Foundations and Trends® in Information Retrieval
Vol. XX, No. XX (2016) 1–43

DOI: 10.1561/XXXXXXXXXX

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.

Paul: This suggests that lab studies are in scope.

Jin: I think it's hard to draw boundaries, except for its goals.

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. The overall outcome of this trend is the advent of new IR evaluation paradigms which are more user-centric, diverse and agile.

Paul: Doesn't this suggest offline evaluation doesn't matter? Tone this down?

Jin: We'll talk about its limitations later

This survey aims to provide an overview of recent research in IR evaluation pertaining to the trends above. 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.

now Publishers Inc.. *Offline Evaluation for Information Retrieval*. Foundations and Trends® in Information Retrieval, vol. XX, no. XX, pp. 1–43, 2016.

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 implicit measurement of real users' experiences of the system in a natural usage environment [Hofmann et al., 2016]. That is, online evaluation directly employs user behavior in natural environment for evaluation.

As large-scale online services become commonplace, online evaluation became a viable option for companies who is running service with large user base. In 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.

1.1.2 Hybrid Approaches

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

Paul: really? How about "...is based on abstractions of real users"?

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 user behavior. ⋆

Paul: how is that a limitation? need to be explicit

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.

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. However, the distinctions are getting blurred as search engines increasingly serve more complex set of results, and SERP (search engine results page) or session-level evaluation is drawing more attention. In fact, some recent research has tried to use task completion settings for system-to-system comparison [Xu and Mease, 2009]. We will return to this point in Chapter 2.

1.1.3 When to Use Offline Evaluation

At this point, a reader may ask: when should we use online vs. offline evaluation? While online metrics are certainly valuable and must-have when feasible, there are reasons we may need explicit input from human judges. First, in initial stages of system development we simply might not have real users to study. More importantly, traces of 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 click is certainly an indication that user is interested in the result, it is not clear whether the clicked result actually led to satisfaction. Also, click is often concentrated on the top of the page, making it difficult to interpret. That is, the ambiguity and bias inherent in user behavior often make it hard to infer true quality of our products.

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. By contrast, the data collected from online system is valid for the evaluation of the system user interacted with, and the data should be collected for every new system to be developed.

⋆

Emine: Actually there are more reasons than that. I will send a table that contain advantages/disadvantages and we might want to include that.

1.2 Offline Evaluation for IR

1.2.1 Traditional Approaches in Offline Evaluation

Paul: to write

⋆

The field of IR has rich tradition in evaluation.

Conceptual Model

- Labels/Metrics based on Query-URLs
- Test collections
- Concept of relevance

History

- TREC and related evaluation venues Sanderson [2010]
- Refer to Borlund [2003b] Cleverdon [1967] Voorhees and Harman [2005]

1.2.2 Recent Trends in Offline Evaluation

Paul: perhaps: developments since TREC got established

⋆

So far we have looked at traditional approaches in IR evaluation. While this tradition has served the community well for the past few decades, there has been several trends which necessitates the change in the role and method of IR evaluation. In this section, we outline recent trends and delve into their implications for offline evaluation.⋆

Emine: “recent”

User-Centric Evaluation

First and foremost, online search engines with large-scale user base has become commonplace, enabling online evaluation based on user behavior. This availability of user data has opened up possibilities to validate assumptions of offline evaluation with actual user data. Also, recent work on evaluation metrics have embraced online user data to tune parameters of the metrics.

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]).

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

As a side note, there has been an increasing efforts to combine online and offline evaluation. These include ways to use online user data for offline evaluation Li et al. [2015] Li et al. [2010] Chuklin et al. [2015], or ways to collect feedback directly from user Kim et al. [2016].

(mentions of user study / iir papers)

Diverse Endpoints and Search Scenarios

There are also new endpoints for search beyond desktop web browser such as mobile phone and conversational agents. This opened up a whole venue of research which focuses on different interaction method and user experience in respective endpoints. For instance, mobile device has much smaller screen dimensions and the interaction is based on touch, and conversational agents use natural language, often in voice, to interact with the user.

Even for web search itself, the types of search results has diversified beyond the list of web documents to include other results types such as images, videos, news and even direct answers. This diverse set of results types and user interface design 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 SERP, making it hard to apply click-based evaluation techniques Li et al. [2009] Diriye et al. [2012].

IR evaluation research has responded to this needs with various lines of work. There has been increased⋆ interests on whole-page evlauation Emine: and optimization Zhou et al. [2012], which encompasses wide variety of page elements beyond web results.

Task and Session-level evaluation Kanoulas et al. [2011a], Carterette et al. [2014] also drew interests, with TREC tracks of the same name . Finally, there has been a new line of work focusing specifically on mobile interfaces Verma et al. [2016], or evaluation of search with spoken agents Kiseleva et al. [2016].

Crowdsourcing / Agile Evaluation

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.

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

Given this opportunities and challenges, there has been quite a few research work Alonso [2012] about how to collect high-quality labels with least efforts. 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.

1.3 Scenarios for Offline Evaluation

We have outlined basic concepts and recent trends for offline evaluation so far. The goal of this paper is to provide a practical guide in conducting offline evaluation for both academic and industry practitioners. Since there can be various scenarios in conducting offline evaluation, here we outline possible ones which we cover in this paper.

In classical IR research, a typical evaluation scenario is to improve the performance of a system given a test collection and a predetermined set of evaluation metrics. For instance, in TREC Web Track, participants are given a collection representative of the Web, and then asked to submit the results for their systems in designated format and due date, which then will be evaluated on metrics like NDCG Järvelin and Kekäläinen [2002] or ERR Chapelle et al. [2009].

While academic IR research has developed well-accepted evaluation practice as above, the situation is a lot more ill-defined and varied from practitioners' standpoint. There are multiple components in a modern IR system such as web search engine, and each requires different emphases and considerations. For instance, one can think of component-level (i.e., query suggestions) evaluation as opposed to system-level evaluation.

Also, building a working system serving 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 and so on. 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 setting is that the evaluation is an on-going process which takes multiple iterations over the lifetime of the service, as opposed to one-off research project. This necessitates the development of so called *evaluation pipeline* where any new system can be evaluation 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 setting at Chapter 4.

1.4 General Framework for Offline Evaluation

In this section, we describe a general framework for offline evaluation in detail. The goal is to propose a general framework which can encompass diverse set of scenarios outlined above.

1.4.1 Definitions

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

Search Task A search begins with user's information needs, which we call a search task. Search task can be represented as a description of information needs, or queries user would have used in actual information seeking.

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

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

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

Test Collection A test collection is a collection of judgments with a specific evaluation goal. An evaluation metric summarizes the outcome

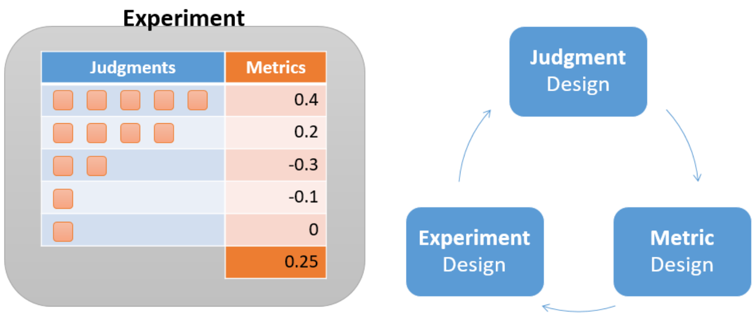

Figure 1.1: Overview of Offline Evaluation.

of an experiment, and an appropriate statistical test needs to be accompanied to make a claim about the validity and reliability of the findings.

Emine: I am not sure what an experiment refers to here

Paul: Can we refer to this as an *evaluation*? "Experiment" has a particular meaning; and in particular, (a) we need not be varying anything when doing offline evaluation (so it's not experimental) and (b) when we do, this measurement/evaluation can be a part of the larger experiment. Calling this an "experiment" is I think wrong, it's a measurement tool/technique

Jin: I propose calling it Test Collection. What do you think?

⋆ ⋆ ⋆

1.4.2 Evaluation Process

Given the components above, here we discuss the general process for offline evaluation. At a high level, offline evaluation is composed of three steps 1) judgment design 2) metric design 3) experiment design. Alternatively, you can consider the whole process in terms of collecting data (judgments), combining them into meaningful numbers (metrics), drawing conclusions (experiment). 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 train 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?

Designing Test Collections

Lastly, judgments and metrics should be combined into a test collection 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?

1.5 The Organization of this Paper

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. ⋆ ⋆

Emine: We already had a part describing the organization. In general, this section feels a bit repetitive given the text in first page

Paul: I disagree, that was in the abstract; it makes sense here (as well) I think. Unless I missed something?

2

Human Judgments

The goal of collecting a human judgment is to get an accurate measurement of search engine results quality for given set of search tasks. A canonical example is collecting a binary relevance judgment for a document given a TREC-style search topic. The form of human judgment 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 list of possible search tasks about the topic of *crowdsourcing* on the left side, and a few samples from existing web search results for query 'crowdsourcing' on the right side.

Emine: Would it be better to show a more standard judging UI here? Like a query and a web page?

Paul: Can we use another topic? We also discuss crowdsourcing below, which may be confusing. Let's use something which is clearly from a user not a researcher, e.g. "rules of soccer". I like the diagram otherwise I think

Jin: @emine we do show judging I/F later. @paul let's keep it this way unless you strongly disagree – i've used this topic throughout the chapter so it's not trivial to change

⋆ ⋆ ⋆

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. You can use either a (potentially ambiguous) keyword query or a well-defined topic description. You can collect judgment for a web document or any SERP element including instant answers or the list of news articles.

The rest of this chapter is to give you guidance in designing a hu-

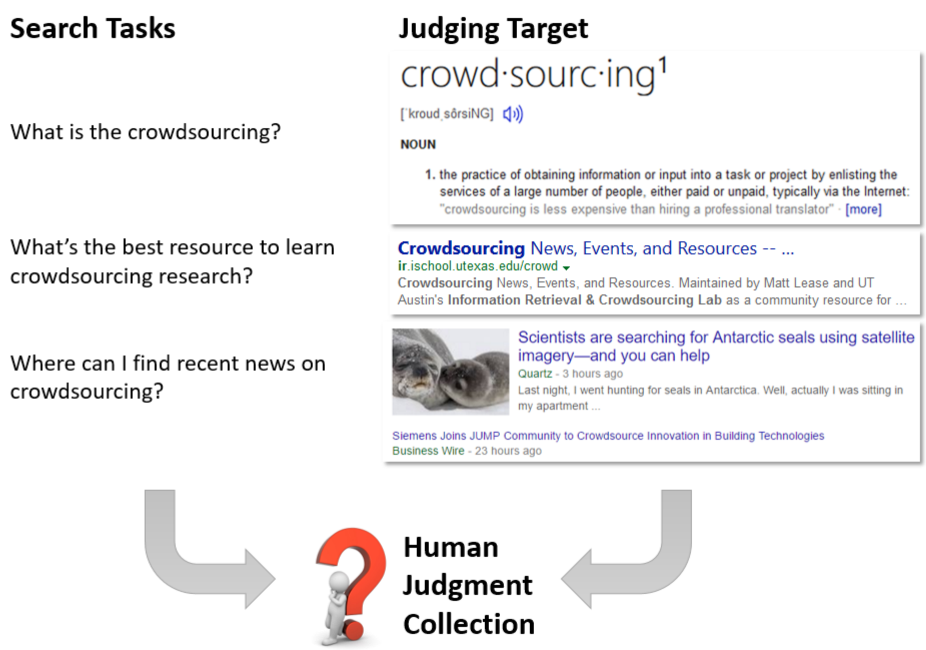

Figure 2.1: Overview of human judgment collection.

man judgment, 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.

2.1 Collecting Search Tasks

Before considering judgment design, one needs to collect search tasks on which search results will be evaluated. Search tasks represents users' information needs that needs to be satisfied by the search results. In an industry 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 issued, 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 – in building a new product, or in academic setting. It is essential to collect hypothetical search tasks in such cases, which is called simulated search or work (where work includes search and other things) tasks. Borlund [2003a] summarizes the role of simulated work task as follows:

> 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 IR literature has long debated over the definition of the search task, as summarized in Kelly [2009]. For our purpose, it is sufficient to understand it as the representation of information needs which a human judge can use to perform a search and⋆ judge the quality of results. ⋆ ⋆

Paul: and/or? Often no searching is done by a judge

Emine: Is that the definition of task? May be we should use a more proper definition?

Paul: task $\neq$ need, but I think this use is blessed by so much past use

The design of search tasks needs many considerations which can critically affect evaluation results. First, there is the question of where the task is originated from and how much the judge is interested in the task. Edwards and Kelly [2016] shows 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, expected the tasks to be easier, and had higher engagement in terms of time spent.

Another dimension of task creation is the complexity, which again has many dimensions. Kelly et al. [2015] looked at this problem using a cognitive complexity framework. They found that participants spent more efforts (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 sum, these results show that the characteristics of search task 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.

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.

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* per-

Paul: A random sample, you mean? Deduplicated? Balanced/stratified?

formance, but there are scenarios where average performance is not informative.

For example, a recent paper from 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. In these case a sample focusing on 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.

Paul: so, stratified and re-balanced?

Jin: more details?

⋆ ⋆

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

Paul: empty cite?

These characteristics of user queries can be a significant source of noise because 1) there can be many query forms for given information needs, as shown by Bailey et al. [2015], 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 description can also cause large variability in judgment 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.

2.2 Designing a Judging Interface

Once the search tasks are collected, we are ready to design judging interface. There are several main considerations in designing a judging interface as listed below. We cover these in what follows.

1. How do we describe the context of search task?
(user location, session history, etc.)

2. What should be the target of judgment?
(webpage, SERP elements or whole SERP)

3. What is the quality dimension we want to measure?
(relevance, usefulness, novelty, etc.)

4. What should be the scale of judgment?
(absolute vs. relative, likert vs. numeric)

2.2.1 Judging Context

There are many contextual variables that affects user satisfaction on a given search results. Users' knowledge and preference, the timing and the location of the search, just to name a few. Even with well-defined search task, it is hard to specify all these factors, let alone with terse keyword queries. Providing some of these information to human judge can potentially reduce user-judge gap, thereby increasing the quality of judgment.

The choice of judging context depends on evaluation goal – what do you want judges to be aware about the search task given? 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 location of the task, too. Note that, if possible, the location information should be collected along with user queries to get a realistic sample of actual user locations.

Relevance judgment can certainly get affected by what user already did during the session, so it is reasonable to present some part of user session as a judging context. In fact, recent work has proposed various types of judging context from users' session context. Chandar and Carterette [2013] used a document as a context with a goal to collect judgments when the context document has already been read. They proposed an evaluation framework for 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, it should be noted that more context means more efforts for judges in digesting the information and applying them for judgments. Moreover, more context can increase judging cost by adding further source of variability. That is, instead of collecting judgment for every search task, now that judgments should be collected for every query and context pairs, which can potentially make the evaluation prohibitively expensive.

Therefore, one should carefully consider the value-cost trade-off in adding the context to a judging task. 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 user satisfaction. However, they recommend using usefulness evaluation only for post-hoc analysis of the experiments due to high cost associated with using the whole session as a context.

2.2.2 Judging Target

Judging target defines the basic form of judgment. In what granularity the judgment should be collected (judging unit), and whether the judgment should be given for single item, or a set of items (judgment type).

Figure 2.2: Various judging units for web search results.

Judging Unit

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

Basically, judging unit should be determined by the goal of evaluation: if you care about the quality of ranked list, collecting judgment for each web search result seems like a natural choice. If the presentation of SERP is primary concern, SERP should be the right unit for judgment.

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

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

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

Smaller judging unit also means better reusability of individual labels, because you can combine labels for each SERP element to evaluate arbitrary SERP configuration. This means that the cost of collecting judgments can be amortized over multiple experiments. In fact, query-URL relevance judgment has 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, smaller judging unit makes an assumption that each label can be collected independent of other element. This is hardly true in a typical search scenario where the concept and criteria of relevance can evolve over time. On this regards, larger judging unit has the benefit of providing rich context for judges. Also, larger judging unit can capture the interaction between elements – i.e., redundancy among documents in a ranked list.

In literature, as briefly mentioned above, document-level judgment is most prevalent. However, there has been a few papers which deal with SERP-level evaluation. Bailey et al. [2010] introduces a judgment scheme which can capture the interaction among SERP elements as well as element-level quality.

SERP-level judgments were introduced in Thomas and Hawking [2006], where they used pairwise judging 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.

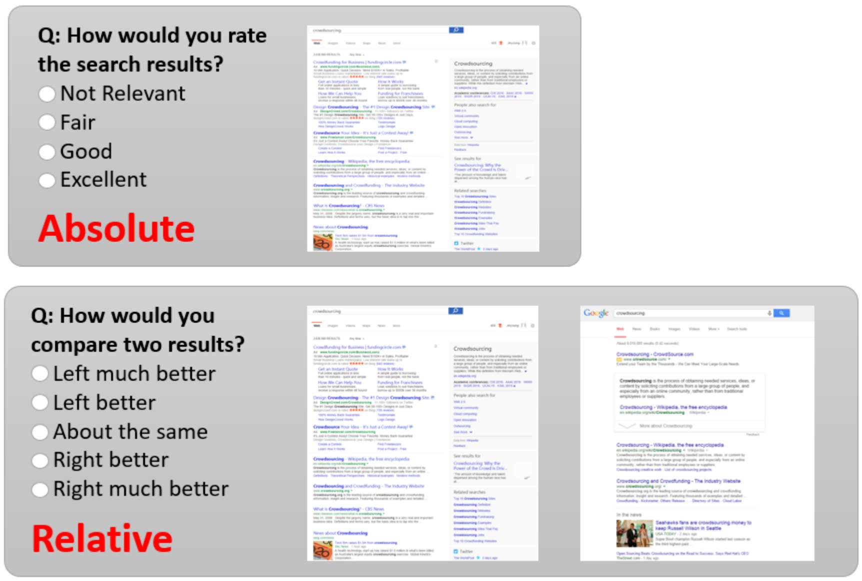

Figure 2.3: Absolute vs. Relative judgments

Absolute vs. Relative Judgment

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

Now, how should one choose between absolute vs. relative judgment? In general, making an absolute scale judgment requires having objective criteria among different levels, whereas relative judgment can avoid the issue. Carterette et al. [2008] also suggested that relative judgment is more accurate for document-level judging.

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

On the other hand, absolute judgments are reusable in that you can compare among any items for which you have item-level labels, whereas you need to collect labels for every pair of items. Therefore, if you want to reuse judgments in a production environment where multiple generations of ranking techniques should be compared against each other, absolute judgment might save the 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 useful to the potential user. However, it is not a trivial task for judge given contextual and multi-faceted nature of relevance. (Borlund [2003a])

Effort based judgments Yilmaz et al. [2014b] Verma et al. [2016]

Relevance vs. Usefulness-based evaluation Zhou et al. [2012] Kim et al. [2016]

Multi-dimensional judgment collection Golbus et al. [2014] Kim et al. [2013]

Novelty and diversity Chandar and Carterette [2013]

2.2.4 Additional Issues

Judging scale Turpin et al. [2015] Järvelin and Kekäläinen [2002]

Decision Criteria
- Judgment goal (target / decision)
- Judging effort/time

- Outcome reliability/intepretability
- Reusability

2.3 Collecting Judgments

Choosing Judges:
- Crowd vs. Expert Kazai et al. [2013] Alonso and Mizzaro [2012]
- Query owner vs. non-owners Chouldechova and Mease [2013]

Reducing noise in judging:
- Anchoring bias in judging Shokouhi et al. [2015]
- Multiple judgments and majority voting, etc. Venanzi et al. [2014] Aroyo and Welty [2013b] Aroyo and Welty [2013a]

Efficient judgment collection using Crowdsourcing
- Design decisions that need to be tackled Blanco et al. [2011] Kazai et al. [2012] Alonso [2012] Alonso et al. [2015] Scholer et al. [2013a]
- Incentivising judges and how to make it more attractive (payment / I/F) Megorskaya et al. [2015] Davtyan et al. [2015] Rokicki et al. [2014] Eickhoff et al. [2012]

2.4 Open Issues

Collecting labels for contextual / personalized search results
- using judgments with search context?

Collecting labels for whole SERP / non-document results

Collecting labels for non-traditional endpoints (i.e., conversational agent)
- Judgment for Desktop vs. Mobile environment **?**

Session/Task-based evaluation Moraveji et al. [2011] Xu and Mease [2009]

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

Experiments 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

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.

Omar Alonso and Stefano Mizzaro. Using crowdsourcing for {TREC} relevance assessment. *Information Processing {and} Management*, 48(6):1053 – 1066, 2012. ISSN 0306-4573. . URL http://www.sciencedirect.com/science/article/pii/S0306457312000052.

Omar Alonso, Catherine C. Marshall, and Marc Najork. Debugging a crowdsourced task with low inter-rater agreement. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '15, pages 101–110, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3594-2. . URL http://doi.acm.org/10.1145/2756406.2757741.

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.

Lora Aroyo and Chris Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013, 2013a.

Lora Aroyo and Chris Welty. Measuring crowd truth for medical relation extraction. In *2013 AAAI Fall Symposium Series*, 2013b.

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, 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, 2015. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2767728.

Roi Blanco, Harry Halpin, Daniel M. Herzig, Peter Mika, Jeffrey Pound, Henry S. Thompson, and Thanh Tran Duc. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 923–932, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. . URL http://doi.acm.org/10.1145/2009916.2010039.

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, 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, 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*, 2014. 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.

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.

Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 871–880, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5. . URL http://doi.acm.org/10.1145/2348283.2348400.

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.

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

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.

Jin Young Kim, Jaime Teevan, and Nick Craswell. Explicit in situ user feedback for web search results. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16. ACM, 2016.

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.

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.

Markus Rokicki, Sergiu Chelaru, Sergej Zerr, and Stefan Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1469–1478, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2598-1. . URL http://doi.acm.org/10.1145/2661829.2661946.

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 TrendsÂő in Information Retrieval*, 4(4):247–375, 2010. ISSN 1554-0669. . URL http://dx.doi.org/10.1561/1500000009.

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, 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. .

Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 565–574, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. . URL http://doi.acm.org/10.1145/2766462.2767760.

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`.

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`.