finally, we introduce a conceptual framework and the outline for the rest of this paper.

1.1 Evaluation Paradigms in IR

Evaluating a search system, or any system that supports information access such as recommendation or filtering, is a complex problem. The performance of a search system is dependent on various contextual factors, such as the task at hand, the user's preference, abilities, location and other characteristics, and even the timing of the interaction. Also, the ultimate source of ground truth, the user's judgment, is subjective, volatile, and often hard to come by.

1.1.1 Offline vs. Online Evaluation

In order to meet these challenges, IR researchers have built a rich evaluation tradition. Most of this work has been based on a few simplifying

assumptions. The document collection is static and the user's information need is represented as a description or a keyword query. The user's judgments in situ are replaced with judgments collected post-hoc and from third parties, often in the form of binary or numeric-scale labels.

We can define this evaluation paradigm as *offline evaluation* [Sanderson, 2010] in that the evaluation of the system can happen without requiring an actual user. This makes offline evaluation particularly suitable for early-stage evaluation of an IR system, when users are hard to come by. Another typical characteristic of offline evaluation is that the test collection (a set of tasks, judgments and documents) is 'reusable', in that once built it can be used to evaluate new systems; because many factors are controlled, evaluations are also commensurable across time and between researchers.

An evaluation paradigm contrasting with offline evaluation is called *online evaluation*. In a recent survey on this topic, online evaluation is defined as the evaluation of a fully functioning system based on measurement of real users' interactions with the system in a natural environment [Hofmann et al., 2016]. That is, online evaluation directly employs user behavior in natural environment for evaluation.

As large-scale online services are commonplace now, online evaluation has become a viable option for companies with running services with large user bases. In the literature, there has been a plethora of papers on methodologies for online evaluation. While online evaluation has benefits in using data readily available as a by-product of serving users, this dependence on user behavior also creates limitations for online evaluation, which we will discuss later in this section. ⋆

Maarten: The relation between online and offline evaluation needs a more thorough treatment. The discussion would have to include correlations between offline metrics and online metrics.

Now, let us compare two evaluation paradigms – offline and online evaluation. Table 1.1 summarizes the advantages and disadvantages of online vs. offline evaluation.

Offline evaluation typically requires access to explicit judgments of relevance obtained from relevance assessors, or judges. Obtaining judgments is an expensive procedure; hence, online evaluation tends to be cheaper compared to offline evaluation. Furthermore, online evaluation is based on signals that directly come from real users, which can enable us to get a more realistic signal of user satisfaction.

Table 1.1: Pros and cons of offline vs. online evaluation

	Online	Offline
Pros	Very little marginal cost Based on actual user behavior	No need for production system Easy to try new ideas Amortized cost / reusability
Cons	Need for production system Need for large-scale data Noisy interpretation of behavior	High marginal cost of label collection Need for judging infrastructure Need for 'ground truth' judgments Difficult to model real users behavior

On the other hand, online evaluation requires a running system as it is based on signals from real users. First, in initial stages of system development we simply might not have real users to study. Furthermore, small companies and academics may not have access to a large volume of users to be able to collect reliable signals. On the other hand, with the availability of various crowdsourcing services, it is relatively easy to collect labels from human judges.

Another major problem in online evaluation is that usually a significant amount of usage data is needed before one can reach reliable signals of satisfaction. ⋆ Hence, online evaluation tends to be very slow, which may not be suitable for evaluating the quality of new methodologies quickly.

Maarten: please qualify, see early work by Joachims

More importantly, signals obtained from real users tend to be noisy and traces of user behavior are often insufficient to measure a user's true satisfaction. As an example, let's take clicks on results for evaluating a search engine. While a click is certainly an indication that the user is interested in the result, it is not clear whether the clicked result was actually satisfying. Also, clicks are often concentrated on the top of the page regardless of result quality making them difficult to interpret. All in all, the ambiguity and bias inherent in user behavior often make it hard to infer the true quality of search engine. ⋆

Maarten: What is "true quality"? In the eyes of the users? Of the judges? Of the managers?

Another consideration is the reusability of the data collected. In offline evaluation, typically the label is collected at the level of individual information item (i.e., document) and the system is evaluated by its ability to put more relevant items on top. This means the labels can be reused to evaluate new systems that produce different rankings of

the same items. By contrast, the data collected from online system is typically valid for the evaluation of the system user interacted with, although there are new research to address these issues.

Offline evaluation, on the other hand, tends to be fast once explicit judgments of relevance are obtained from relevance assessors. Once these judgments are collected, they can be used to evaluate the quality of sytems quickly. This makes offline evaluation very suitable for trying new ideas, and the initial cost can be amortised over many experiments.

One major drawback of offline evaluation is the expense of collecting these explicit judgments, or the ground truth. Obtaining relevance judgments can be slow and expensive, and has to be repeated if the notion of relevance changes. Furthermore, these explicit judgments of relevance tends to come from a third-party assessor, as opposed to the real user of the system. Hence, the assessor may have a different understanding than an actual user as to what documents should be considered relevant. Finally, depending on domain (i.e., medical or engineering) or task (i.e., personalized search), it is difficult to find capable assessors. ⋆⋆

Maarten: plus difficulties of getting expert labels in some domains/tasks (personalized search) plus dynamics of relevance
Jin: added

Finally, offline evaluation metrics tend to be based on *models* of user behavior, as opposed to behavior signals obtained from real users and modeling users can be quite challenging due to the variance in behavior and expectations of real users. Hence, evaluation metrics based on user models may not necessarily reflect user satisfaction. Much recent work in offline evaluation focuses on this issue, which we will review later in Chapter 2.

1.1.2 Hybrid Approaches

So far we have compared two evaluation paradigms – offline and online evaluation – with distinctive characteristics. Offline evaluation is based on human judges as substitution of real users, and has strengths in experimental control and reusability. Online evaluation is based on user behavior, and has strengths in fidelity and cost.

While these two approaches comprise the majority of evaluation efforts, there have been several approaches trying to find a middle ground. Click modeling [Chuklin et al., 2015] and counterfactual online

evaluation [Li et al., 2015, 2010], for example, re-use online user data for future evaluation. These approaches, while enabling the re-use of online user data, are still limited in that they are based on implicit signals from user behavior. For instance, it is not trivial to decide whether a user indeed found the clicked document relevant or not, even with all the contextual information.

⋆ ⋆

Another related line of work is *user study-based evaluation* [Bron et al., 2013, Liu et al., 2014, Shah and González-Ibáñez, 2011], which is widely used in interactive IR studies [Kelly, 2009]. In such work, a group of participants are typically brought into a lab environment and asked to perform a set of (usually predetermined) search tasks. It is common for this type of study to collect both behavior and labels from the participants to get a more complete picture of search activity.

Maarten: Not sure. You want to make this argument. The exact same thing can/should be said about expert judges. They are not not real users so their judgments aren't necessarily an indication that users will find a document relevant/useful.

Jin: The argument has been softened a bit

User studies bear similarities with offline evaluation in that they typically involve some form of explicit judgments, but their emphasis is more on understanding some aspect of users' search behavior, as opposed to comparative evaluation among search systems. Also, user studies tend to be limited in scale (typically less than 100 participants) and based on a biased, possibly not representative, sample of participants (typically people within the same institution).

⋆ ⋆

However, the distinctions are getting blurred as search engines increasingly serve more complex results, and SERP (search engine results page) or session-level evaluation is drawing more attention. In fact, some recent research has tried to employ task completion environment with human subjects for system-to-system comparison [Xu and Mease, 2009]. Also, crowdsourcing techniques are reducing barriers in getting access to a large number of subjects with diverse backgrounds. We will return to this point in Chapter 2.

Maarten: but then TREC is even worse off, just a handful of assessors...

Jin: Crowdsourcing can help here...

1.1.3 Combining Approaches for Evaluation

Given a variety of options for evaluation – online, offline and even hybrid ones, one may be confused to choose which one. However, the very existence of these different approaches predicate the multifaceted

nature of the problem. Then, it makes sense to combine approaches for evaluation in order to get a full picture of the quality. Here we describe two approaches in combining multiple approaches in evaluation.

1.1.4 Funnel Approach for Combination

In fact, it has been a common practice for IR evaluation in industry settings to combine offline and online evaluation in sequence. Since the number of algorithms approved for final deployment is many fewer than the algorithm sent for initial test, this whole process is sometimes called *evaluation funnel*.

The process starts with a set of new ranking techniques which are candiates for deployment. First, they are evaluated against existing ranking technique (baseline) using an offline evaluation method. This ensures that the new techniques meet the minimal quality bar before they are exposed to users, and the cost of evaluation at this can be kept low if the labels can be reused across evaluations.

Once we have a subset of algorithms which passes the bar using offline evaluation, they are ready for online evaluation, which can be done by showing them to the user in a controlled experiment setting. This step ensures that the new algorithm does make positive user impact measured in the gain of various online meatrics, which then would be 'shipped' to the users.

Now, for this funnel approach to work, it is important that the evaluation results earlier at the funnel (offline evaluation) should be show reasonable agreement with the results at the later stages (online evaluation). Ideally, the offline evaluation results should be a lenient filter which includes all the techniques with positive online evaluation results, so that it can reduce the number of algorithms sent for online evaluation.

For this reason, understanding the relationship between online and offline evaluation methods is important, and there has been several studies examining this issue. Huffman and Hochster [2007] examines the relationship per-query relevance measures and session-level user satisfaction, finding that the relationship is quite strong, and that including the session-level information makes it even stronger.

Radlinski et al. [2008] showed paired experimental design for online evaluation (i.e., by interleaving two ranked lists) gives reasonable agreement with retrieval quality. Radlinski and Craswell [2010] further develops this idea to compare the sensitivity between metrics, showing that offline evaluation based on 5,000 judges queries have sensitivity equivalent to 50,000 user impressions.

1.1.5 User Modeling Approach for Combination

'Funnel approach' above combines offline and online evaluation in sequence to find a set of ranking techniques that satisfies both. Alternatively, one can build a metric based on some model of user to capture the elusive concept of user satisfaction, and the combining data from online (user) and offline (judges) is often the key in building such user model.

For offline evaluation, online user behavior can inform the various parameters of the metric so that the resulting metric values better reflect user satisfaction. Recent work on offline evaluation metrics has embraced online user data to tune parameters of the metrics [Carterette et al., 2011, 2012, Smucker and Clarke, 2012, Yilmaz et al., 2010, for example].

For online evaluation, labels from human judges (or from user themselves, if available) can be used to build a model of successful search session. [Hassan et al., 2010, Hassan, 2012] built a model of satisfaction based on human annotation of user sessions, or in-situ feedback from user, respectively. Ageev et al. [2011] built a game-style interface where the goal is to perform a search task and then rate the experience, and then built a model of success which predicts the ratings given a sequence of behavior.

1.1.6 Summary

So far we have looked at various approaches in IR evaluation. Online evaluation based on user data and offline evaluation based on human judgments are two dominant approaches, and there many approaches which has the characteristics of both. Given these approaches with different characteristics, it makes sense to combine a few methods in many

cases. In summary, here is our list of recommendations for deciding on which approach, or combination of approaches to use.

1. Start with offline evaluation when usage data is not available, or the amount of user data is not sufficient to draw meaningful conclusions.

2. Even if sufficient usage data is available, consider offline evaluation if you want to impose certain policy on evaluation results, or if there a certain area which online evaluation cannot cover.

3. Consider user study if certain aspect of user behavior needs to be better understood, or detailed feedback of user experience is required.

4. Consider combining both offline and online evaluation approaches in funnel, or building a user model which leverage both data sources.

1.2 General Framework for Offline Evaluation

We have outlined major evaluation paradigms for IR evaluation so far. The goal of this paper is to provide a practical guide to conducting offline evaluation for both academic and industry practitioners. Here we outline two major scenarios which we cover in this paper.

In traditional IR research, a typical evaluation scenario is to improve the performance of a document retrieval system given a test collection and a pre-determined set of evaluation metrics. For instance, in the TREC Web Track, participants are given a collection of web documents, and then asked to submit the results for their systems in a designated format. These are then evaluated on metrics like NDCG [Järvelin and Kekäläinen, 2002] or ERR [Chapelle et al., 2009].

While academic IR research has developed well-accepted offline evaluation practices for document retrieval based on explicit labels, there are many evaluation scenarios not addressed in research from a practitioners' standpoint. There are multiple components in a modern IR system such as a web search engine, and designing evaluation for

each requires different emphases and considerations. For instance, evaluating a query suggestion system can be quite different from evaluating a document ranking system.

Also, building a working system serving a large number of real users takes several stages of development. The evaluation at early stages of development would be more exploratory in nature, whereas at later stage the focus would shift to making ship decisions. We can call the former *information-centric* evaluation in that the goal is to collect information helpful for system development and debugging, where the latter can be considered *number-centric* in that the goal is to get reliable performance numbers for decision making.

Another characteristics of IR evaluation in industry settings is that the evaluation is an on-going process which takes multiple iterations over the lifetime of the service, as opposed to a one-off research project. This necessitates the development of so called *evaluation pipelines* where any new system can be evaluated on a ongoing basis.

Since the goal of this paper is to meet the need of practitioners as well as academic researchers, we describe decisions one needs to face in conducting offline evaluation across various scenarios outlined above. We also focus on considerations in designing a evaluation pipeline in industry settings at Chapter 5.

Dealing with evaluation problems across many scenarios requires a general framework of thinking. For the rest of this chapter, we introduce definitions and general process in offline evaluation which will constitute the framework.

1.2.1 Definitions

First, here are a few definitions that will be used throughout this paper. These comprise the components of offline evaluation.

Evaluation Goal What is trying to be achieved through the evaluation? What is the coverage, criteria, and budget for evaluation?

Search Task A search task is the user's information needs. And it is typically represented as a description or as a query.

Search Engine A search engine is an IR system that include both the interface and algorithm under the hood, and the goal of evaluation is to evaluate some aspect of search engine.

Judging Target Judging target denotes a result produced by an IR system, and the item which is evaluated. It can be of any granularity – a snippet, a web document, or entire SERP.

Human Judgment A human judgment is an assessment of a *judging target* by a human judge, in the context of a *search task*, over some dimension of quality.

User Satisfaction User satisfaction is considered as the goal of evaluation. It can be elicited directly from user, or be inferred by human judges given the transcript of search session.

Evaluation Metric An evaluation metric (or metric in short) summarizes judgments into a single score. The design of an evaluation metric depends on the type of judgments being collected, and the model of user behavior.

Experiment We define an experiment as a collection of search tasks, judging targets, and human judgments with a specific evaluation goal. An evaluation metric summarizes the outcome of an experiment with a test collection, and an appropriate statistical test can be used to make a claim about the validity and reliability of the findings.

1.2.2 Evaluation Goal

The first step in offline evaluation is defining a clear evaluation goal. Evaluation goal itself is a multifaceted concept which can include success criteria, coverage and criteria. These are all critical questions that can determine many parameters of evaluation. Here we look into each in detail.

Evaluation Types First and foremost, evaluation goal should include the motivation and success criteria. While there can be many reasons for performing an evaluation – understanding the performance, finding defects, making a shipping decision, etc., one can broadly define two types of evaluation – exploratory vs. confirmatory, borrowing from the two major types of statistical analysis.

The goal of exploratory evaluation is to find the information about the performance of search engine in question. This can include assessing the performance of the search engine in overall, finding glaring defects to be fixed before shipping the product to the customers. In contrast, the goal of confirmatory evaluation is to derive numbers that can be a basis of decision making. This mostly takes the form of delta between two or multiple search engines.

These two evaluation types take different criteria of success, which results in different approaches. For exploratory evaluation whose focus is the discovery, the design of judging interface should emphasize the collection of detailed information about the judging target, whereas for confirmatory evaluation the focus should be mostly on collecting accurate labels for metric calculation.

Evaluation Coverage The effectiveness of a search engine is heavily dependent on various factors such as topic, user location and preference. For instance, a search engine which is excellent in sports and entertainment can be lousy in academic topics due to the type and characteristics of document collection. Therefore, it is crucial to design evaluation so that these various aspects are covered.

Evaluation Criteria The quality of a search engine can be defined in many dimensions. Topical relevance would be a main concern for any search engine, but there are other criteria which can be the focus of evaluation, such as novelty, freshness and authority of results. Also, some evaluation criteria such as diversity and topic coverage can be defined at the granularity of results set.

1.2.3 Evaluation Process

Given the evaluation goal, here we discuss the general process for offline evaluation. At a high level, offline evaluation based on human judgment is composed of three steps: 1) judgment design, 2) metric design and 3) experiment design. Alternatively, you can consider the whole process in terms of collecting data (judgments), combining them into meaningful numbers (metrics), carry out experiments to test hypotheses and draw conclusions (test collection). Now we discuss major considerations in each step.

Designing Human Judgments

In the first step, the details of human judgment should be defined, which is the basic unit of offline evaluation. Human judgments capture the quality of the results for given search tasks. Here are major considerations in this step:

1. How do you define and collect search tasks?
2. What should be your judging unit?
3. How do you design judging interface?
4. How do you hire and manage judges?

Designing Evaluation Metrics

The second step in offline evaluation is selecting or designing a evaluation metric. Metrics summarize the information from individual labels into meaningful numbers. This is essentially the question of how to combine labels to meaningful numbers.

1. How do you transform the labels from human judges?
2. How do you define user models in combining labels into a metric? ⋆
3. How do you estimate the parameters for the user model?

Maarten: You have not told us how or why user models enter the picture. Or what they are.

Designing Experiments

Lastly, judgments and metrics should be combined to achieve the goal of evaluation. Since this is an iterative step which takes several stages of refinement, here we describe methods and criteria in doing so.

1. How do you size the test collection to fulfill your evaluation goal?
2. How do you evaluate the validity of the outcome?

1.2.4 Summary

In this section, we introduced definitions and general process in offline evaluation which will constitute a general framework of offline evaluation. In what follows, we make the discussion more concrete by describing the trends in offline evaluation using this framework.

1.3 Trends in Offline Evaluation for IR

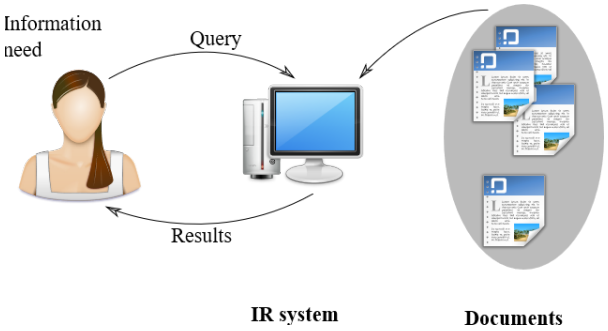
Information retrieval has a rich tradition of evaluation, both online and offline, and this tradition has been responsible for some of the rapid advances in search technology of the past two decades [Rowe et al., 2010]. Below we survey traditional approaches to offline evaluation, and consider trends in recent years which suggest new roles and methods.

1.3.1 Traditional Approaches to Offline Evaluation

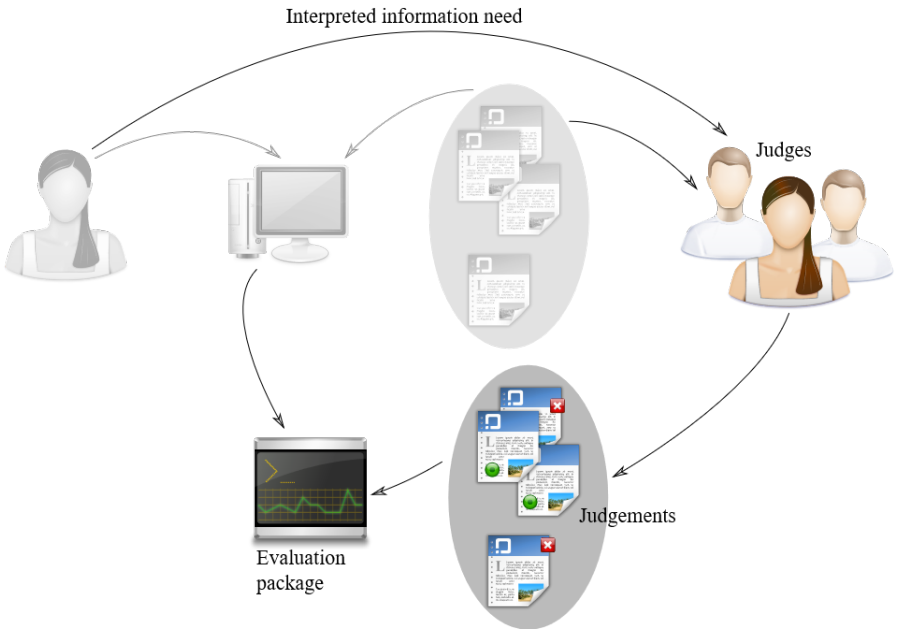
The traditional offline approach to IR evaluation is the test collection, or "Cranfield", approach first described by Cleverdon [1967] and refined through exercises such as the Text REtrieval Conference (TREC; see Voorhees and Harman [2005] for an overview). We will summarise this approach here, noting that Sanderson [2010] provides a historical summary and comprehensive discussion.

⋆⋆ In its most basic form, we can think of an information retrieval system as providing the right information to the right person in the right way. A user has an information need; they express this as a query; and the system will draw on the collection of information to produce some set of results (Figure 1.1a).

Maarten: That's a narrow view of IR: why not "provide the right information to the right person in the right way". So this could be about documents, but it could also be about answers, entities, etc.
Jin: done



(a) A simplified model of a retrieval system in context.



(b) Extension of the model, illustrating third-party relevance judges and the formation of a test collection.

Figure 1.1: Test-collection-based evaluation of an information retrieval system.

The test collection approach simulates this model by using the judgments of relevance and evaluation metrics that aim at measuring the quality of the results presented to the user (Figure 1.1b). The query, document collection, and retrieval system are as before but three components are added:

⋆

Judges interpret the user's information need, for example on the basis of the query or other context; and consider the extent to which each document in the collection answers this need.

Maarten: Should you not more explicitly and formally define what these core concepts are: need, query, inut for a need, test collection plus the relation between them?

Judgments record this information obtained from the judges for each (query, document) pair.

Evaluation is then a matter of aggregating the recorded judgements for the set (or ranking) of documents retrieved by the system; or comparing the documents retrieved with the documents judged as relevant.

For example, precision can be calculated by counting the number of retrieved documents which were marked as relevant; recall can be calculated by comparing the number of retrieved documents judged relevant with the overall number judged relevant; or rank-sensitive metrics, such as average precision or reciprocal rank, can look at the judgment for each retrieved document in turn.

Abstractions

⋆ In principle, the judgements formed are complete—that is, the collection includes a judgment for each possible (query, document) pair. Although this was true of some of the earliest exercises [Cleverdon et al., 1966], it is clearly impossible for today's much larger document sets. Two shortcuts have allowed researchers to collect useful judgments regardless. First, a simplifying assumption is that every search is over the same, fixed, document collection: that is, the documents do not change over time and nor are they different for different information needs (or users—as would be the case for personal or corporate collections).

Maarten: should you not put that in the big picture from the start? in a sense, we are sampling everywhere, queries, documents, judges. it would be good to point out the uncertainties that come with this from the start

Even allowing for this, it is clearly impossible to judge each of millions (or billions) of documents for any arbitrary information need. *Pooling* provides a common shortcut. In a pooled evaluation, each of a number of search systems provides its own ranking of documents, perhaps by running the same query on each. Every document which appears in the top k in any ranking is then judged, so if N systems contribute to the pool there are at most Nk judgments to be made: likely fewer, as some documents will appear in more than one list.

A related assumption is that each need can be captured in a single expression: that is, most (although not all) collections include a single query for each need. Although this is clearly a very small sample of possible inputs for the need, and although relatively small changes to a query can result in large changes in measured effectiveness [Bailey et al., 2015b], a large sample of needs can still capture useful variation.

Judging is also abstracted from real users in real contexts, in order to collect judgments at scale. The largest assumption here—indeed, an assumption relied upon by most effectiveness metrics—is that judgments are independent. That is, it is assumed that the extent to which a document is relevant to a need is independent of any other document which might be returned, or the order in which they are seen by the user. Notable exceptions are techniques from Golbus et al. [2014] and Chandar and Carterette [2013], who used relevance judgments based on other documents.

The notion of "relevance" is also normally abstracted. Although in reality relevance is complex, multi-facteted, and highly contextual [Borlund, 2003b, Saracevic, 2016], judges are often given much simpler instructions which can for example boil down to what Borlund [2003b] calls "intellectual topicality" alone. Recent work such as that by Mao et al. [2016] and Verma et al. [2016] also aims at extending this narrow definition of relevance. We delve into this in Chapter 2.

Test Collections

A central concept in traditional IR evaluation is the *test collection*. A test collection is the combination of

- a fixed set of documents;
- a set of information needs or topics, typically each with an associated query; and
- a set of relevance judgments which detail the relevance of at least some documents to each need.

Because they involve a static representation of an information-seeking session, test collections can be distributed;⋆ the judgments therein are reusable; and, in combination with one or more effectiveness metrics, they make it simple to compare systems. ⋆

Maarten: what does that mean

Test collections have been especially valuable for evaluation as they are easy to re-use: typically the limiting factor is just physical (or network) distribution of the documents themselves.⋆ Since they are so abstracted, they are self-contained, and it is trivial to compare results across systems, times, or laboratories.

Maarten: Explain how your survey relates to Mark SandersonâĂŹs survey and why we need your survey now.

Maarten: nope: pool bias

A noteworthy example of this is the Text REtrieval Conference (TREC) series, run annually by the (U.S.) National Institute of Standards and Technology (NIST). Since 1992 these conferences have been based around shared evaluations, using test collections so that each participating system can be directly compared to others Voorhees and Harman [2005]. The model has been adopted by a number of other conferences including the NTCIR Workshop[1], the Conference and Labs of the Evaluation Forum (CLEF)[2], and the Forum for Information Retrieval Evaluation (FIRE)[3]. These collections now include genera such as the web, microblogs, genomics, tourism, email, and others and in virtue of their scope and portability have become standard tools for information retrieval research.⋆

Maarten: Think of Hersh paper about relation between offline and user studies. Think also of relation between offline and online: often online is the key metric (REF?); e.g., work by Aleksandr Chuklin.

1.3.2 Recent Trends in Offline Evaluation

So far we have looked at traditional approaches in IR evaluation. While this tradition has served the community well for the past few decades,

[1] http://research.nii.ac.jp/ntcir/index-en.html
[2] http://www.clef-initiative.eu/
[3] http://fire.irsi.res.in/

there has been several trends which necessitate a change in the roles and methods of IR evaluation. In this section, we outline recent trends and discuss implications for offline evaluation.

User-Centric Evaluation

First and foremost, online search engines with large-scale user bases have become widely available and used, enabling online evaluation based on user behavior. This availability of user data has opened up the possibility of validating the assumptions of offline evaluation with actual user data. Recent work on evaluation metrics has embraced online user data to tune parameters of the metrics [Carterette et al., 2011, 2012, Smucker and Clarke, 2012, Yilmaz et al., 2010, for example].

The overall outcome of this trend is the advent of new IR evaluation paradigms which are more user-centric, diverse and agile. Here, being user-centric means that the evaluation process is based on a model of user behavior, or/and aims to improve user satisfaction or other user-visible measure such as engagement or task completion [Scholer et al., 2013b]. ⋆

Maarten: this is a very diffuse definition of "user-centric, diverse and agile" evaluation. Can you split out the three notions and be more precise about the definitions of each? Also, what is a user-visible measure? Please define.

This has already led to new methodologies to better estimate user satisfaction and behavior in judgment collection [Verma and Yilmaz, 2016, Verma et al., 2016] or metric design [Yilmaz et al., 2010, Carterette et al., 2011, Chapelle et al., 2009]. Also, some recent work has looked at cross-metric correlation, aiming to align IR evaluation with user satisfaction or some proxy of it [Al-Maskari et al., 2007, Radlinski and Craswell, 2010].⋆

Maarten: There is quite a bit more on this.

Diverse Endpoints and Search Scenarios

There are also new endpoints for search beyond desktop web browsers, such as mobile phones and conversational agents. This has opened up a whole area of research which focuses on different interaction methods and user experiences across endpoints. For instance, mobile devices have much smaller screen dimensions and the interaction is based on touch, while conversational agents use natural language, often in voice, to interact with the user.

Even for web search itself, the types of search results have diversified beyond the list of web documents to include other result types such as images, videos, news and even direct answers. This diverse set of results types, and corresponding user interface designs, breaks many assumptions of traditional IR evaluation, providing rich opportunities for exploration. In particular, many of these 'answers' can directly satisfy users' information needs on the SERP, making it hard to apply click-based evaluation techniques [Li et al., 2009, Diriye et al., 2012].⋆

⋆

Maarten: See Chuklin et al, CIKM 2016

IR evaluation research has with various lines of work. There has been increased interest in whole-page evaluation and optimization [Zhou et al., 2012], which encompasses a wide variety of page elements beyond web results. Task and session-level evaluation has also drawn interest [Kanoulas et al., 2011a, Carterette et al., 2014b], with TREC tracks of the same name [Carterette et al., 2014a]. Finally, there have been new lines of work focusing specifically on mobile interfaces [Verma et al., 2016], or evaluation of search with spoken agents [Kiseleva et al., 2016].

Maarten: I suggest organizing this differently: in every paragraph, first mention problem/challenge, then mention recent work that addresses this challenge, then scope: either point to later in the survey in case you are addressing the problem or explicitly say that you are not addressing it

Crowdsourcing / Agile Evaluation

⋆

Maarten: I don't see how the novelty or exploratory nature of new endpoints and scenarios calls for an agile manner of collecting labels. Vague. What do you mean "with less investments". Simply that it should be cheaper? Or that TREC style judging does not scale for financial reasons? Aren't new devices and crowdsourcing orthogonal?

These diverse new endpoints and scenarios for search required ways to collect labels in a more agile manner, because many of these services are new and exploratory by nature, with less investments compared to well-established ones like web search. Also, in academic settings, it has been difficult to recruit participants with diverse backgrounds at scale.

Fortunately, services such as Amazon Mechanical Turk have provided new ways for human judgments of any kind to be collected at a large scale. These services are called 'crowdsourcing', in that they pull the 'wisdom of the crowd' for tasks needing human intelligence. Accompanying this alternative data collection method is a challenge in quality control, since the labeling work is completed by a remote worker on the internet.

Given these opportunities and challenges, there has been quite a good deal of research on collecting high-quality labels with low ef-

fort [Alonso, 2012]. Popular approaches include using overlapping judgments to identify ground truth labels [Venanzi et al., 2014], or identifying the quality of judges based on their behaviors [Kazai and Zitouni, 2016]. We cover some of these methods in Section 2.3.1.

1.3.3 Summary

1.4 The Organization of this Survey

In the following chapters, we describe each process of offline evaluation in detail so that a reader can design his or her own evaluation pipeline following the flow of this paper. Chapter 2 deals with gathering judgments, which need to be created for the purpose. Chapter 3 considers steps in designing an effective metric. Chapter 4 covers the methods in designing and analyzing experiments. Finally, Chapter 5 describes evaluation practices from major companies in search and recommendation area. ⋆

Maarten: How is this consistent with earlier statements and definitions that all seem to focus on *document retrieval*,

2

Human Judgments

The goal of collecting human judgments is to estimate the satisfaction of actual users of a search system, by asking explicit questions to judges (or assessors) who simulate the actual users. A canonical example is collecting a binary relevance judgment for a document given a search topic. The form of human judgments can be quite varied, however, depending on the type of search task and judging target.

We will start with an example to make the discussion more concrete. Figure 2.1 shows a judging interface for evaluating the quality of a search engine results page (SERP) given the query 'crowdsourcing'. This example presents basic ingredients in collecting human judgments – search tasks and judging targets. From this example one can imagine a myriad of possibilities in designing a human judgment task.

While this is a simple example, it presents numerous trade-offs one can make in collecting human judgments. You can use either a (potentially ambiguous) keyword query, a well-defined topic description, or a description of a larger information-seeking task. You can collect judgments for a web document or any SERP element, including instant answers or a list of news articles. Queries, topics, or tasks could be created in many ways. And so on.

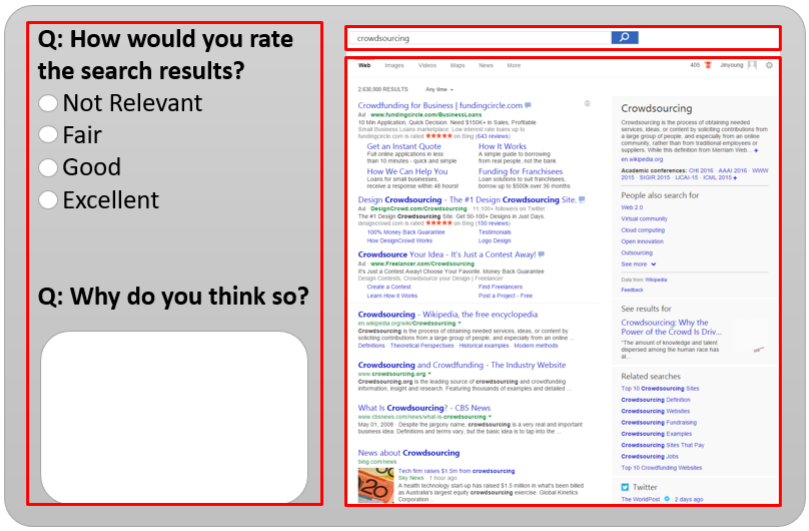


Figure 2.1: An example UI for human judgment collection.

The rest of this chapter is to give you guidance in collecting human judgments, in the light of recent literature on this topic. We will look over how to collect search tasks and how to determine a judging target. Various considerations in designing a judging interface will be examined, as well as methods for finding and managing human judges.

2.1 Collecting Search Tasks

Before considering judgment design, one needs to collect or construct search tasks against which search results will be evaluated. Search tasks are users' information needs that are typically represented as a description or as a query. In a setting where the search engine is used by actual users, the job of collecting search tasks can be as simple as sampling from queries users issue, whereas without access to such resources one needs to create tasks based on assumptions of target users and information needs.

2.1.1 Creating Search Tasks

In many cases one needs to perform offline evaluation without a working system – e.g. in building a new product, or in an academic setting. In such cases it is essential to collect hypothetical search tasks, often called simulated search or work tasks (where work includes search and other things). Borlund [2003a] summarizes the role of simulated work tasks as follows: ⋆

Maarten: I don't learn from 2.1.1 how I should go about creating a search task.

> "A simulated work task situation, which is a short 'cover story', serves two main functions: 1) it triggers and develops a simulated information need by allowing for user interpretations of the situation, leading to cognitively individual information need interpretations as in real life; and 2) it is the platform against which situational relevance is judged. Further, by being the same for all test persons experimental control is provided. Hence, the concept of a simulated work task situation ensures the experiment both realism and control."

'Task' can mean different things for different people, and the IR literature has seen long debate over the definition of search task (see Kelly [2009] for a summary). For our purpose, it is sufficient to understand it as a information need which can be represented in a way that a human judge can use to judge the quality of given result.

The design of search tasks takes a few considerations which can critically affect evaluation results. First, there is the question of where the task comes from and how much the judge is interested in or knowledgable about the task, or the corresponding domain. Edwards and Kelly [2016] show that judges' interests in the task has effects on how they perceive and perform the tasks. Judges in general had more knowledge on the tasks they were interested in, perceived the tasks as easier, and had higher engagement in terms of time spent. It is also known that judges' knowledge of the task can affect the quality of the outcome, with small but measureable differences between experts and non-experts [Bailey et al., 2008]. ⋆ ⋆

Maarten: What are the implications for setting up an assessment exercise yourself? How to take these findings into account?
Jin: I would recommend have a separate group of people designing evaluation, but this would be feasible mostly for industry settings.

Another dimension of task creation is complexity, which again has many aspects. Kelly et al. [2015] looked at this problem using a cognitive complexity framework. They found that participants spent more effort (queries, clicks and time to completion) in performing tasks with higher cognitive complexity (create, evaluate and analyze) than tasks with lower cognitive complexity (apply, understand, remember). In order to ensure the representativeness of the evaluation outcome, it would be sensible to balance the tasks of varying complexity in a way that matches actual users' workloads.

In summary, these results show that the characteristics of search task can affect the quality of corresponding human judgments, and therefore is an important dimension in designing an offline evaluation. It is recommended to collect information about task characteristics and design experiments accordingly so that one can control the effect of these factors in evaluation. ⋆

Maarten: Can you make this more precise? Give more details? What does "important" mean? This is all fairly abstract?

2.1.2 Sampling Query Logs

Assuming you have a working search engine with real users, it is natural to collect search tasks from query log data. While this is a seemingly straightforward task, there are a few considerations. We outline some below, along with recommendations based on recent studies. ⋆

Maarten: I would expect the outcome of 2.1.2 to be a clear recipe for sampling. But that 's not the case.

Evaluation Goals The appropriate sampling strategy depends on evaluation goals. In a typical scenario, it is reasonable to start with a *representative* sample of the traffic. Measurements based on this sampling strategy would lead to the characterization of *average* performance, but there are scenarios where average performance is not informative.

For example, Zaragoza et al. [2010] suggested techniques to identify segments useful for measurement. They introduce the notion of 'disruptive sets', which are a set of queries with high quality results in one engine, but not in another. Using a disruptive set, one can focus on the set of queries with a goal to gain competitive advantage.

Other goals can also dictate the choice of sample. For instance, in industry one often targets a specific query segment (e.g., queries with

fresh or local search intent); or perhaps on *hard* queries where there is more room for improvement and metrics are more sensitive. In these cases sampling from the particular segment maximizes the evaluation efficiency.

Characteristics of Search Traffic The characteristics of search traffic also needs to be considered. Baeza-Yates [2015] shows that web search query logs follow a power distribution, with longer tails. He suggests a sampling technique to generate a sample that follows this distribution. The main idea is to bin the queries based on the frequency, which allows the sampled queries to match the distribution of original query set.

Query vs. Task Description While it is possible to ask judges to imagine a search task given a query, it is open to question whether using a query to represent an information need is worthwhile.⋆ Unlike search tasks, which should contain sufficient details of user context and information need, queries in a typical search engine are often in an abbreviated form, ambiguous and/or with typographical errors. ⋆

Paul: I've tried to rephrase this

Jin: any recommended citation? i.e., % of queries with errors

These characteristics of user queries can be a significant source of noise because 1) there can be many query forms for the same information need [Bailey et al., 2015a], and 2) inferring true information needs from queries can be hard. On the other hand, Yilmaz et al. [2014a] argued that the choice of intent descriptions can also cause large variability in evaluation results and therefore the judging should be done based on queries.

All in all, despite limitations, user queries are still the most readily available sources of task information, and therefore are widely used for judging search results. One can mitigate the noise and ambiguity of the search query by training judges and presenting possible meanings of the query – i.e., a SERP from a commercial search engine. We discuss this in detail in Section 2.2.1. ⋆

Maarten: I did not find a discussion of noise and ambiguity in 2.2.1

2.1.3 Summary

In summary, here is our list of recommendations for collecting tasks for human judgments.

1. Decide whether to create search task or sample from query logs.
   (a) If query logs are available and queries are easy to understand, using query as the search task would be fine.
   (b) If query logs are available yet queries are not easy to understand, consider using queries to generate simulated search tasks.
   (c) If query logs are not available, consider reaching out potential users to collect search tasks.
2. In sampling queries from logs, use appropriate sampling strategy.
   (a) If the goal is to collect a representative sample of traffic, use random sampling.
   (b) If the goal is to collect a biased sample of traffic according to certain criteria, use stratified sampling.
3. Always collect metadata (task type, user con text, etc.) along with search tasks to facilitate further analysis. User context can also be presented as a part of the search task.

2.2 Designing a Judging Interface

Once the search tasks are collected, we are ready to design a system to gather judgments. There are several main considerations in designing a judging interface: we cover these in what follows.

1. How do we describe the context of a search task?
   (user location, preferences, previous queries in the session, etc.)
2. What should be the target of each judgment?
   (webpage, SERP elements or whole SERP)
3. What should be the scale of judgment?
   (absolute vs. relative, numeric scales vs. Likert-type scales vs. magnitude estimation)

4. What are the quality dimensions we want to measure? (relevance, usefulness, novelty, trustworthiness, etc.)

2.2.1 Judging Context

There are many contextual variables that affect user satisfaction with any given search result: users' knowledge and preference, language, timing and location of the search, just to name a few. Even with well-defined search tasks, it is hard to specify all these factors, let alone with terse keyword queries. Providing some of this contextual information to judges can potentially reduce the user-judge gap, thereby increasing the judgment quality. ⋆ ⋆

Paul: refs?
Maarten: Measured how?

The choice of what context to provide depends again on the evaluation goal – what do you want judges to know about the search task? For instance, if you think user location is crucial in judging the relevance of results (which is the case in many tasks), you should present the user's location alongside the query text. Note that, if possible, the location information should be collected along with user queries to get a realistic sample of actual user locations.

Relevance judgments are also affected by what user already did during the session, so it is reasonable to present some part of user session as judging context. Several authors have examined this. Chandar and Carterette [2013] used a document as context, with the goal of collecting judgments when the context document has already been read. They proposed an evaluation framework for novelty and diversity evaluation which captures subtopics implicitly and at finer-grained levels. Golbus et al. [2014] also experimented with using a document as a context, and found that the metrics based on conditional judgments correlate better with user preference at SERP-level.

While one may assume that adding more and more context can only increase the quality of judgments by reducing the user-judge gap further, it should be noted that more context means more effort for judges in digesting and applying the information. Moreover, more context can increase judging cost by adding a further source of variability. That is, instead of collecting judgment for every search task, these judgments should now be collected for every query and context pairs, which can

potentially make the evaluation prohibitively expensive.

Therefore, one should carefully consider the cost/value trade-off in adding the context to a judging task. As an extreme example, Mao et al. [2016] used the entire session as a judging context for collecting judgments on usefulness (as opposed to relevance) and found that usefulness metrics show higher inter-assessor agreement and better correlation with task-level satisfaction elicited from actual users. However, since adding the whole session as judging context increase both the effort needed for individual judgment and the number of judgments required, they recommend using usefulness evaluation only for post-hoc analysis of the experiments.

2.2.2 Judging Target

Judging target defines the basic form of judgment (i.e., what to present and how many), and it is the most critical decision as the details of judging interface depends on it. While creating a judging interface, we should also decide the granularity of judgment, and whether the judgment should be given for a single item, or a set of items.

Judging Unit

The judging unit is the unit at which judgments should be collected, i.e., at what granularity do we want to collect judgments? In web search, for example, the judging unit can be a webpage, SERP elements or a whole SERP, as shown in Figure 2.2.

The judging unit should be determined by the goal of evaluation: if you care about the quality of a ranked list, collecting judgments for each individual result seems like a natural choice. If the presentation of the whole SERP is a primary concern, the entire SERP might be the right unit.

On the other hand, if the judging target is reasonably complex with multiple sub-components, it is also possible to collect judgments at smaller units (i.e., SERP elements) and then calculate scores for large unit (i.e., the whole SERP) by combining unit scores in a sensible way. This is how most IR evaluation metrics (i.e., MAP or NDCG)

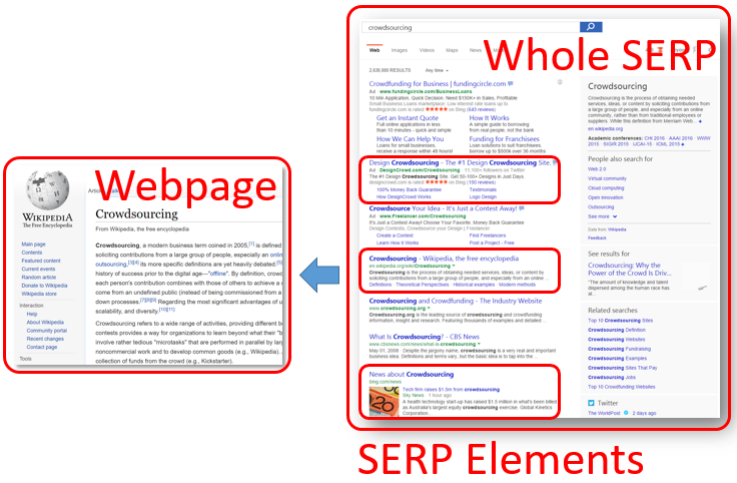


Figure 2.2: Various judging units for web search results.

work.

Now, if we want to collect judgments for SERPs, should we collect element-wise judgments and then combine, or collect single SERP-level judgments? This question can be generalized into the decision of judging unit when the judging target is complex. There is no hard and fast rule to determine the right judging unit, but here we describe a few trade-offs.

A smaller judging unit means a simpler judging task, which can be faster and more reliable ⋆. However, the number of judgments to evaluate a larger unit (i.e., a SERP) can be quite high if the judging unit is small, making overall judging cost higher than collecting a single judgment for the whole larger unit.

Paul: reference? or other evidence?

A smaller judging unit also means better reusability of individual labels, because you can combine labels for each SERP element to evaluate arbitrary configurations (e.g., arbitrary rankings of URLs on a SERP). This means that the cost of collecting judgments can be amortized over multiple experiments. In fact, query-URL relevance judgments have been so widely used in TREC and other settings because it allows the creation of test collection which can be used to evaluate

any ranked list.

On the other hand, using a smaller judging unit makes an assumption that each label can be collected independent of other elements – for example, that the quality of an item at rank 2 on a SERP can be assessed without knowing anything about ranks 1 or 3. This is hardly true in a typical search scenario where the concept and criteria of relevance can evolve over time. In this regards, larger judging units have the benefit of providing rich context for judges.

More importantly, larger judging units can capture various set-level properties – including the comprehensiveness, redundancy between elements. For instance, SERP-level judging can reveal whether the SERP captures all the reasonable intents for a given query. Also, the redundancy among documents in a ranked list can be captured only at the list-level.

In past work, as briefly mentioned above, document-level judgment is most prevalent. However, there has been some works that focus work SERP-level evaluation. Bailey et al. [2010] introduce a judgment scheme which can capture the interaction among SERP elements as well as element-level quality. SERP-level judgments were introduced by Thomas and Hawking [2006], who propose a pairwise judging interface in order to minimize the complexity of defining judging criteria (more about this in the following section). Several other works including Kim et al. [2013] refined this idea to include dimensional relevance judgments as well as overall SERP-level comparison. ⋆

Paul: I will add work by Falk et al. on judging snippets

Absolute vs. Relative Judgments

Another consideration in determining a judging target is the type of judgment, which can be either absolute or relative. In absolute judging, judgments are collected for a single judging target, whereas relative judgment asks for a pairwise preference between two targets. Figure 2.3 shows the two types of judgments in evaluating web search results.

Now, how should one choose between absolute and relative judgments? In general, absolute judging requires objective criteria to distinguish amongst different levels, whereas relative judgments can avoid the issue. Carterette et al. [2008] have also suggested that relative judg-

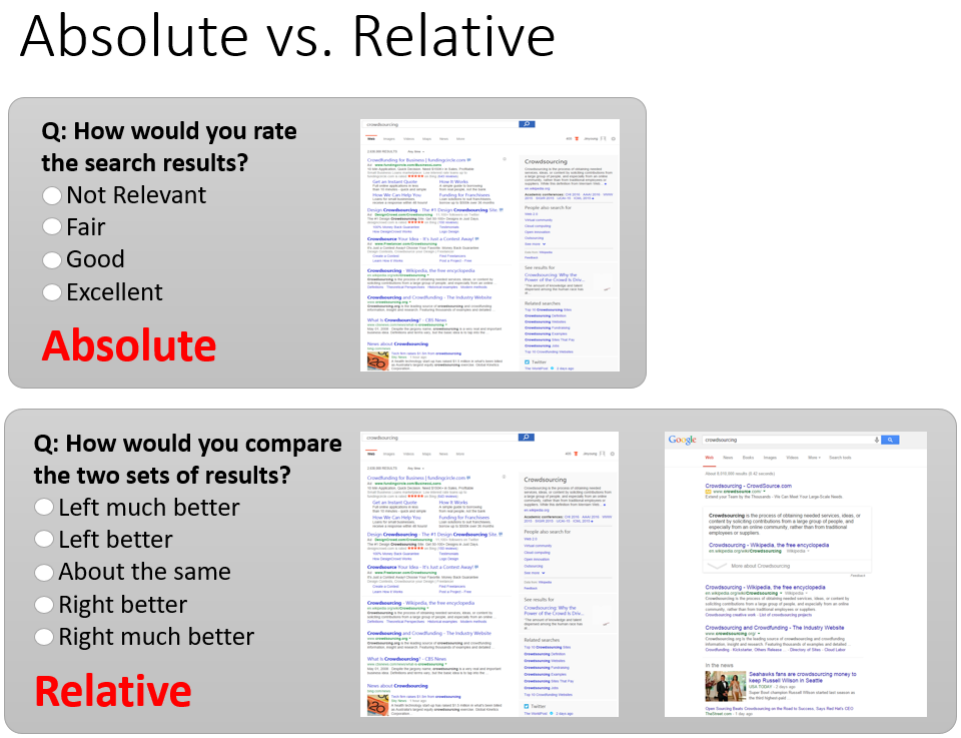


Figure 2.3: Absolute vs. relative judgments.

ments tend to be more accurate for document-level judging, while Kazai et al. [2013] found that a pairwise judging interface improves crowdsourcing quality so that it can be on par with that of trained judges.

Relative judgments have been used in various evaluation settings. Chandar and Carterette [2013] employed document-level pairwise judging using another document as a context, to evaluate novelty and diversity. Arguello et al. [2011] proposed an evaluation scheme for aggregated search based on pairwise preference judgment at element level, and Zhou et al. [2012] used SERP-level pairwise preference judgments as part of the evaluation framework for aggregated search.

On the other hand, the number of relative judgments grows with the square of the number of items. Since preferences may be weak, and may also be nontransitive, in principle each possible pair needs to be labeled. Carterette et al. [2008] On the other hand, absolute judgments are reusable in that you can compare among any items for which you have item-level labels. Therefore, if you want to reuse judgments in an environment where multiple generations of ranking techniques should be compared against each other, absolute judgments may save cost in the long run. This is also the reason that TREC has employed absolute judgment since its inception.

2.2.3 Judging Criteria

The central assumption of offline evaluation is that human judges can represent real users, and we often want judges to tell us if the judging target would be relevant to the potential user. ⋆ However, this is not a trivial task for judges given the contextual and multi-faceted nature of relevance [Borlund, 2003a], and for example Chouldechova and Mease [2013] report increased judging quality when done by query owners (users who did the search themselves) compared to query non-owners.

Maarten: Should relevance be the core criterion here? Why not "utility" (see eg Belkin).

Also, while the concept of relevance is broad, it typically specifies the relationship between an information need and an object, and is not sufficient to capture the true value of the item in the context of a search session. Therefore, it has been argued that IR as a field should move beyond relevance to evaluate usefulness in the context of search tasks [Belkin, 2015]. The TREC Session track [Carterette et al., 2014a]

and TREC Task Track Yilmaz et al. [2015] is another movement in the same spirit.

Recent work has tried to address this problem from multiple angles. The role of user effort and effort-based judging has been proposed [Yilmaz et al., 2014b, Verma et al., 2016], where it is shown that effort should be incorporated as an additional factor in human judgment to build retrieval systems that optimize user satisfaction. Carterette [2011] also analyzed existing evaluation metrics from the view of expected utility and expected efforts. Golbus et al. [2014] and Kim et al. [2013] also experimented with multi-dimensional judgment collection, which is useful in finding the relationship between different aspects of relevance.

Another thread of work looked at relevance judgments in the context of other items, or even the whole session. Chandar and Carterette [2013] proposed judging methods for novelty and diversity, where they employed preference-based judgment between document A and B in the context of a third document (C). The resulting method has the benefit of allowing the evaluation of novelty and diversity without requiring the collection of sub-topical judgments.

Mao et al. [2016] proposed collecting usefulness judgment in the context of whole session. They showed that high relevance by assessors is a necessary but not sufficient condition for high usefulness for users, and that usefulness judgments better correlate with behavioral signals such as click cumulative gains. But since usefulness judgments are costly to collect, they advised only collecting them post-hoc.

Overall, the current literature suggests many ways to set judging criteria for relevance, with different methods having different emphases. If the goal is to focus on query-document relevance, a simple interface as seen at the top of Figure 2.3 will do. However, one can add another document or even whole session history as a context if the goal is to capture the value of the item in the context of a broader search task. ⋆

Maarten: In terms of practical hands on advice on setting up label collection efforts, this is a bit vague.

2.2.4 Summary

In summary, here is our list of recommendations for designing an interface for human judgments collection.

1. Present each search task with context to reduce variability of the results.

2. Use the smallest judging unit at the beginning to collect fine-grained information with least amount of noise.

3. Consider collecting more set-level (coarse-grained) judgments as well to capture interactions among items.

4. Use absolute rating scale when it is possible to define clear criteria for each rating. Use relative rating scale otherwise, especially when employing crowd judges.

5. Judging interface design is an iterative process. Test multiple versions with small group of judges before scaling up.

2.3 Collecting Judgments

Once the judging interface is designed, the next task is to find judges to work with. There are quite a few options from which you can find judges, but you can roughly put them into four categories: 1) team members who work on the project, 2) expert judges who typically sit in-house with the team, 3) crowd judges who work remotely and can be reached via platforms like Amazon Mechanical Turk, 4) people who actually use the system.

How should we decide on which option to choose? First, it is recommended to start some judging exercise with the team (Group 1) before outsourcing the judging task, because you need to make sure you provide high-quality interfaces and descriptions to get judgments of reasonable quality. But this approach soon hits scalability issues⋆, so we focus on expert judges (Group 2) and crowd judges (Group 3) in this paper.

Maarten: Please explain. How many judges are needed? You don't say this?

There has been some recent work comparing human judges of different characteristics. Bailey et al. [2008] is a classic work where they found that judges' level of expertise on the domain can result in small yet consistent difference on system scores and rankings. Similarly, Chouldechova and Mease [2013] looked at judgments done by

query owners (users who did the search themselves) vs. query non-owners, where they concluded that query owners are can distinguish a higher quality set of search results from a lower quality set in a blind comparison.

However, neither finding domain experts nor using search tasks from judges themselves are feasible if you need judgments at large scale, or the goal is to collect judgments from representative sample of user traffic. Typically the options available are either in-house judges with some form of training or crowd judges.

Among these groups, Kazai et al. [2013] found that trained judges are significantly more likely to agree with each other and with users than crowd workers. But when they compared third-party judgments with clicks from real users, they found that the judgments from trained judges does not necessarily show higher agreement with metrics based on user clicks.⋆

Maarten: Implications for setting up your own labeling effort?

2.3.1 Crowdsourcing Relevance Judgments

Collecting labels from humans used to require finding and managing a group of people one by one, which is often an expensive and time-consuming process. Compared to this, crowdsourcing – hiring subjects from remotely using services such as Amazon Mechanical Turk – has a clear benefit in cost and scalability, and therefore it has gathered a lot of attention from research community, including a large body of work produced in IR community as well. Alonso [2012] provides a comprehensive survey of research and best practice in this area.

Along with the availability of cheap workforce from across the globe, the challenge in managing the quality of outcome has emerged. A standard approach in reducing errors has been aggregating redundant judgments from a group of independent assessors, and several works has focused on collecting and aggregating redundant labels.

Venanzi et al. [2014] proposed a community-based Bayesian label aggregation model which is based on finding latent groups among crowd workers and aggregating labels based on them. Davtyan et al. [2015] proposed using textual similarity to aggregate crowd judgments, where the relevance labels from similar documents are propagated. Companies

such as Crowdflower[1] provide services by which high quality labels are automatically calculated based on redundant judgments.⋆

Maarten: What's the point? How are readers of this survey going to benefit from this comment?

Another approach to improving the quality of crowdsourced judgments is by improving the judging interface design workflow by which crowd judges work on judging work. This section already dealt with design decisions on judging interface design, and Kazai et al. [2012] provide further guidance in deciding the complexity of judging tasks and the amount of payment per judgment.

Several authors have recently investigated workflow design for crowdsourcing. At microscopic level, Scholer et al. [2013a] and Shokouhi et al. [2015] looked at the effect of previous assessments on the quality of a judgment, and showed that the human annotators are likely to assign different relevance labels to a document depending on the quality of the last document they had judged for the same query. At a macroscopic level, Megorskaya et al. [2015] explored various parameters in designing workflow, and argue for having a communication channel between judges and 3–5-way overlap in a production environment.

2.3.2 Summary

In summary, here is our list of recommendations for collecting human judgments.

1. Always start the judging task internally and with small number of judges before scaling up to avoid wasting judging efforts.

2. For simpler judging tasks, consider crowdsourcing. Vary the amount of overlap to find the right trade-off between the judging cost and the precision of the outcome.

3. For more involved judging tasks, consider hiring in-house judges. Try to hire people with domain expertise to further improve the quality.

[1] https://www.crowdflower.com/

2.4 Open Issues

So far in this section, we looked at issues in collecting human judgments, and provided guidance based on latest research. However, search is rapidly evolving and as such new research areas are emerging. Before moving on to the next topic, here we discuss several open issues.

New Judging Targets Most existing research considers document-level judging. But modern SERPs contain rich results beyond documents, such as instant answers and multimedia results. Extending document-based judging model into these new judging targets would be an interesting problem. This includes judging methods for captions, instant answers and rich SERPs with all these elements.

New Endpoints for Search Smart phones are becoming standard devices for accessing the internet; and recently conversational agents have become a major focus for many tech companies. We are yet to learn how these new environments can affect judgment collection. Recent work such as that by Verma and Yilmaz [2016] and Kiseleva et al. [2016] provide some hints at what needs to change for these new environments.⋆

Maarten: This is too short/abstract to be meaningful.

New Judging Methods for Personalized Search Standard judging methods collect labels given a search task and a single, or a pair of, search results. However, this model may not work in environments where search is highly contextual and personal.⋆ Several recent works such as those by Xu and Mease [2009] and Moraveji et al. [2011] explored task-based judgment collection, where judges perform search given a (possibly personalised) search engine to make their judgments. ⋆ ⋆

Maarten: Explain.

Maarten: Make this a useful bit of information for your readers.

Maarten: Missing: a look ahead, i.e., a statement on how the choices made in this chapter (setting up the label collection) affects the next two stages in the offline evaluation pipeline, and vice versa.

3

Evaluation Metrics

The second step in offline evaluation is selecting or designing a meaningful evaluation metric. This is essentially the question of how to combine labels to meaningful numbers. For traditional IR evaluation where the labels are collected at query-URL level, combining labels to a metric requires quite a few assumptions, or even a user model. In this chapter, we go over the various considerations of IR metric design, as well as the user models behind these metrics. We briefly survey some established metrics but spend more time on recent developments: explicit models of user behavior, deriving metrics from these, and open issues including session-level measurement, dealing with variation, and considering rich SERPs. (20-25 pages)

3.1 Basic IR evaluation metrics

- Metrics based on absolute judgments (e.g. Cooper [1973])

- Metrics based on preference-based judgments, including e.g. aggregated in-situ side-by-side Thomas and Hawking [2006]

- Ranking-based metrics (Tau/TauAP)

- Criticisms: especially reproducability/replicability

3.2 Metrics based on simple aggregation of labels/qrels

- Set-based: P, R
  - Rank-based: P@k, AP, RR
  - Criticisms: what tasks and behaviors are modeled here?

3.3 Models of behavior

Evaluation metrics that are based on explicit models of user behavior
  - The cascade model and variants
  - Weights, the C/L/W framework [Moffat et al., 2013]
  - ERR, EBU, GAP, Time-biased gain, etc.
  - Weighted precision metrics such as RBP, INST; notion of residual [Moffat and Zobel, 2008, Moffat et al., 2015]
  - α-NDCG, IA metrics, etc.
  - Cost-based/economic models and the prospects of metrics from these
  - Session-level metrics Kanoulas et al. [2011b] Järvelin et al. [2008]

3.4 Model fitting

Fit of metrics to models; estimating the distribution of parameters/metric values based on user data

Carterette et al. [2011], Moffat et al. [2013]

3.5 Open issues

Open issues in behavior models and the corresponding metrics
  - Whole-page quality
  - Caption effects
  - Variation between users: behaviors, learning styles, cognitive styles, topic expertise, search system expertise, expectations of the system, query variation, ...
  - Duplication in SERPs
  - Learning (?)
  - Non-traditional tasks and novel UIs

- Choosing between metrics; sensitivity; finding evidence any of them correlates with user behavior or other important dependent variables

- Measuring things outside the SERP: query formulation, source/engine selection

4

Experiments

An experiment is defined as the collection of labels and metrics defined on top of them. We first look over many considerations in order to design an experiment given a budget and time constraint. We then focus on a set of analyses we can perform once the data is collected, followed by the ways of reporting experimental results. ($\approx$ 15 pages)

4.1 Designing an Experiment

- How to select queries?
    - How many queries? Sakai [2014]
    - How many documents? Carterette et al. [2009a]
    - How to distribute judgment efforts across queries and documents? Carterette et al. [2009b], Yilmaz and Robertson [2009]

4.2 Analysis of Experimental Results

Survey of research results Sakai [2016]

Drawing conclusions from metrics

- Hypothesis Testing Dinçer et al. [2014]

- Comparison of different types of significance tests Smucker et al. [2009]

Various analysis methods
- Power analysis Sakai [2014]
- Failure analysis
- Sensitivity and Reliability analysis Urbano et al. [2013]
- Informativeness (MaxEnt) Aslam et al. [2005]
- ETC Bron et al. [2013] Boytsov et al. [2013] Robertson and Kanoulas [2012]

Reporting results
- Effect sizes and distributions, vs point estimates and p values

4.3 Open Issues

- Reusability for SERP/task-level evaluation
- Beyond significance testing – bayesian alternatives?
- Reusability / Generalizability of experimental results

5

IR Evaluation in Practice

In this chapter, we review evaluation practices happening in both academia and industry. We first cover evaluation practices from academia, including recent TREC tracks, data generation efforts. We also look at evaluation efforts in related area such as recommendation systems and conversational agents. We then turn to evaluation practices from industry including major players in search and recommendation based on published papers and articles.

5.1 Evaluation Practices from Academia

Emerging TREC tracks
- Task track
- Microblog track
- Live QA track
- Contextual suggestions track

Dataset generation efforts
- Living labs for IR [1]

[1] http://living-labs.net/

- Data set shared by industry [2]

Evaluation in related domains
- Aggregate search Zhou et al. [2013]
- Recommendation systems Gunawardana and Shani [2015]
- Conversational agents

5.2 Evaluation Practices from Industry

How are the practitioners doing? (≈15 pages)
- Google [3] [4]
- Bing [5]
- Netflix Gomez-Uribe and Hunt [2015] [6]
- Facebook [7]
- Pinterest [8]
- LinkedIn [9]
- Startups [10]
[11]

Common features: combine online and offline evaluation
- Offline evaluation for early iteration
- Online evaluation for final ship decisions

[2] http://jeffhuang.com/search_query_logs.html
[3] How Search Works (Google) https://www.google.com/insidesearch/howsearchworks/thestory/
[4] Updating Our Search Quality Rating Guidelines https://webmasters.googleblog.com/2015/11/updating-our-search-quality-rating.html
[5] The Role of Content Quality in Bing Ranking (Bing) http://bit.ly/1T1BaYN
[6] The Netflix Tech Blog: Learning a Personalized Homepage http://techblog.netflix.com/2015/04/learning-personalized-homepage.html
[7] Who Controls Your Facebook Feed (Slate) http://slate.me/1T1BbvU
[8] Machine Learning at Pinterest http://www.slideshare.net/HiveData/the-hive-think-tank-machine-learning-at-pinterest-by-jure-leskovec-61383413
[9] http://www.slideshare.net/dtunkelang/search-quality-at-linkedin
[10] The Humans Hiding Behind the Chatbots http://www.bloomberg.com/news/articles/2016-04-18/the-humans-hiding-behind-the-chatbots
[11] 10 Data Acquisition Strategies for Startups http://bit.ly/28IHlC7

6

Conclusions

In this chapter we conclude this survey by providing the summary of contents so far. We also provide a brief outlook toward the future of offline evaluation for IR.

6.1 Summary

Recap: general Components of Offline Evaluation
- Experiment
- Search Task (Query / context)
- Evaluation Metric
- Judging Method (Interface / rating scale)

6.2 Future of Offline Evaluation for IR

Emerging trends in the tech ecosystem
- Mobile-first: different interfaces and information needs
- Natural-language interaction: Bots and Conversational agents
- End-to-end support for task completion: e.g., restaurant reservation

Future of Offline Evaluation

- Evaluation of search agents (as well as engines)
- Evaluation of various information 'cards'
- Evaluation of end-to-end task completion

Future of Offline Evaluation Research

- Need for Academy-Industry collaboration

References

Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 345–354, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. . URL http://doi.acm.org/10.1145/2009916.2009965.

Azzah Al-Maskari, Mark Sanderson, and Paul Clough. The relationship between ir effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR*, SIGIR '07, pages 773–774, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. . URL http://doi.acm.org/10.1145/1277741.1277902.

Omar Alonso. Implementing crowdsourcing-based relevance experimentation: an industrial perspective. *Information Retrieval*, 16(2):101–120, 2012. ISSN 1573-7659. . URL http://dx.doi.org/10.1007/s10791-012-9204-1.

Jaime Arguello, Fernando Diaz, Jamie Callan, and Ben Carterette. A methodology for evaluating aggregated search results. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 141–152, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-20160-8. URL http://dl.acm.org/citation.cfm?id=1996889.1996909.

Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu. The maximum entropy method for analyzing retrieval measures. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, pages 27–34, 2005. . URL http://doi.acm.org/10.1145/1076034.1076042.

Ricardo Baeza-Yates. Incremental sampling of query logs. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1093–1096, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2776780.

Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance assessment: Are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 667–674, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4. . URL http://doi.acm.org/10.1145/1390334.1390447.

Peter Bailey, Nick Craswell, Ryen W. White, Liwei Chen, Ashwin Satyanarayana, and S. M.M. Tahaghoghi. Evaluating search systems using result page context. In *Proceedings of the third symposium on Information interaction in context*, IIiX '10, pages 105–114, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0247-0. . URL http://doi.acm.org/10.1145/1840784.1840801.

Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. User variability and ir system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 625–634, New York, NY, USA, 2015a. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2767728.

Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. User variability and IR system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 625–634, New York, NY, USA, 2015b. ACM. .

Nicholas J. Belkin. Salton award lecture: People, interacting with information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1–2, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2767854.

Pia Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003a. URL http://informationr.net/ir/8-3/paper152.html.

Pia Borlund. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, May 2003b. ISSN 1532-2882.

Leonid Boytsov, Anna Belova, and Peter Westfall. Deciding on an adjustment for multiplicity in ir experiments. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 403–412, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL `http://doi.acm.org/10.1145/2484028.2484034`.

Marc Bron, Jasmijn van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. Aggregated search interface preferences in multi-session search tasks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 123–132, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL `http://doi.acm.org/10.1145/2484028.2484050`.

Ben Carterette. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 903–912, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. . URL `http://doi.acm.org/10.1145/2009916.2010037`.

Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there. In *ECIR*, pages 16–27, 2008.

Ben Carterette, Virgiliu Pavlu, Hui Fang, and Evangelos Kanoulas. Million query track 2009 overview. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*, 2009a. URL `http://trec.nist.gov/pubs/trec18/papers/MQ09OVERVIEW.pdf`.

Ben Carterette, Virgiliu Pavlu, Evangelos Kanoulas, Javed A. Aslam, and James Allan. If I had a million queries. In *Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings*, pages 288–300, 2009b. . URL `http://dx.doi.org/10.1007/978-3-642-00958-7_27`.

Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 611–620, 2011. . URL `http://doi.acm.org/10.1145/2063576.2063668`.

Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 135–144, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. . URL http://doi.acm.org/10.1145/2396761.2396782.

Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. Overview of the trec 2014 session track. Technical report, DTIC Document, 2014a.

Ben Carterette, Evangelos Kanoulas, Mark M. Hall, and Paul D. Clough. Overview of the TREC 2014 session track. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*, 2014b. URL http://trec.nist.gov/pubs/trec23/papers/overview-session.pdf.

Praveen Chandar and Ben Carterette. Preference based evaluation measures for novelty and diversity. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 413–422, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL http://doi.acm.org/10.1145/2484028.2484094.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 621–630, 2009. . URL http://doi.acm.org/10.1145/1645953.1646033.

Alexandra Chouldechova and David Mease. Differences in search engine evaluations between query owners and non-owners. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 103–112, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1869-3. . URL http://doi.acm.org/10.1145/2433396.2433411.

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. *Click Models for Web Search*. Morgan & Claypool, 2015. ISBN 9781627056489. .

C. W. Cleverdon. The cranfield tests on index language devices. *Aslib*, 19: 173–192, 1967.

Cyril W. Cleverdon, Jack Mills, and E Michael Keen. *Factors determining the performance of indexing systems; Volume 1: Design*. The College of Aeronautics, Cranfield, 1966.

William S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100, 1973. ISSN 1097-4571. . URL http://dx.doi.org/10.1002/asi.4630240204.

Martin Davtyan, Carsten Eickhoff, and Thomas Hofmann. Exploiting document content for efficient aggregation of crowdsourcing votes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 783–790, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3794-6. . URL `http://doi.acm.org/10.1145/2806416.2806460`.

B. Taner Dinçer, Craig Macdonald, and Iadh Ounis. Hypothesis testing for the risk-sensitive evaluation of retrieval systems. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 23–32, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. . URL `http://doi.acm.org/10.1145/2600428.2609625`.

Abdigani Diriye, Ryen White, Georg Buscher, and Susan Dumais. Leaving so soon?: understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1025–1034. ACM, 2012.

Ashlee Edwards and Diane Kelly. How does interest in a work task impact search behavior and engagement? In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 249–252, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3751-9. . URL `http://doi.acm.org/10.1145/2854946.2855000`.

Peter B. Golbus, Imed Zitouni, Jin Young Kim, Ahmed Hassan, and Fernando Diaz. Contextual and dimensional relevance judgments for reusable serp-level evaluation. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 131–142, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2744-2. . URL `http://doi.acm.org/10.1145/2566486.2568015`.

Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4):13:1–13:19, December 2015. ISSN 2158-656X. . URL `http://doi.acm.org/10.1145/2843948`.

Asela Gunawardana and Guy Shani. Evaluating recommender systems. In *Recommender Systems Handbook*, pages 265–308. Springer, 2015.

Ahmed Hassan. A semi-supervised approach to modeling web search satisfaction. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 275–284, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. . URL `http://doi.acm.org/10.1145/2348283.2348323`.

Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. Beyond dcg: User behavior as a predictor of a successful search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 221–230, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-889-6. . URL http://doi.acm.org/10.1145/1718487.1718515.

Katja Hofmann, Lihong Li, and Filip Radlinski. Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 2016.

Scott B. Huffman and Michael Hochster. How well does result relevance predict session satisfaction? In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 567–574, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. . URL http://doi.acm.org/10.1145/1277741.1277839.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. . URL http://doi.acm.org/10.1145/582415.582418.

Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, chapter Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions, pages 4–15. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-78646-7. . URL http://dx.doi.org/10.1007/978-3-540-78646-7_4.

Evangelos Kanoulas, Ben Carterette, Paul D. Clough, and Mark Sanderson. Evaluating multi-query sessions. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1053–1062, 2011a. . URL http://doi.acm.org/10.1145/2009916.2010056.

Evangelos Kanoulas, Ben Carterette, Paul D Clough, and Mark Sanderson. Evaluating multi-query sessions. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1053–1062. ACM, 2011b.

Gabriella Kazai and Imed Zitouni. Quality management in crowdsourcing using gold judges behavior. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 267–276, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3716-8. . URL http://doi.acm.org/10.1145/2835776.2835835.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178, 2012. ISSN 1573-7659. . URL `http://dx.doi.org/10.1007/s10791-012-9205-0`.

Gabriella Kazai, Emine Yilmaz, Nick Craswell, and S.M.M. Tahaghoghi. User intent and assessor disagreement in web search evaluation. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 699–708, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. . URL `http://doi.acm.org/10.1145/2505515.2505716`.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1âĂŤ2): 1–224, 2009.

Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. Development and evaluation of search tasks for iir experiments using a cognitive complexity framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, pages 101–110, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3833-2. . URL `http://doi.acm.org/10.1145/2808194.2809465`.

Jinyoung Kim, Gabriella Kazai, and Imed Zitouni. Relevance dimensions in preference-based ir evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 913–916, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL `http://doi.acm.org/10.1145/2484028.2484168`.

Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 45–54, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4069-4. . URL `http://doi.acm.org/10.1145/2911451.2911521`.

Jane Li, Scott Huffman, and Akihito Tokuda. Good abandonment in mobile and pc internet search. In *32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–50, 2 Penn Plaza, Suite 701, New York 10121-0701, 2009. URL `http://portal.acm.org/citation.cfm?id=1571941.1571951`.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Lihong Li, Jin Young Kim, and Imed Zitouni. Toward predicting the outcome of an a/b experiment for search relevance. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 37–46, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3317-7. . URL http://doi.acm.org/10.1145/2684822.2685311.

Chang Liu, Jingjing Liu, and Nicholas J. Belkin. Predicting search task difficulty at different search stages. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 569–578, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. . URL http://doi.acm.org/10.1145/2661829.2661939.

Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 463–472, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4069-4. . URL http://doi.acm.org/10.1145/2911451.2911507.

Olga Megorskaya, Vladimir Kukushkin, and Pavel Serdyukov. On the relation between assessor's agreement and accuracy in gamified relevance assessment. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 605–614, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2767727.

Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1), 2008. paper 2.

Alistair Moffat, Paul Thomas, and Falk Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 659–668, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. . URL http://doi.acm.org/10.1145/2505515.2507665.

Alistair Moffat, Peter Bailey, Falk Scholer, and Paul Thomas. INST: An adaptive metric for information retrieval evaluation. In *Proceedings of the Australasian Document Computing Symposium*, 2015.

Neema Moraveji, Daniel Russell, Jacob Bien, and David Mease. Measuring improvement in user search performance resulting from optimal search tips. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 355–364, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. . URL `http://doi.acm.org/10.1145/2009916.2009966`.

Filip Radlinski and Nick Craswell. Comparing the sensitivity of information retrieval metrics. In *SIGIR*, pages 667–674, 2010.

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 43–52, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. . URL `http://doi.acm.org/10.1145/1458082.1458092`.

Stephen E. Robertson and Evangelos Kanoulas. On per-topic variance in ir evaluation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 891–900, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. . URL `http://doi.acm.org/10.1145/2348283.2348402`.

Brent R Rowe, Dallas W Wood, Albert N Link, and Diglio A Simoni. Economic impact assessment of NIST's Text REtrieval Conference (TREC) program: Final report, July 2010. URL `trec.nist.gov/pubs/2010.economic.impact.pdf`.

Tetsuya Sakai. Designing test collections for comparing many systems. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 61–70, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. . URL `http://doi.acm.org/10.1145/2661829.2661893`.

Tetsuya Sakai. Statistical significance, power, and sample sizes: A systematic review of sigir and tois, 2006-2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 5–14, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4069-4. . URL `http://doi.acm.org/10.1145/2911451.2911492`.

Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. ISSN 1554-0669. . URL `http://dx.doi.org/10.1561/1500000009`.

Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd international ACM SIGIR*, SIGIR '10, pages 555–562, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0153-4. .

Tefko Saracevic. The notion of relevance in information science: Everybody knows what relevance is. but, what is it really? 8(3):i–109, September 2016. .

Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S. Lee, and William Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 623–632, New York, NY, USA, 2013a. ACM. ISBN 978-1-4503-2034-4. . URL http://doi.acm.org/10.1145/2484028.2484090.

Falk Scholer, Alistair Moffat, and Paul Thomas. Choices in batch information retrieval evaluation. In *Proceedings of the Australasian Document Computing Symposium*, 2013b.

Chirag Shah and Roberto González-Ibáñez. Evaluating the synergic effect of collaboration in information seeking. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 913–922, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. . URL http://doi.acm.org/10.1145/2009916.2010038.

Milad Shokouhi, Ryen White, and Emine Yilmaz. Anchoring and adjustment in relevance estimation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 963–966, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2767841.

Mark D. Smucker and Charles L. A. Clarke. Stochastic simulation of time-biased gain. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2040–2044, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. . URL http://doi.acm.org/10.1145/2396761.2398568.

Mark D. Smucker, James Allan, and Ben Carterette. Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 630–631, 2009. . URL http://doi.acm.org/10.1145/1571941.1572050.

Paul Thomas and David Hawking. Evaluation by comparing result sets in context. In *Proceedings of the 15th ACM CIKM*, CIKM '06, pages 94–101, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2. .

Julián Urbano, Mónica Marrero, and Diego Martín. On the measurement of test collection reliability. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 393–402, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. . URL http://doi.acm.org/10.1145/2484028.2484038.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 155–164, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2744-2. . URL http://doi.acm.org/10.1145/2566486.2567989.

Manisha Verma and Emine Yilmaz. Characterizing relevance on mobile and desktop. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, pages 212–223, 2016. . URL http://dx.doi.org/10.1007/978-3-319-30671-1_16.

Manisha Verma, Emine Yilmaz, and Nick Craswell. On obtaining effort based judgments for information retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, pages 277–286, 2016. . URL http://doi.acm.org/10.1145/2835776.2835840.

Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press, 2005.

Ya Xu and David Mease. Evaluating web search using task completion time. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 676–677, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. . URL http://doi.acm.org/10.1145/1571941.1572073.

Emine Yilmaz and Stephen Robertson. Deep versus shallow judgments in learning to rank. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 662–663, 2009. . URL http://doi.acm.org/10.1145/1571941.1572066.

Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1561–1564, 2010. . URL http://doi.acm.org/10.1145/1871437.1871672.

Emine Yilmaz, Evangelos Kanoulas, and Nick Craswell. Effect of intent descriptions on retrieval evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 599–608, New York, NY, USA, 2014a. ACM. ISBN 978-1-4503-2598-1. . URL http://doi.acm.org/10.1145/2661829.2661950.

Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. Relevance and effort: An analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 91–100, New York, NY, USA, 2014b. ACM. ISBN 978-1-4503-2598-1. . URL http://doi.acm.org/10.1145/2661829.2661953.

Emine Yilmaz, Evangelos Kanoulas, Manisha Verma, Ben Carterette, Nick Craswell, and Rishabh Mehrotra. Overview of the trec 2015 tasks track. Technical report, 2015.

Hugo Zaragoza, B. Barla Cambazoglu, and Ricardo Baeza-Yates. Web search solved?: All result rankings the same? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 529–538, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. . URL http://doi.acm.org/10.1145/1871437.1871507.

Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M. Jose. Evaluating aggregated search pages. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 115–124, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. . URL http://doi.acm.org/10.1145/2348283.2348302.

Ke Zhou, Mounia Lalmas, Tetsuya Sakai, Ronan Cummins, and Joemon M. Jose. On the reliability and intuitiveness of aggregated search metrics. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 689–698, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. . URL http://doi.acm.org/10.1145/2505515.2505691.