11-2009

A Coherent Measurement of Web-Search Relevance

William Mahoney
University of Nebraska at Omaha, wmahoney@unomaha.edu

Peter Hospodka
University of Nebraska at Omaha

William Sousan
University of Nebraska at Omaha

Ryan Nickell
University of Nebraska at Omaha

Qiuming Zhu
University of Nebraska at Omaha, qzhu@unomaha.edu

A Coherent Measurement of Web-Search Relevance

William R. Mahoney, Peter Hospodka, William Sousan, Ryan Nickell, and Qiuming Zhu
University of Nebraska at Omaha

Abstract – We present a metric for quantitatively assessing the quality of Web searches. The relevance-of-searching-on-target index measures how relevant a search result is with respect to the searcher's interest and intention. The measurement is established on the basis of the cognitive characteristics of common user's online Web-browsing behavior and processes. We evaluated the accuracy of the index function with respect to a set of surveys conducted on several groups of our college students. While the index is primarily intended to be used to compare the Web-search results and tell which is more relevant, it can be extended to other applications. For example, it can be used to evaluate the techniques that people apply to improve the Web-search quality (including the quality of search engines), as well as other factors such as the expressiveness of search queries and the effectiveness of result-filtering processes.

Index Terms—Human factors, network interfaces, search methods.

I. Introduction

When studying methods and techniques for improving Web-based search performance, people often question how to quantitatively assess the amount of improvement that a new technique brings. The same question is also asked when a new set of keywords is used or a revision/refinement of the search terms is made. It is useful to have a methodology for the measurement of the search results. Naturally, a portion of this measurement could be subjective. For example, if you ask a typical Internet user which search engine they prefer, you usually receive a very definite answer. The follow-up question, regarding why the user prefers one search portal over another, is difficult for the user to verbalize or to quantify. How does one measure and assess the search process in terms of the quality of the search? Some common factors are as follows:

1. efficiency and effectiveness—measured according to how much effort the user must spend on searching in order to locate the desired information and how many times the user has to type/retype and revise/refine the search terms;
2. relevance—measured according to the percentage of entries in the returned list which meet the user's expectation and how the entries are ordered or ranked in the returned list such that the relevant information is easily accessible.

It is well known that search results on the Web have a high duplicate rate in terms of data sources in certain domains or specific areas. For example, the national and international news coverage is often duplicated by various sites mirroring the major news carriers. Due to the vast amounts of information on almost all topics, one cannot systematically review the entire set of results; therefore, the user must rely on the ordering of the results. According to the study of attention psychology, the estimates for the length of human-attention span are highly variable and range from 30 s–1 min per year of age in young

children about the age of 12 to a maximum of around 20 min in adults [17]. It is therefore often the case that when the first several results are not relevant to what the user desires, then the user would prefer to refine the search terms and simply try again rather than scroll down the list to find the useful results; users feel that the latter just wastes time. Users thus often perform frequent research and reretrieval based on the current results in order to narrow down the outcome.

This paper presents an investigation of measurement metrics that can be used for assessing and evaluating the relevance of the search results. With the use of these metrics, Web users might more easily measure their search performance. This paper is based on the findings that the quality (or improvement) of a search can be evaluated according to the following factors.

1. The number of times that the user has to revise/retype his/her query to get the desired (target) information—assuming that the information is available somewhere. Finding relevant information from the Web is often an iterative process. However, it is not in the user's best interest to repeatedly try different words and to have to think about what different combinations of these keywords might yield the best information.

2. How easily the user can find the desired entry from the list of entries returned by the search process—the desired/target entries should appear at the top of the result list, with the most relevant preferably the topmost result.

3. How many relevant entries are shown on the first page of the search results; users tend to not want to advance to the next page of results because it is time consuming.

4. How many entries the user has to browse through in order to find the desired sites from the list.

The problem of search relevance touches upon the aspect of how a Web-search process (a search engine, a search query, or a function/feature applied in the search process) alleviates the information-overloading problem (not presenting excessive irrelevant information) and returns the information that the user needed or intended to retrieve. The outcome of our investigation is not intended to be another search-engine evaluation methodology. We show in Section IV that others are working on this problem. Rather, we want to explore a way that can be used to assist the users in improving their Web-searching effectiveness, such as determining how search terms can be better expressed so as to generate a better result (more accurately defined), how query processing can be improved (e.g., “quick” versus “advanced” search), and how differences in the search terms can be utilized to affect the desired retrieval results. Our objective is to quantify the quality, in terms of efficiency and effectiveness, of Web-search-based information retrieval. This, in turn, will be used to explore means whereby searching can automatically expand the query expression to place proper annotations on the query and the target document or information source, to insert more semantic constraints to the query, to semantically analyze and tag the information resources, etc. That is, the outcomes of this paper can be applied, with respect to the techniques and approaches, to achieve the following.

1. Improve the search efficiency by reducing the number of tries (requery and revising queries) that a user needs to use in order to retrieve pertinent information from the Web.

2. Reduce the complexity of queries that the user has to configure to get the information that he/she needs.

3. Improve the search effectiveness by reducing the complexity of query construction, such that a simple query will cover a range of relevant terms and retrieve/return a sophisticated set of relevant information entries, i.e., improving the chances of hitting the desired information from each search.

4. Improve the quality of search-result presentation by ordering the return entries in terms of the following: 1) relevance to query according to the query semantics and information/document tagging and 2) clusters of result entries. The most relevant (or user intended) entries of search result should be presented on the top of the list and most visible to the user.

We assume that the following are true in the study presented in this paper.

1. When assessing the search quality, we assume that the data resources (Web) are not intentionally smeared/corrupted or skewed/distorted.

2. The effectiveness and efficiency of a Web search is determined by how much relevant/useful information that a user obtains among the number of returns that the user checks—not on how many links that the search process returns nor on the percentage of relevant information out of the total information that the process returns.

Section II reports in more detail the motivation behind the need for these measurement tools, describes the aim of this paper, and details previous work in this research area. Section III covers several items—the development of our index function, a user survey that we conducted in order to obtain raw data, and the results of applying the function to the survey data. One set of measurements in particular is described, and we demonstrate in comparison that our metric is dissimilar and serves a different purpose. Our conclusions are given in Section IV. An Appendix to the document contains a sample of our user survey.

II. Background, Motivation, and the Focus

A. Background and Motivations

As it was pointed out by Bar-Ilan *et al.*, "Searching is a major activity on the Web, and the major search engines are the most frequently used tools for accessing information. Because of the vast amounts of information, the number of results for most queries is usually in the thousands, sometimes even in the millions. On the other hand, user studies have shown that users browse through the first few results only. Thus, results ranking is crucial to the success of a search engine." [2].

Current commercial search engines (such as Google) tend to return to the top of the list the entries that have been visited by more Web users, along with some other attributes of the data resources (Web pages) [5]. This strategy does not necessarily guarantee the retrieval of data that the user intended, particularly when searching for data sources and documents with semantic concerns (such as scientific and professional articles). Overall, the following scenarios have been observed.

1. Search engines have the potential to inadvertently narrow the user's search to an area other than the one of interest, simply because the correct keywords were not utilized in the search process.

2. The use of keywords can be ambiguous. One or more keywords, even when combined, may have more than one meaning. One way to deal with this ambiguity is to search by meaning (semantics) rather than by keyword.

Instead of just entering keywords, which may be ambiguous, the users may need to have a means to somehow specify the context or concept in which their interest lies. That is, potential ways of improving search include allowing the user to specify more precisely the context, concept, or idea that they are searching for and narrowing the search results more accurately to what the user originally desired [1], [19], [22]. For example, the word "tank" has a vastly different meaning to an army general than a cattle rancher. Enabling the search engine to understand terms in the proper context should yield improved results for the user's queries. Certain Internet companies already rely on a form of this context-sensitive searching [14]. Techniques to consider include indexing methods, ontology methods, collaborative tagging, theme detection, recommender systems, knowledge structures, use of tailored agents, extraction methods, clustering methods, standards, and others [11], [15], [25].

Thus, the following two tasks are necessary in order to measure whether these improvements will work and yield better search results.

1. We need measures and metrics for search-result quality assessment and evaluation.
2. We need to conduct experiments and tests by applying these measures to carefully selected queries and then monitor the results.

According to our study, the most popular assessment approaches include the following: 1) relevance of the results—based on human judgment; 2) randomly selecting "pseudorelevant" information—not based on human judgment; and 3) relevance in terms of known information contents and targets. Instead of human judgment, there is a need for the third approach to rely on the use of benchmarks and standard testing resources.

B. Focus

It is known that whether a returned result is a match or a mismatch to the query is obviously a subjective matter. However, in order to gauge the effectiveness of new search methodologies, one must have some means whereby the subjective relevance given by an individual can be translated into an overall "search-quality" result in order to determine whether different evaluation methods are truly effective. The measurement index function that we are looking for should thus be as follows:

1. independent of the search engines used;
2. simple to calculate in an "online" manner;
3. able to generate an appropriate overall value based upon individual result rankings;
4. able to assign appropriate weights, where top-listed correct results are valued highly.

The creation of this index function is the main objective of this paper and will be presented in the following sections.

Many aspects need to be considered when evaluating the efficiency and effectiveness of a search engine or a search process. These include the time that is necessary to execute the search, how accurate the

search is in terms of selecting the correct usage of the terms, what percentage of retrieved information is useful to the user, and so on. We do not consider the search time in our measurements, although we do note that the response time should be reasonable and stable, as pointed out by studies in human–computer interaction [7]. The former criterion is of greater concern to us, namely, how many of the total hits were relevant, how to define a hit, how to make an objective measurement, and how many were faulty results. The raw quantity of results returned by the search process should not be considered. Further items to consider include whether the search missed any target information—some data that should be found and returned by the searching but were not returned. This is difficult to judge, unless you know the data sources in advance and that the query is correct.

However, as we mentioned, the usefulness of a returned entry is a subjective matter. With the same search query and search returns, user Al may find more useful entries than Betty. An entry that is useful to Al may be of no use to Betty and vice versa. Thus, we need to collect numeric counts from many users over a period and form a statistical measurement to judge the effectiveness of a search process.

An additional factor to consider is how to present the data to the user so that they can quickly locate the information desired in order to prevent information overloading. If the users can retrieve the desired information rapidly, from the first few returned results for example, then the search was more efficient. If the correct result is, in fact, located by the search engine but is presented far down within the list of results, then the user must still examine many results within the list to determine whether the desired data are present. This requires them to examine a link and skip the many results that are irrelevant. The search in this case is inefficient if there are no desired results but is also inefficient, even if the results are eventually located.

Initially, one can consider the quantity of results returned to the user and the (potentially much smaller) quantity of results that were useful. Consider the following: If a user is looking for a specific article dealing with a certain topic, how many articles does the user have to browse through from the returned set before locating the correct or a pertinent article? Alternatively, if the user is looking for a particular Web site, or a kind of Web site, how many useful Web links were retrieved relative to the total number presented by the search? We can call this metric the Relevance-of-Searching-on-Target (RoSoT)—which could be a percentage count, a probability estimate, or a statistics evaluation.

It has been difficult in the past to judge the efficiency of information retrieval and the effectiveness of Web searching in terms of at which order and position the relevant and useful entries are returned. For example, if Al found that search-result numbers 1 and 3 were interesting for one query, is this more or less successful than Betty finding all of entries 2, 5, and 6 that are relevant for the same or some other queries? Currently, existing measurement techniques mentioned earlier do not address this question.

The need for a coherent index of search-quality (relevance) measurement is also evidenced by the fact that increasingly diverse information sources are used, and this causes the search results to be more deluged and overwhelming. While the quality and capabilities of search engines are improving over time, by, for example, learning from the user's past behavior, the effectiveness of these improvements becomes harder and harder to judge if without a quantitative measurement. As the search engines focus more on what the user needs, it is important to have an effective means to measure and assess the quality of these improvements overall or individually. Such measurement and evaluation, as our coherent index serves, will also provide meaningful hint and guidance for further efforts.

C. Previous Work on Measurement Metrics

There are many prior literature works dealing with search performance but mostly from the standpoint of evaluation of the search engines themselves and not from the perspective of the user. In fact, entire books have been written, which detail the algorithms used by Google and others [12], and research papers such as by Su *et al.*[21] specifically use metrics to not judge the search results but the user satisfaction with the search engine. Thus, prior work generally deals with the question of whether "search engine *X* is better than search engine *Y*." We first mention several of the available research papers in the area of measurement metrics and then look at one in particular for comparison purposes.

Bar-Ilan *et al.* [2] presented a number of measures that compare rankings of search-engine results. They applied the measures to five queries that were monitored daily for two periods of 14 or 21 days each. Rankings of the different search engines (Google, Yahoo!, and Teoma for text searches and Google, Yahoo!, and Picsearch for image searches) were compared on a daily basis in addition to longitudinal comparisons of the same engine for the same query over time. The results and rankings of the two periods were compared as well. In a separate paper, Bar-Ilan *et al.* [3] also measured how similar were the rankings of search engines on the overlapping results. The said paper compared the rankings of results for identical queries retrieved from several search engines. The method was based only on the set of URLs that appeared in the answer sets of the engines being compared. For comparing the similarity of rankings of two search engines, the Spearman correlation coefficient was computed. When comparing more than two sets, Kendall's *W* (coefficient of concordance) was used. These were well-known measures, and the statistical significance of the results could be computed. The methods were demonstrated on a set of 15 queries that were submitted to four large Web-search engines. The findings indicated that the large public search engines on the Web employ considerably different ranking algorithms.

In a later work by Bar-Ilan *et al.* [4], the authors investigated the similarities and differences between rankings of search results by users and search engines. Sixty-seven students took part in a three-week-long experiment, during which they were asked to identify and rank the top ten documents from the set of URLs that were retrieved by three major search engines (Google, MSN Search, and Yahoo!) for 12 selected queries. The URLs and accompanying snippets were displayed in random order without disclosing which search engine(s) retrieved any specific URL for the query. The authors computed the similarity of the rankings of the users and search engines using four nonparametric correlation measures that complement each other. The findings showed that the similarities between the users' choices and the rankings of the search engines were low. The authors also examined the effects of the presentation order of the results and of the thinking styles of the participants. Presentation order influenced the rankings, but overall, the results indicated that there was no "average user," and even if the users had the same basic knowledge of a topic, they evaluated information in their own context, which was influenced by cognitive, affective, and physical factors. This was the first large-scale experiment in which users were asked to rank the results of identical queries. The analysis of the experimental results demonstrated the potential for personalized search.

Awad and Khan [1] studied the issue of how to use multiple evidence combination and domain knowledge. The prediction is also used by Varadarajan *et al.* in [22], which discusses their findings on the degradation of the quality of search results across multiple pages and proposes a technique to generate composed pages by extracting and stitching together relevant pieces from hyperlinked Web pages. Their

work considers ranking the composed pages with an experimentally evaluated heuristic algorithm for efficiently generating the top composed pages. The work by Smyth [19] focused on applying the recorded user's search activities—the queries that they submit and the results that they select—to build a relevance model to improve Web search. The collaborative Web-search approach provides a valuable form of search knowledge and makes adaptive search possible. Lancieri and Durand [11] investigated Internet-user behavior through a comparative study of the access traces and application to the discovery of communities. It presented a comparative analysis of Internet-navigation traces (URLs versus keywords) to characterize individual or group-of-users' behavior when accessing the Web, based on the study of access redundancy and from the angle of time evolution. In [15], a seamless combination of context and policy to manage behaviors that Web services expose during composition and in response to changes in the environment was exposed. In this approach, behavior management and binding are subject to executing policies of types permission, obligation, restriction, and dispensation. A hot research topic of how to configure Web services to meet the demands of user requirements and limitation of resources is considered by Xiong *et al.*[25]. An optimal algorithm was presented to help choose the best configuration with the highest quality of service to meet users' nonfunctional requirements.

A frozen 18.5-million-page snapshot of part of the Web has been created by Hawking *et al.* [10] and touted by Voorhees [24], among others, in order to enable and encourage meaningful and reproducible evaluation of Web-search systems and techniques. This collection is being used in an evaluation framework within the Text Retrieval Conference (TREC) and will hopefully provide convincing answers to questions such as, “Can link information result in better rankings?,” “Do longer queries result in better answers?,” and “Do TREC systems work well on Web data?” The snapshot and associated evaluation methods are described, and an invitation is extended to participate. Preliminary results are presented for an effectiveness comparison of six TREC systems working on the snapshot collection against five well-known Web-search systems working over the current Web. These suggest that the standard of document rankings produced by public Web-search engines is by no means state of the art.

Soboroff and Cahan [20] have used the TREC data to create “a new evaluation methodology that replaces human relevance judgments with a randomly selected mapping of documents to topics.” Hawking *et al.* [9] evaluated the effectiveness of 20 public search engines using TREC-inspired methods and a set of 54 queries taken from real Web-search logs. The Web is taken as the test collection, and a combination of crawler and text retrieval systems was evaluated. The engines were compared on a range of measures derivable from binary relevance judgments of the first seven live results returned. Statistical testing reveals a significant difference between engines and high intercorrelations between measures. Surprisingly, given the dynamic nature of the Web and the time elapsed, there was also a high correlation between the results of this paper and that of a previous study. For nearly all engines, there was a gradual decline in precision at increasing cutoffs after some initial fluctuation. The performance of the engines as a group is found to be inferior to that of the group of participants in the TREC-8 Large Web task, although the best engines approach the median of those systems. The shortcomings of the current Web-search evaluation methodology were identified, and recommendations were made for future improvements. In particular, this paper and its predecessors dealt with queries that were assumed to derive from a need to find a selection of documents that were relevant to a topic. By contrast, real Web searches reflected a range of other information-need types that require different judging and measures.

Vaughan [23] proposed a set of measurements for evaluating Web-search engine performance. Some measurements were adapted from the concepts of recall and precision, which were commonly used in evaluating traditional information retrieval systems. Others were newly developed to evaluate search-engine stability, an issue that is unique to Web information retrieval systems. An experiment was conducted to test these new measurements by applying them to a performance comparison of three commercial search engines: Google, AltaVista, and Teoma. Twenty-four subjects ranked four sets of Web pages, and their rankings were used as benchmarks against which to compare search-engine performance. Results show that the proposed measurements are able to distinguish search-engine performance very well.

In [13], again, the topic was Web-search engines, particularly the challenges in indexing the World Wide Web, the user behavior, and the ranking factors used by these engines. Ranking factors were divided into query-dependent and query-independent factors, the latter of which had become more and more important in recent years. The possibilities of these factors were limited, mainly of those that were based on the widely used link popularity measures. The paper concluded with an overview of the factors that should be considered to determine the quality of Web-search engines.

McCown and Nelson [16] provided the first in-depth quantitative analysis of the results produced by the Google, MSN Search, and Yahoo! application-programming and Web-user interfaces (APIs and WUIs, respectively). Google, Yahoo!, and MSN Search all provided both WUIs and APIs to their collections. Whether building collections of resources or studying the search engines themselves, the search engines request that researchers use their APIs and not “scrape” the WUIs. However, anecdotal evidence suggests that the interfaces produced different results. The authors had queried both interfaces for five months and found significant discrepancies between the interfaces in several categories. In general, they found MSN Search to produce the most consistent results between their two interfaces. The findings suggest that the API indices are not older, but they are probably smaller for Google and Yahoo!. The authors also examined how search results decay over time, and built predictive models based on the observed decay rates. Based on the findings, it can take over a year for half of the top ten results to a popular query to be replaced in Google and Yahoo!; for MSN Search, it may take only two to three months.

In evaluating search engines for the Web, Chu and Rosenthal [18] described their research method in very simple terms: “Relevance of retrieved Web records was determined separately by both authors on the basis of the up to 10 Web records we downloaded for each query.” The authors did not try to read the full-text Web documents by following the links provided because of time considerations and reliability of the Web linkages. In order to delineate the overall performance of each Web engine that they examined, the authors not only computed precision scores for each individual query but also calculated average precision among all ten searches for every search engine included in the study. In addition, the authors tabulated the mean precision for each sample query so that some light could be shed on the suitability of using Web-search engines for certain questions.

One focus of the work by Beg [6] was to “outline a procedure for assessing the quality of search results obtained through several popular search engines.” To accomplish this, they “watch the actions of the user on the search results presented before him in response to his query and infer the feedback of the user therefrom.” In this approach, the actual activity of the user is monitored, and the quality of the search results is inferred from the actions of the user. This is an “implicit” ranking that is then used to

assess the quality of the search results by the various search engines under study. Our work does not rely on this implicit inference; we actually use the user survey to assess each document returned in the search.

III. Coherent Index for RoSoT

In order to evaluate potential index functions for the measurement of RoSoT, first, it is necessary to initially conduct research in the form of surveys to gather user data. Second, it is necessary to propose several potential RoSoT index functions. Third, we need to determine which of the index functions is appropriate. The latter is accomplished by evaluating the functions relative to the initial data collected from the user surveys.

A. Data Collection

To establish a proper RoSoT measurement indexing function, we first surveyed students at our campus since these data are easy to obtain. A copy of the survey form is shown in the Appendix. The survey asked the students to first select a search term that was something that they were familiar with but had not looked up on the Web prior to taking the survey. Suggestions included solar power for homes, the birth date of certain celebrities, and so on. They were then asked to rank each of the search results on a scale from 1 to 5, according to the following scale.

1. The link was broken or had nothing to do with the search terms.
2. The search terms were included in the page, but it was not something that really pertained to the search.
3. It pertained to the search but was not that interesting relative to what I was looking for.
4. It is interesting and, for the most part, seems to match the search terms.
5. It is exactly what I was searching for.

Duplicate entries were marked on the form, and in the process of data entry, we discounted these to a result of 1; effectively, they are thus counted as if they were not relevant. Students were also asked to make sure that they addressed the “overall” rank at the top of the page. This represents an indication of the quality of the search, again with 5 as the highest and 1 as the lowest:

$$
\begin{aligned}
1 &= \text{The search yielded no URLs that matched.} \\
2 &= \text{The search yielded a few that matched,} \\
&\quad \text{but I had to hunt for them.} \\
3 &= \text{This was a pretty “average” search result.} \\
4 &= \text{The search returned mostly what I wanted;} \\
&\quad \text{many pages were relevant.} \\
5 &= \text{The search returned what I wanted.}
\end{aligned}
$$

The experiment is thus to take the *individual result rankings* and attempt to use our function to predict the *overall quality ranking* selected by the survey participant. In this way, we expect that, in the future, a value resulted from simply ranking the individual results can be utilized to judge whether one search technique is better than another.

We were also careful to separate the survey users into groups in order to check whether there were variances within groups. For example, we can attempt to determine the following thoughts.

1. Do Ph.D. students have higher (or lower) expectations than others?
2. Do undergraduate students have higher (or lower) expectations than others?
3. Is there a difference between entry-level university students versus juniors and seniors?
4. Do these results match "others" outside of the aforementioned groups?

Our hope is that our index function accurately represents results "across the board" in all groups. However, this may or may not be the case, and any variance within user communities would need to be accounted for.

B. Selection of a Uniform Index Function for RoSoT

Once the survey results were collected, our research focused on the definition of a function that matched this sample data as closely as possible. Thus, we developed a set of numeric functions that measure the qualitative results of the Web searches. As noted, previous research on search-engine evaluation relies on comparisons—comparing how the same entries appear at different places of different search processes. There is no metric or uniform index given for each search result, independent of other search attempts. We hope that, by using our "uniform index function," we can give a quantitative evaluation for each search process and the returned result, not based on comparisons but simply based on the orders of the entries returned and their relevance to user interest. Specifically, the search index must be able to achieve the following.

1. Place a high value in terms of the order of entries returned.
2. Distinguish totally relevant or partially relevant entries.
3. Distinguish duplicate entries.

The first of these is critical: A search that a user deems successful involves no need to modify the search terms and yields a result at or very close to the top of the results. The number of search results *N* is the parameter to the uniform index function. An index function with a relatively slow decrease is not preferred because it does not fall off rapidly with larger values of *N*.

Our selection of the relevance index function is determined according to the psychological study of human cognition and attention. It is known that human attention in a function is a reverse order with respect to the time line [8], [17]. Considering the time spent on browsing the search results, it increases proportionally with the number of search results *N*. That is, human attention decreases as the number of search results to be evaluated by the user increases. Thus, our initial study looked at a number of potential functions and then focused on the following three functions:

1. $X(N) = 1/N$;
2. $X(N) = 1/N^{1/2}$;
3. $X(N) = X(N-1) * D$; $X(1) = 1.0$, with constant $D < 1.0$.

It can be seen that functions 1) and 2) are based on qualitative understanding. They differ in terms of the delineation of their exponential factors, giving differing quantitative lines. This is because the psychology study only provides a qualitative principle of the attention span, and thus, no exact quantitative factor can be drawn from them. Function 3) is based on the principle that attention fades (decreases) in proportion to a factor that is relevant to time—here the amount of entries viewed/examined.

These index functions are parameterized such that multiple relevant entries far down in the search results can be valued higher than an initial correct single entry. This implies that searches with a high number of returned results would have an advantage over searches with fewer entries, even if the latter is more accurate. However, an index function with a too rapidly decreasing order (e.g., $X(N) = X(N-1) * D$, with $D \leq 0.5$) applies too much of a penalty on entries below the top result.

Since the index function that we look for measures RoSoT, we call it the RoSoT index.

Result	$X(N) = D^{(N-1)}$	$X(N) = 1/N$	$X(N) = 1/N^{1/2}$
1	1.0000	1.0000	1.0000
2	0.7549	0.5000	0.7071
3	0.5699	0.3333	0.5774
4	0.4302	0.2500	0.5000
5	0.3248	0.2000	0.4472
6	0.2452	0.1667	0.4082
7	0.1851	0.1429	0.3780
8	0.1397	0.1250	0.3536
9	0.1055	0.1111	0.3333
10	0.0796	0.1000	0.3162

Table 1: Values for Sample Index Functions

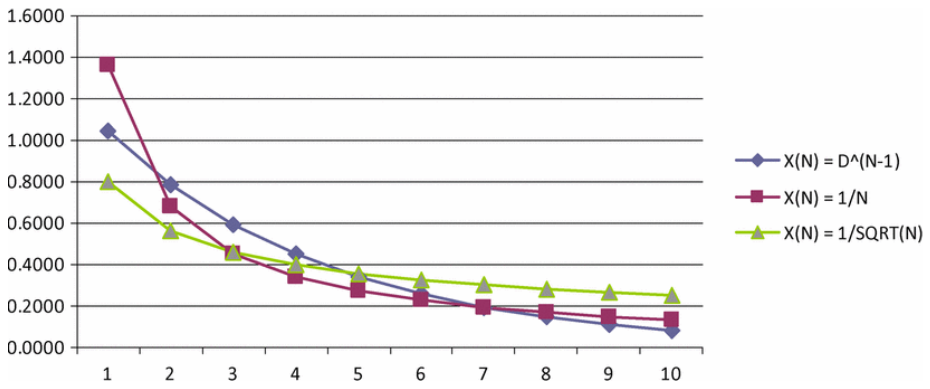

Figure 1: Index functions and various *N*'s.

To better determine which index function fits more properly with the user cognitive characteristics in terms of the survey results, we desire to resolve first a set of criteria instead of repetitively trying different functions. Among the important criteria for our index function, we consider the following.

1. $X(N) \geq X(N+1) > 0.0$. That is, a matching entry N should always be better than the next result.
2. $X(N) < X(N+1) + X(N+2)$. That is, an entry is not necessarily better than two subsequent matching results.
3. $X(N) \geq X(N+2) + X(N+3)$, i.e., entry N should be better than two results at two positions farther down.
4. $X(N) < X(N+2) + X(N+3) + X(N+4)$.
5. $X(N) \geq X(N+3) + X(N+4) + X(N+5)$.
6. $X(N) < X(N+3) + X(N+4) + X(N+5) + X(N+6)$.
7. In general, $X(N) \geq \sum_{i=A}^{2A-1} X(N+i)$ and $X(N) < \sum_{i=A}^{2A} X(N+i)$ for $A = 1, 2, \ldots$.

General property 7), as well as 4), 5), and 6), is automatically satisfied and guaranteed when properties 2) and 3) are satisfied; we only need to find a function that satisfies properties 2) and 3). We also note that, obviously, property 1) is satisfied when 2) and 3) are satisfied.

Through the study, we found the following functions:

1. $X(N) = e^{-(N-1)/C}$, with $2.078 \leq C \leq 3.556$;
2. $X(N) = D^{(N-1)}$, i.e., $X(N) = X(N-1) * D$, with $0.618 \leq D \leq 0.755$, matches the aforementioned requirements the best.

In fact, we have $e^{-(N-1)/c} = D^{(N-1)}$ when $C = -1/ln(D)$ or $D = e^{-1/c}$. Thus, either of the two formats can be utilized to describe the same idea.

Result	$X(N) = D^{(N-1)}$	$X(N) = 1/N$	$X(N) = 1/N^{1/2}$
1	1.0431	1.3657	0.7967
2	0.7874	0.6829	0.5634
3	0.5944	0.4552	0.4600
4	0.4487	0.3414	0.3984
5	0.3388	0.2731	0.3563
6	0.2557	0.2276	0.3253
7	0.1930	0.1951	0.3011
8	0.1457	0.1707	0.2817
9	0.1100	0.1517	0.2656
10	0.0830	0.1366	0.2519

Table II: Normalized Values for Sample Index Functions

Let *X(N) < X(N + 1) + X(N + 2)* for *X(N) = D(N−1)* function, in which case we have *D(N−1) < D(N) + D(N+1)*. Solving for *D*, we have *D2 + D − 1 > 0* for a position valued solution, with *D >0.618*. Using the relation *C = −1/ ln(D), C > 2.078* for *X(N) = e−(N−1)/C* to satisfy *X(N) < X(N + 1) + X(N + 2)*.

Let *X(N) ≥ X(N + 2) + X(N + 3)* for the *D(N−1)* function, in which case we have *D(N−1) ≥ D(N+1) + D(N+2)*. Solving this equation for *D, D3 + D2 − 1 ≤ 0*, and for a position valued solution, we have *D ≤ 0.755*, and *C ≤ 3.556*. That is, when D >0.618 and C > 2.078, we haveX(N) < X(N + 1) + X(N + 2) for the e−(N−1)/C and D(N−1) functions. When D ≤ 0.755 or C ≤ 3.556, we have X(N) ≥ X(N + 2) + X(N + 3) for the e−(N−1)/C and D(N−1) functions.

Table I shows the relative weights for the three index functions, with each row in the table representing successive URL results.

Fig. 1 shows the relative values obtained from Table I in a graphical format.

The summary value of the relevant entries is the index of the search. For example, if user Al finds entries 1 and 10 in his interest, the total score using the *X(N) = D(N−1)* function is *X*(1) + *X*(10) = 1.0 + 0.080 = 1.080; i.e., this search has a quality (or success—efficiency and effectiveness) index of 1.080. All three candidate functions are used in the same manner for our testing. However, for practical reasons, each function in this paper is normalized such that $\sum_{N=1}^{10} X(N) \approx 4.0$. This is accomplished by scaling *X(N) = D(N−1)* by 1.0431, *X(N) = 1/N* by 1.3657, and *X(N) = 1/N1/2* by 0.7967, as shown in Table II.

Some observations of this index function include the following.

1. The application of the numeric assignment allows us to measure the search quality/improvement without depending on how many total entries have been examined. The

value is determined by how many entries are of interest to the user and what their relative positions are within the results.

2. When comparing the results of two searches, examining more results does not penalize the score generated by the function. We do not need to take into account how many entries were examined. The total number of entries looked at should not be much of a significant factor—as long as a certain number of relevant/interest entries have been located.

3. No averaging is needed.

4. How much effect should the latter entries have on the overall quality of the search? The effect should be small but not completely insignificant or zero.

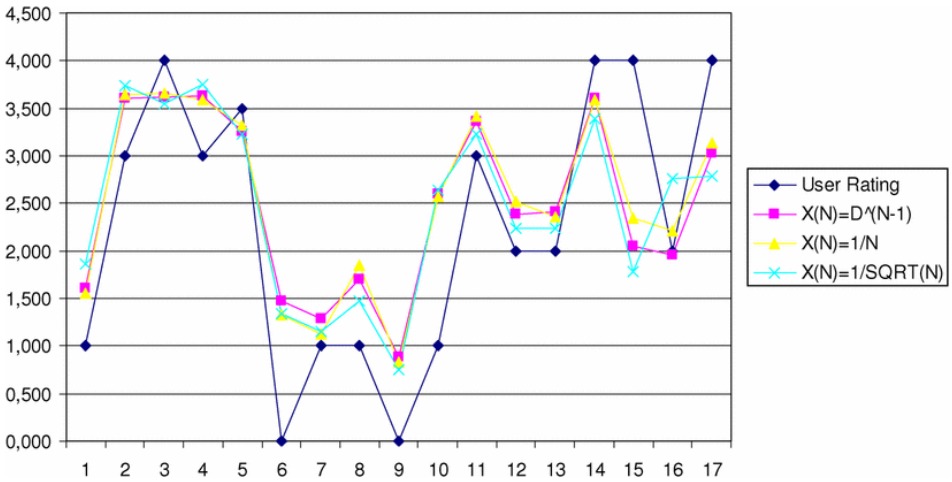

Fig. 2. Ph.D.-student response versus index ranking.

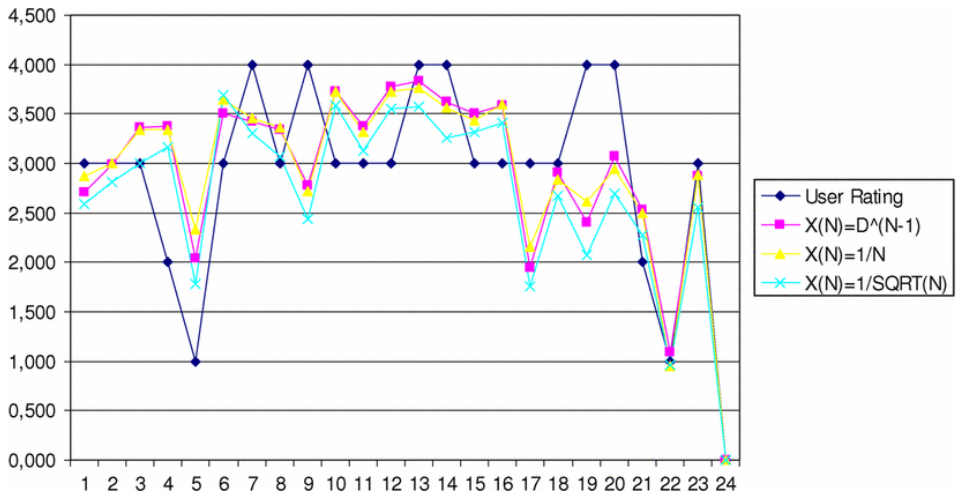

Figure 3: First-/second-year undergraduate student response versus index ranking.

C. Results, Analysis, and Insights

After obtaining the results from user surveys, the respondents were classified into groups according to the aforementioned criteria. In order to normalize the data, we adjusted the user rankings from a 1–5 scale to a 0–4 scale in order to match the indexing function; since few users examined more than the first ten URL results, and since the total of the rank index for 1–10 is made 4.0 by scaling, this seemed reasonable.

Shown hereinafter is a graphical representation of the results when Ph.D. students were asked to take the survey. In addition to the actual rankings made by the Ph.D. students, we show the three functions detailed earlier for comparison purposes.

The three functions tested differ primarily in the rate at which they decrease as more URLs are considered. They also differ in the weight assigned to the initial sample returned from the browser, with a heavy weight being specifically assigned to the initial result by the $1/N$ algorithm. With respect to the survey data, we expected that the initial search result would be the most heavily weighted, but of course, it is not clear without studying the extent to which this weighting would affect the overall match of the function.

Fig. 2 shows one of the sample group's responses that were generated by the Ph.D. students in our department. Upon the initial examination, it appears that all three functions are relatively good at predicting the user's overall impression of the search. One thing to notice about these results is that all three functions tested tend to overestimate the satisfaction when the search results are deemed completely “bad” by the user. Several people conducted searches (in Fig. 2, consider users 6 and 9) and indicated that the results were of no use. However, based on the individual responses URL-by-URL, all three functions predicted at least some satisfaction with the outcome; the users did not feel this way.

Thus, from the graphical representation, it is clear that the candidate functions do, in general, represent the responses from the Ph.D. students, albeit with some issues stated previously; the functions tend to overestimate when the user indicates poor individual search results. We repeated the experiment with a combination of first- plus second-year students, third- plus fourth-year students, and others. These results are shown in Figs. 3 and 4.

From the figures, it is clear that each function does a fair job of predicting what the user indicated was the value of the search. Table III summarizes the errors in the three candidate functions. The mean error is simply the average difference between what the function predicted versus what the user indicated on the survey. The deviation is $(1/n)\sum x - \overline{x}|$, where, again, x is the error quantity

From Table III, we can see that the overall closest function for predicting the user's responses is obtained by using the function $X(N) = D^{(N-1)}$. Our first observation is that users tend to rank the overall search results using extremes. Consider the “other” group shown earlier in Fig. 5. Of the 37 responses in this category, 20 rated the overall search results as either “1 = The search yielded no URLs that matched” or “5 = The search returned what I wanted” (this is normalized to zero and four in the graphs). Further research may prove that using the $X(N) = D^{(N-1)}$ index function could yield better results if we broaden the range of the function after the prediction so that these extremes are accommodated.

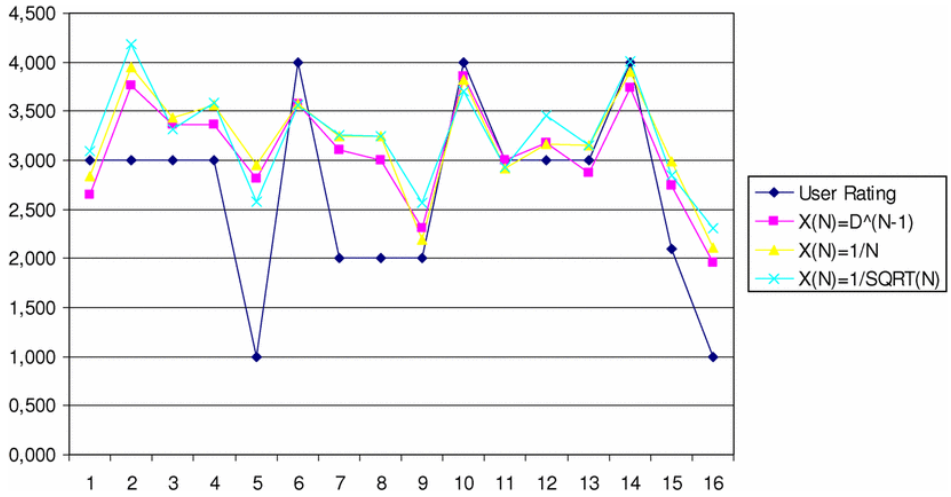

Figure 4: Third-/fourth-year undergraduate-student response versus index ranking.

Classification	Mean Error			Deviation of Error		
	$D^{(N-1)}$	1/N	$1/N^{1/2}$	$D^{(N-1)}$	1/N	$1/N^{1/2}$
1st / 2nd Year Undergraduates	0.569	0.566	0.596	0.349	0.357	0.395
3rd / 4th Year Undergraduates	0.550	0.615	0.643	0.370	0.462	0.433
Ph.D. Students	0.702	0.672	0.761	0.397	0.359	0.408
Others	0.634	0.638	0.674	0.350	0.361	0.363
Overall	0.616	0.622	0.664	0.363	0.378	0.397

Table III: Mean Error and Deviation for Three Index Functions

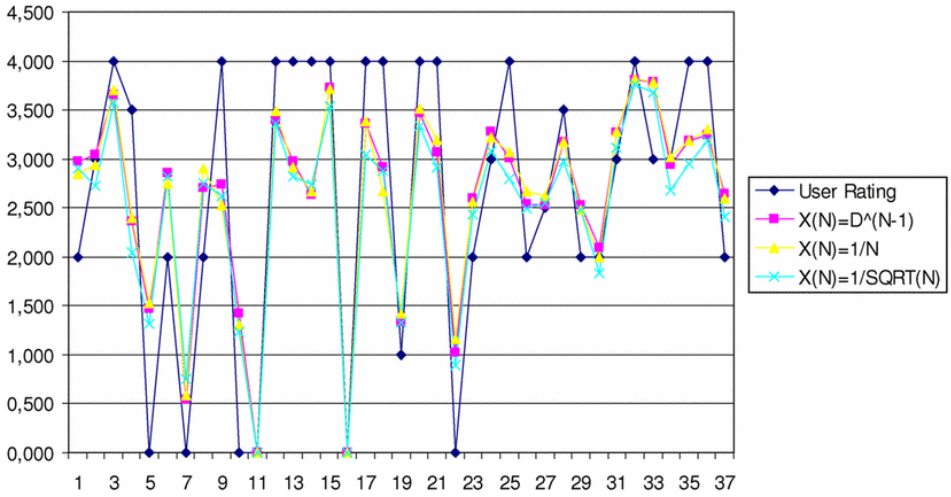

Figure 5: Other response versus index ranking.

We do notice the differences of the results among the surveyed user groups, namely, the "undergraduate-student" group, the "Ph.D.-student" group, and the "others" group. It is a bit surprising that the difference between the "undergraduate-student" group and the "Ph.D.-student" group is even bigger than the "undergraduate-student" group and the "others" group. However, these differences do not affect our selection of the best index function because the differences are consistent with respect to the three functions tested. On the other hand, even though we have our selection of the best fit of the function, the differences also tell us that there is room for improvement.

It is also seen from our experimental results that there is really no significant difference between the functions of $D^{(N-1)}$ and $1/N$. This could be a valuable insight for us to consider in our future study. It could lead us to the exploration of more rapidly decreasing functions or the adjustment of the parameters of the aforesaid functions in order to more closely fit the coherent user psychological behavior or attention models.

Functionality	Bar-Yossef & Gurevich's Measurements	RoSoT Index
Measuring the relevance to user intentions of search (information the user looking for)	NO	YES
Measuring the quantitative degree of matching of the search results with respect to user's intention of search	NO	YES
Measuring the web site index freshness and popularity	YES	NO
Measuring individual search result (i.e., exact measurement rather than statistical estimate)	NO	YES
Ranking the returned entries in terms of relevance	NO	YES
Measuring the size of returning corpus	YES	NO
Dealing with density of duplicates in the corpus	YES	YES
Dependent on the corpus size	YES	NO
Can be used to compare Quality of Search Engines	YES	YES
Can be used to compare quality of search queries	YES	YES
Dependent on knowledge of how search engine works	NO	NO
Affected by the use of filters on the search result	YES	NO
Human inputs incorporated	NO	YES

Table IV: Comparison of the RoSoT Index with Bar-Yossef and Gurevisch's Measurements

D. Comparison of Our RoSoT Index to That of the Work by Bar-Yossef and Gurevich

The prior research listed previously in this paper is principally concerned with the determination of "which search engine is better." None of these directly addresses the problem that we are describing here, i.e., the determination of user satisfaction based on the individual search results. One paper that is at least somewhat related to ours is by Bar-Yossef and Gurevich [5], which addresses the problem of measuring global quality metrics of search engines, including corpus size, index freshness, and the density of duplicates. The authors presented two new estimators that were able to overcome the bias introduced by previous research. Their estimators are based on a careful implementation of an approximate importance sampling procedure. Comprehensive theoretical and empirical analyses of the estimators demonstrate that they have essentially no bias, even in situations where document degrees are poorly approximated.

The paper by Bar-Yossef and Gurevich addresses the "degree mismatch problem," in which, with respect to a certain document, the number of queries that would (or should) retrieve the document is compared

against the number of queries that actually do retrieve the document. For example, given a pool of potential queries and a certain document x, the degree of x is |q| where $q \subseteq P$, and each query in q retrieves x. The authors use this primarily for an estimation of the effectiveness of the search engine; however, we note that the inverse is also of importance—for a certain query in the set q, which is known to retrieve document x, where in the overall results will x appear? This is where our research fits; the prior research focuses on estimating the degree, whereas we focus on the rank of x within the results of some query within q.

Duplicate density, which is briefly mentioned in the prior work, is associated with our research since we treat all but the first match in a duplicate set as a complete search miss. In our case, duplicate density has a direct impact on the RoSoT index result because of this elimination. In the case of Bar-Yossef and Gurevich, the focus is on estimating the duplicate density of the corpus maintained by the search engine; again, we are considering the opposite of this, the position ranking of the duplicate within the results of query q.

Finally, like our method, Bar-Yossef and Gurevich did not rely on specific knowledge of how the search engine works; they simply examined the results. However, they stated that these metrics were relevance neutral and that no human judgment is required for computing them; we are focusing on which documents are relevant to a user that obviously requires human judgment. Unfortunately, it is impractical to quantitatively compare their measurement metric with our RoSoT index due to the natural differences of the parameter sets and the measurement functions involved. At the same time, it is possible to compare and summarize the differences in the approaches between the Bar-Yossef and Gurevich method versus our RoSoT method; these differences are presented in Table IV.

Section IV: Conclusion

This paper has presented a technique for a prediction of the user's overall search value, based on the quality of individual search results as ranked by the user. Our research properly addresses the problem of how to measure the relevance of the Web-search results with respect to the searcher's interest and expectation. The RoSoT index represents a quantitative evaluation of the search effectiveness from the user's perspective. It can be utilized, for example, in the determination of whether search results are improved from the perspective of the users. The indexing function is not intended to be used for the comparison of different search engines but is better used as a quantitative evaluation for the overall search results. It is also independent of the search engine used, as it only describes the satisfaction of the user, not the effectiveness of the search engine itself (although these may, of course, be related). The index function is also stand-alone and does not rely on a corpus or database of information sources created for testing purposes, as some previous methods did.

Specifically, we cite the following strengths of our approach.

1. The function is based on human preferences; we do not rely on particular search engines or sources of experimental setting.
2. The function is stand-alone in the sense that no other data are needed other than the individual rankings of the returned entries by whatever search engine is used.

3. The function is simple to compute once the rankings of a selected number of returned entries are obtained, without the need to know the total number of entries returned and how many returned entries are relevant.
4. The function is intuitive—much more weight is given to correct results when they are encountered early in the returned set. This mimics the user's perspective, where an early good match is extremely important.
5. No standard database of test data is necessary. We also do not require any knowledge about the data themselves, as the function results are based purely on the user responses to individual search results.
6. The index function is an online function as opposed to any offline functions.

We do note the following weakness in our approach.

- The index function obviously requires the user to provide the worthiness (a ranking value) of individual returned results, i.e., by its very nature, subjective.

As another note, the concept about "the relevance measure of a Web search" is itself a subjective matter. It is known that, for a given search session (given set of keywords), the returned results would be "very relevant" or "relevant" to some people, but "less relevant" or "not relevant at all" to some other people, depending on the background, the context, and the intentions of the searcher. The merit of our evaluation index is that it takes into account this differentiation, so that the measurement is coherent to the nature of human cognitive behavior and attention psychology.

Nonetheless, we need to investigate further to determine whether such factors as age, income, education, social status, etc., have an impact on the search results and, in turn, whether the RoSoT function correctly predicts the results. At this time, we have utilized principally university students and others such as staff and faculty and have not researched the impacts of these other factors.

Our research is designed as a method to evaluate the quality of search results. In the future, we can utilize the index function to determine whether methods such as ontology-based Web-search methodologies and context-mediated queries are generating improved or inferior search results. Not like some other research in this area that was designed for the comparison of search engines and often used standardized testing data sets, the RoSoT index function is applicable to a broad range of situations where an assessment of a Web-search quality needs to be independent of the specific search engine and specific data sources on which the function is tested.

Footnotes

This paper was recommended by Associate Editor Y. Wang.

The authors are with the University of Nebraska at Omaha, Omaha, NE 68182 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Appendix

Table V: Sample Data-Collection Form

Search Quality Evaluation Data Collection Sheet

Date: _____

Search engine: _____

Search terms: _____

Overall rating:_____

For the overall rating, use this scale:
1 = The search yielded no URLs that matched.
2 = The search yielded a few that matched, but I had to hunt for them.
3 = This was a pretty "average" search result.
4 = The search returned mostly what I wanted; many pages were relevant.
5 = The search returned what I wanted.

URL #	Measure of relevance	Is it a duplicate?	Comment
	1. The link was broken, or had nothing to do with the search terms. 2. The search terms were included in the page, but it was not something that really pertained to the search. 3. It pertained to the search, but was not that interesting relative to what I was looking for. 4. It is interesting and for the most part seems to match the search terms. 5. It is exactly what I was searching for.		
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			

References

[1] M. A. Awad and L. R. Khan, "Web navigation prediction using multiple evidence combination and domain knowledge," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 6, pp. 1054–1062, Nov. 2007.

[2] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, "Methods for comparing rankings of search engine results," *Comput. Netw.*, vol. 50, no. 10, pp. 1448–1463, Jul. 14, 2006.

[3] J. Bar-Ilan, "Comparing rankings of search results on the Web," *Inf. Process. Manag.*, vol. 41, no. 6, pp. 1511–1519, Dec. 2005.

[4] J. Bar-Ilan, K. Keenoy, E. Yaari, and M. Levene, "User rankings of search engine results," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 9, pp. 1254–1266, Jul. 2007.

[5] Z. Bar-Yossef and M. Gurevich, "Efficient search engine measurements," in *Proc. WWW*, Banff, AB, Canada, May 8–12, 2007, pp. 401–410.

[6] M. M. S. Beg, "A subjective measure of Web search quality," *Inf. Sci.*, vol. 169, no. 3/4, pp. 365–381, Feb. 2005.

[7] A. Dix, J. Finlay, G. D. Abowd, and R. Beale, *Human–Computer Interaction*., 3 ed. Englewood Cliffs, NJ: Prentice-Hall, 2004.

[8] D. Groome, *An Introduction to Cognitive Psychology: Processes and Disorders*. Philadelphia, PA: Psychology Press, Aug. 2006.

[9] D. Hawking, N. Craswell, P. Bailey, and K. Griffiths, "Measuring search engine quality," *Inf. Retr.*, vol. 4, no. 1, pp. 33–59, Apr. 2001.

[10] D. Hawking, N. Craswell, P. Thistlewaite, and D. Harman, "Results and challenges in Web search evaluation," *Comput. Netw.*, vol. 31, no. 11, pp. 1321–1330, May 1999.

[11] L. Lancieri and N. Durand, "Internet user behavior: Compared study of the access traces and application to the discovery of communities," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 36, no. 1, pp. 208–219, Jan. 2006.

[12] A. N. Langville and C. D. Meyer, *Google's Page Rank and Beyond: The Science of Search Engine Rankings*. Princeton, NJ: Princeton Univ. Press, 2006.

[13] D. Lewandowski, "Web searching, search engines and information retrieval," *Inf. Serv. Use*, vol. 25, no. 3/4, pp. 137–147, Jul. 2005.

[14] G. Linden, "People who read this article also read. . .," *IEEE Spectr.*, vol. 45, no. 3, pp. 46–60, Mar. 2008.

[15] Z. Maamar, D. Benslimane, G. K. Mostefaoui, S. Subramanian, and Q. H. Mahmoud, "Toward behavioral Web services using policies," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 6, pp. 1–1324, Nov. 2008.

[16] F. McCown and M. L. Nelson, "Agreeing to disagree: Search engines and their public interfaces," in *Proc. Conf. Digit. Libraries*, Vancouver, BC, Canada, Jun. 18–23, 2007, pp. 309–318.

[17] H. E. Pashler, *The Psychology of Attention*. Cambridge,MA: MIT Press, Jul. 16, 1999.

[18] H. Chu and M. Rosenthal, Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology, 1996. [Online]. Available: http://www.asis.org/annual-96/ElectronicProceedings/chu.html

[19] B. Smyth, "A community-based approach to personalizing Web search," *Computer*, vol. 40, no. 8, pp. 42–50, Aug. 2007.

[20] C. N. Soboroff and P. Cahan, "Ranking retrieval systems without relevance judgments," in *Proc. 24th Annu. Int. ACM SIGIR Conf.*, 2001, pp. 66–72.

[21] L. T. Su, H. L. Chen, and X. Y. Dong, "Evaluation of Web-based search engines from the end-user's perspective: A pilot study," in *Proc. ASIS Annu. Meeting*, 1998, vol. 35, pp. 348–361.

[22] R. Varadarajan, V. Hristidis, and T. Li, “Beyond single-page Web search results,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 3, pp. 411–424, Mar. 2008.
[23] L. Vaughan, “New measurements for search engine evaluation proposed and tested,” *Inf. Process. Manag.*, vol. 40, no. 4, pp. 677–691, May 2004.
[24] E. M. Voorhees, “Variations in relevance judgments and the measurement of retrieval effectiveness,” *Inf. Process. Manag.*, vol. 36, no. 5, pp. 697–716, Sep. 2000.
[25] P. Xiong, Y. S. Fan, and M. C. Zhou, “QoS-aware Web service configuration,” *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 4, pp. 888–895, Jul. 2008.