Journal of Web Engineering, Vol.3, No.3&4 (2004) 171-181

WHY IS WEB SEARCH SO HARD... TO EVALUATE?

DANIEL E. ROSE

Yahoo! Inc.
drose@yahoo-inc.com

Received July 26, 2004
Revised November 23, 2004

Web search has several important characteristics that distinguish it from traditional information retrieval: the often adversarial relationship between content creators and search engine designers, the nature of the corpus, and the multiplicity of user goals. In addition to making the search task itself difficult, these characteristics make it particularly hard to evaluate search effectiveness. In this paper, we examine these characteristics and then consider the problems with several different standard evaluation techniques.

Key words: Web search, evaluation, user satisfaction
Communicated by: A Spink and C Watters

1. Introduction

Web search engines are used for a growing variety of tasks by a growing number of people. As search engine designers work to keep up with the needs of their users, they often introduce new features or new algorithms designed to enhance the user experience. But how do we know whether these changes are helping or hurting users? Many different evaluation techniques have been suggested. Some date back to the earliest days of information retrieval, while others come from the realm of usability testing. Unfortunately, none of these techniques is entirely satisfactory when it comes to evaluating web search effectiveness.

We will examine four strategies for assessment – recall and precision, clickthrough, time on task, and surveys and user feedback – and consider the problems that these methods face for the web search task. But in order to understand why these traditional evaluation measures aren't easy to apply, we first need to look at what makes web search a particularly difficult problem.

2. What's So Hard About Web Search

There are many aspects of the web search problem that distinguish it from traditional information retrieval (IR). By "traditional information retrieval," we mean the world of pre-web computer-based systems that provide full-text search of a corpus of automatically indexed documents, and which have been the focus of several decades of research reported in forums such as the ACM SIGIR conferences and NIST's TREC competitions. In contrast, "web search" refers to systems that crawl and index pages on the World Wide Web. Some of these differences between web search and traditional IR are evident to anyone who has used a web search engine; others may only be known to those who have seen the work that goes on "behind the scenes" at web search companies.

2.1 An Adversarial Relationship

It is a well-known phenomenon that authors and readers – or any two people, for that matter – use different words to describe the same thing. This is known as the vocabulary problem. In one well-known study, users were asked to choose a name for something (e.g. a keyword describing a recipe). Over a variety of tasks, there was less than a 20% chance that they chose the same name [7]. In other words, if something is indexed using one person's favorite term, another person using his or her own favorite term as a query will fail to find the item more than 80% of the time. In the study, even the "best" name – the one favored by the largest number of users – still results in a 65% to 85% failure rate. The study authors summarized the results starkly:

> Simply stated, the data tell us that there is no one good access term for most objects. The idea of an "obvious," "self-evident," or "natural" term is a myth! Since even the best possible name is not very useful, it follows that there can exist no rules, guidelines, or procedures for choosing a good name, in the sense of "accessible to the unfamiliar user." [7]

Substituting the word "query" for "name" in the last sentence highlights one difficulty of the information retrieval task.

Librarians and other information professionals originally addressed this problem by using a controlled vocabulary for indexing, and/or having experienced editors assign index terms. Full text information retrieval systems did not have this luxury; their designers struggled to map the intent expressed in user queries to the appropriate documents.

Addressing this challenge was difficult enough when all parties involved wanted to help match user interests and authors' content as accurately as possible. But on the web, many content creators deliberately provide *misleading* content in order to get their pages to be viewed by more readers. An entire industry of Search Engine Optimizers (SEOs) has arisen, promising content providers higher ranking in search results. While some of these services are legitimate attempts to make sure that pages are found by their intended audience, others – often known as spammers -- practice deliberate deceit.

Search engine spam has been around nearly as long as web search engines. A survey of web search technology [10] cites articles in the popular press [6, 12] dating back to 1996. Early spammers tended to be "entrepreneurs, cult recruiters, egocentric Web page authors wanting attention, and technically well-versed, but unbalanced, individuals who have the same sort of warped mentality as inventors of computer viruses" [10]. Today, nearly all spam has an underlying commercial motive.

Search engine spam practices range from merely trying to boost ranking by repetition of content words, to putting names of competitors' products into a page to cause it to be retrieved for the competing product. Some pornographic sites put hundreds of non-pornographic terms on their pages, hoping to attract the attention of a user searching for something unrelated. Spammers also use a technique called cloaking, in which the content displayed to the user is deliberately different from the content made available to the search engine crawler for the same URL.

Since link-based relevance features became popular, many spammers create "link spam" – links created only for the purpose of making their pages appear popular, so search engines will rank them higher. Because some engines only count links from distinct hosts, spammers have created software robots that crawl the web looking for guest book pages, then "sign" the guest book with links back to the pages they want to promote. There are also "link swapping" sites where web site owners agree to link to each other's pages to boost their ranking.

The flip side of content and link spam are traffic bots -- programs that simulate users issuing search engine queries and clicking on results. In some cases, the bots are an SEO's way of seeing how high their pages are ranking. In other cases, the bots' creators are attempting to affect the ranking of these pages in engines that use click rates as a relevance factor.

The result of these techniques is a kind of arms race between search engine designers and spammers. Instead of spending time adding features that would improve relevance for users, many search engine engineers devote all their time to detecting and combating search engine spam.

2.2 Multiplicity of User Goals

In traditional information retrieval, the goal of the user – who was typically a student, an information professional such as a librarian, or an expert researcher – was to find information. On the web, information-finding is only one of many goals users have when they come to a search engine. In a recent study [16], we analyzed logs from the AltaVista search engine, observing not only the queries users issued but also the results found by the engine and the users' subsequent behavior (clicking on results, reformulating the query, and so on). From this analysis we identified ten different user goals underlying web searches, ranging from "getting advice" to "downloading a resource." These are shown in Table 1. These user goals fall into three broad categories: navigational, in which the user's goal is to go to a specific known web site that he or she already has in mind; informational, in which the user's goal is to learn something by reading or viewing web pages; and resource, in which the goal is to obtain a resource available on a web page. Broder suggested a similar division in an earlier analysis of web search behavior [2].

We found that over one-third of the queries had non-informational goals. Even among the informational queries, a large fraction were attempts to locate something the user wanted to buy. In other words, less than one-third of all queries reflected a traditional IR-style information need.

2.3 Corpus Characteristics

The corpora used in traditional information retrieval systems were generally fairly homogeneous and stable. Documents usually were of similar form, genre, and quality. On the web, the reverse is true. Documents have endless variety, from a kindergarten page featuring children's artwork to a PDF spec sheet for an industrial part. An increasing fraction of web pages aren't even documents in the traditional sense; they're dynamically generated content from databases.

Many traditional corpora contained a predefined (if growing) set of content – for example, all federal court decisions or all AP news wire stories from certain years – and users had a clear understanding of what documents each corpus contained. As a result, users could have a high degree of certainty that the documents they expected to find in the corpus actually existed, and to some extent, that a document that could not be found probably did not exist. But on the web, which is unbounded and subject to no centralized organization, aggregation, or filtering of content, no such inferences can be drawn. A searcher doesn't know whether the content sought doesn't exist, or simply isn't being found by the search engine. Perhaps the missing content is not on the web at all; perhaps it is on the web but on sites unknown to the search engine; perhaps it is on sites that are not crawled that deeply.

Furthermore, web content grows stale quickly, with a large fraction of pages disappearing every month. Although the rate of so-called "link rot" varies with the nature of the content, Koehler [11] reports a half-life for about two years for random web pages, while Ntoulos et al. [15] report a half-life of just nine months for pages from "popular" web sites.

SEARCH GOAL	DESCRIPTION	EXAMPLES
1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book
2. Informational	My goal is to learn something by reading or viewing web pages	
2.1 Directed	I want to learn something in particular about my topic	
2.1.1 Closed	I want to get an answer to a question that has a single, unambiguous answer.	what is a supercharger 2004 election dates
2.1.2 Open	I want to get an answer to an open-ended question, or one with unconstrained depth.	baseball death and injury why are metals shiny
2.2 Undirected	I want to learn anything/everything about my topic. A query for topic X might be interpreted as "tell me about X."	color blindness jfk jr
2.3 Advice	I want to get advice, ideas, suggestions, or instructions.	help quitting smoking walking with weights
2.4 Locate	My goal is to find out whether/where some real world service or product can be obtained	pella windows phone card
2.5 List	My goal is to get a list of plausible suggested web sites (i.e. the search result list itself), each of which might be candidates for helping me achieve some underlying, unspecified goal	travel amsterdam universities florida newspapers
3. Resource	My goal is to obtain a resource (not information) available on web pages	
3.1 Download	My goal is to download a resource that must be on my computer or other device to be useful	kazaa lite mame roms
3.2 Entertainment	My goal is to be entertained simply by viewing items available on the result page	xxx porno movie free live camera in l.a.
3.3 Interact	My goal is to interact with a resource using another program/service available on the web site I find	weather measure converter
3.4 Obtain	My goal is to obtain a resource that does not require a computer to use. I may print it out, but I can also just look at it on the screen. I'm not obtaining it to learn some information, but because I want to use the resource itself.	free jack o lantern patterns ellis island lesson plans house document no. 587

Table 1: A hierarchy of search goals (from [16]). The queries are taken from actual examples in the AltaVista query log.

Finally, the contents of many web pages changes almost daily [3]. Although many of these changes are very small [5], there are certain types of sites (news and blogs, for example), where there are substantial changes. It is common for a search engine to retrieve a URL whose actual content no longer matches the query. (This is one of the reasons why some web search engines offer cached versions of the indexed pages.)

Between the rapid addition of new pages, the death of old pages, and the changes to pages that persist, Ntoulos reports that after a year, about 50% of the content of the web is new.

3. Problems With Traditional Evaluation Measures

Keeping in mind the unique characteristics of web search described above, we can now examine specific measures that might be used to evaluate its performance.

3.1 Recall & Precision

The most common measures of effectiveness for traditional information retrieval systems are recall (the proportion of relevant documents that are retrieved) and precision (the proportion of retrieved documents that are relevant). Unfortunately, assessing relevance is a notoriously difficult business. One survey of relevance research published a few years ago [13] identified 130 papers on the topic, dating back to 1958 when relevance became an explicit concept in the field. Part of the problem is that, in a phenomenon similar to the vocabulary problem, the people assigning relevance judgments may not have the same idea of relevance as the target users. In fact, Cuadra and Katter [4] showed long ago that relevance judgments changed when the judges were given different descriptions of the intended use of the documents.

Despite these difficulties, recall and precision have been widely used in the IR research community. When these measures were originally used in the 1960s, a typical corpus consisted of a thousand or so abstracts, or perhaps a few hundred short full-text documents. Judges could then manually determine whether each document in the corpus was relevant to each of a small set of queries. By the late 1980s, the TREC conferences [8] had increased the size of the test collections to half a million documents, and manually judging every one became impossible. Instead, TREC judges assessed the relevance of a small subset of the corpus, namely, the union of all documents ranking in the top 200 of each participating system for a set of 50 test queries.

Unfortunately, this process breaks down when the corpus is several orders of magnitude larger. In fact, the sheer size of the web makes the entire concept of recall problematic. Today, the web technically contains an infinite number of pages, since some sites dynamically generate new content in response to each HTTP request. Even the "static" or "indexable" web contains on the order of five billion pages, with the number continuing to grow rapidly every year. For many typical queries, the number of *relevant* results is greater than the entire size of a typical TREC corpus.[a] At the same time, most search engine users only look at the first page of search results – typically 10 or 20 hits. So the real question from the user's perspective is not "how many of the X million relevant pages were actually retrieved by the engine?" but rather, "do the 10 results in front of me contain what I'm looking for?"

---

[a] Of course, there are other queries for which there are only a few relevant documents, and web search engines often fail to find them. In this case, recall is still a meaningful concept, but – as in TREC – it's quite possible that the sampling procedure used for relevance judgments will not even select relevant documents. One problem specific to the web is that some new sites may not be found until known sites link to them.

This latter question is close to one that we might be able to answer with the precision measure: "*How many* of the 10 results in front of me contain what I'm looking for?" In order to automate this measure, we need to have judges assess the results of the queries and decide which would have been deemed relevant. The problem, of course, is figuring out what the user was looking for.

Traditional precision tests were already problematic due to the relevance assessment problems described above, but at least the expert judges and the actual users could be assumed to be doing roughly the same information-seeking task.[b] But because web search users have so many different goals, two users issuing the same query may be looking for entirely different results. Furthermore, because of the heterogeneity of the web, the results may include a variety of genres and formats that are difficult to assess. (How relevant is a page that contains no content of its own, but has a link to a really good page? How relevant is a page that contained good content when it was indexed, but no longer does?)

If classical precision measures make too many assumptions about users to be useful, why not examine the behavior of the users themselves in assessing the system? We will now consider several measures that do just that.

3.2 Clickthrough

One intuitively appealing measure of user satisfaction is clickthrough, the rate at which users click on search results. The assumption behind the use of clickthrough as a user satisfaction measure is that if the search engine finds the web page the user is looking for, he or she will click on it.

Clickthrough has been used successfully under certain controlled conditions as a proxy for the relative value of different result sets or individual result pages. For example, Joachims [9] has interleaved results produced by different search algorithms for the same queries, and used clickthrough to evaluate their relative performance. It is plausible to conclude that if users consistently click on items returned by one algorithm (or search engine) and not another, then that algorithm (engine) produces superior results. Similarly, if one result for a given query is consistently clicked on more than another, we might conclude that the former is preferred by users to the latter, or at least has a more inviting abstract. It is natural to try to extend this approach to measuring overall user satisfaction.

Unfortunately, there are many problems with using clickthrough as a satisfaction measure. First, and most importantly, the meaning of a click depends entirely on a user's task. For example, a really successful question-answering result is one in which the answer is apparent on the search result page itself (either in a document summary or in additional content provided by the engine), so that no clicks are required. Consider the search session depicted in Figure 1. The user has entered the query **`first woman in space`**, which may be interpreted as an implicit representation of the question, "who was the first woman in space?" In this case, the abstract in the second result refers to "The first woman in space, Valentina Tereshkova, cosmonaut." Thus the information need was satisfied without the need for any result clicks.

---

[b] For example, the user intents that Cuadra and Katter described [4] involved tasks such as finding articles that "speak directly to any one of the categories in the requirement statement," finding articles that would be used for writing a review paper, getting an overview of the field, etc. While all the intents are different, they are all variations of the traditional research task.

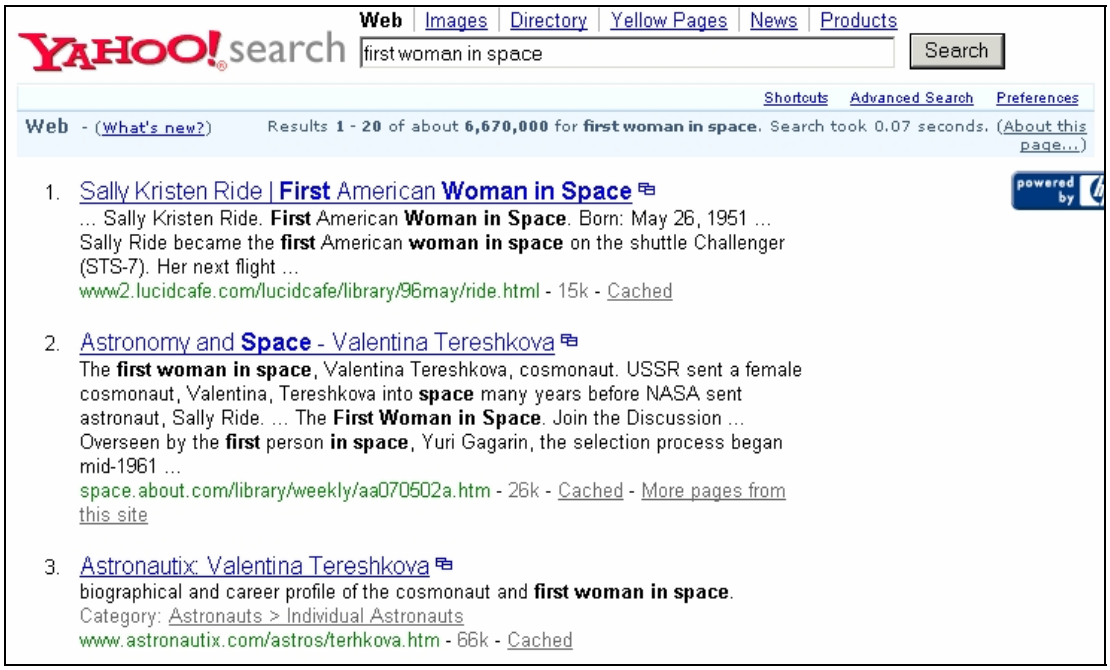

Figure 1: A search result where the title and/or abstract satisfies the information need.

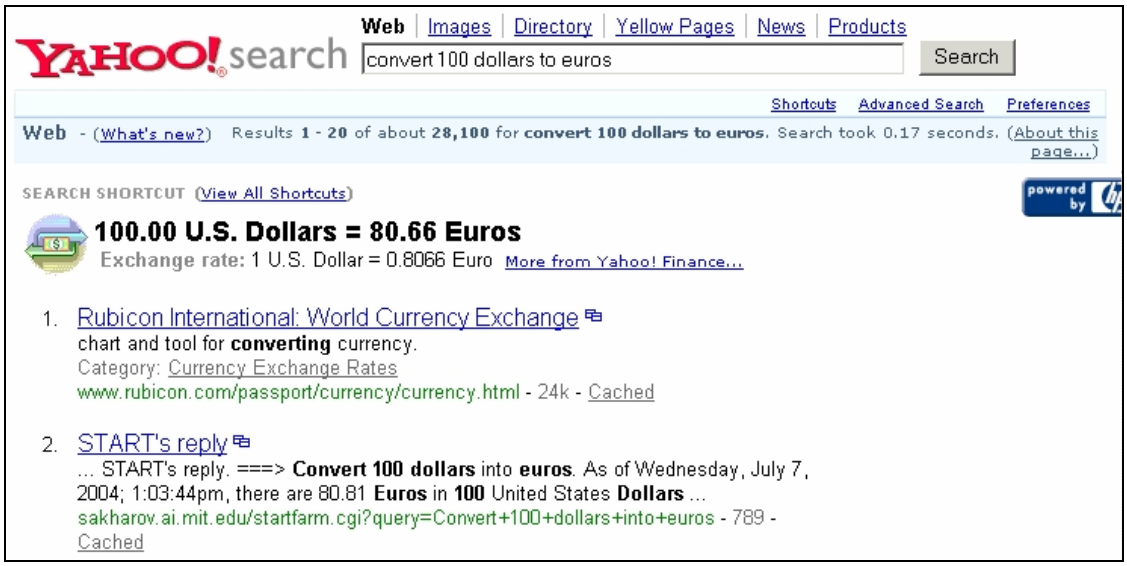

Figure 2: Another query where success is achieved with no result clicks.

Similarly, users are increasingly relying on search engines to find specific pieces of information in databases, or as a result of calculations, that would not be present in the text of an ordinary static web page. Search engines are now providing what Yahoo! calls "shortcuts" – direct answers above the web search results. For example, in Figure 2, the user's query is a request to perform currency conversion (**`convert 100 dollars to euros`**). Not only are no clicks required to satisfy the information need, in this case the user doesn't even need to look at the matching web pages.

In contrast to these "zero-click" interactions, a successful research session often involves visits to several different sites. For example, someone interested in buying a camera may want to visit a page that describes the pros and cons of different types of cameras, then a page with user reviews, then a price comparison page, and so on. A student studying a topic may want several different opinions about it. An entrepreneur may want to find out what other products of a given type are already available, and learn more about them. In all of these cases, users are likely to click many results as part of the successful completion of their task. If they don't click at all, maybe they aren't finding what they're looking for.

But even in situations where the lack of a click indicates failure, it's not always the case that the overall click rate correlates with success. Consider the navigational search task: a user wants to find a particular site she already has in mind. If she finds the site and clicks on it, the click rate goes up. But if the engine does a poor job of ranking and puts other sites first, the user may end up clicking on several different sites before finding the correct one.

Another problem with clickthrough is that users tend to be attracted by novel or provocative content. For example, there is at least some evidence from image search that users click on pornographic images at a disproportionate rate, even when those results are completely irrelevant to the original query. Engines that use clickthrough as a ranking factor may find that pornographic content rises to unwanted prominence.

Finally, clickthrough is subject to manipulation by robot traffic from spammers. Although most search engines work hard to detect and ignore traffic bots, some of these false bot clicks undoubtedly get through. Depending on their prevalence, traffic bots could make accurate measures of clickthrough very hard to determine.

3.3 Time On Task

If user click behavior is not a reliable measure of user satisfaction, perhaps we can look to the techniques of usability testing. After all, this field specializes in observing user behavior in order to learn how to improve systems. Some usability testing methods, such as think-aloud protocols [14], work well in situations where only a small number of test subjects are required (for example, finding interface design errors). However, they do not scale when faced with the more unconstrained task of assessing the overall effectiveness of the system – especially one whose users are conducting as diverse set of tasks as found in web search.

A more scalable approach might be to use another measure from usability testing, *time on task* – the amount of time a user spends performing a certain specified activity. Time-on-task measurement is commonly taught in standard usability testing textbooks (e.g. [17]) dating back long before the web was popular. In a controlled laboratory study, users can be given a set of tasks to be performed using two or more different systems. The effectiveness of the system is assumed to be inversely proportional to the time required. The systems being tested may be anything, from software to physical devices to

packaging. It is easy to imagine applying this to the web search task (e.g. "Find a web site where I can buy product X online," or "Find the address of the admissions office for university Y.") Users may be given instrumented browsers, allowing a much larger number to be studied, and allowing the data to be gathered in the user's natural work environment. In fact, this approach has been used in many studies of alternative search engine designs.

However, further analysis illustrates why time-on-task measures may not be appropriate for measuring search effectiveness. In traditional usability testing, the tasks are typically "work" and involve a clear definition of completion, and the interfaces are typically controls of one sort or another. For example, designers of an airline reservation system might want to see how long it takes a user to book a flight using a particular version of the interface. But in web search, the tasks may be viewed as "play" (entertainment or diversion), or may be open-ended. Furthermore, the "interface" with which the user is interacting includes not just the controls but the content itself – a wealth of potentially interesting information, with controls (links) of its own. If a study finds that users perform a certain task faster on search engine A than on search engine B, it does not necessarily mean that A is more effective or produces greater user satisfaction. In fact, users on system B may be encouraged to explore, iterate, and gradually refine their search – behavior consistent with our understanding of how human information-seeking works [1]. As they learn more about their topic, they may be end up with a better, more thorough answer. Or, these users may simply be enjoying the content they're looking at.

3.4 Surveys and User Feedback

If we can't automatically detect user satisfaction, what about simply asking users to give their own assessment of the experience?

One way to do this is to survey users who are already using the system. For example, suppose we wanted to assess the quality of web search results as judged by users. We might create a feedback mechanism embedded in the search interface, as in the example shown in Figure 3. Here, each search result is followed by a simple form that lets users rate the quality of that result for that query. Alternatively, the survey can be part of a separate interface, such as a pop-up window. Each of these is problematic. In both cases, the survey should not be shown too frequently, or it will annoy users and interfere with their tasks. The embedded survey has a particular design challenge: it must not prevent users from using their normal search results, but it can't be so subtle as to go unnoticed. Furthermore, it appears that this type of survey leads to skewed results, since only users who have strong opinions (particularly dissatisfaction) are likely to reply. A pop-up survey, though more intrusive, is harder to overlook – yet many users have learned to close pop-up windows without even thinking about it, and a growing number have browser plugins (usually toolbars) that automatically block pop-ups. In either case, the number of usable responses is likely to be quite low.

One problem might be that users have no incentive to spend time evaluating the system or filling out a survey. After all, this effort does not get them closer to satisfying their information-seeking goal. Suppose users were given an incentive to treat the survey as their primary task? This is an alternative approach to explicitly gathering user input. For example, one way to assess the quality of web search results is to pay users to judge them. There are many variations of this method – for example, we might compare two search engines by having users indicate which of two search result lists were preferred for a given query. (The results would be shown using a neutral interface, with branding and style information that might bias the user removed.)

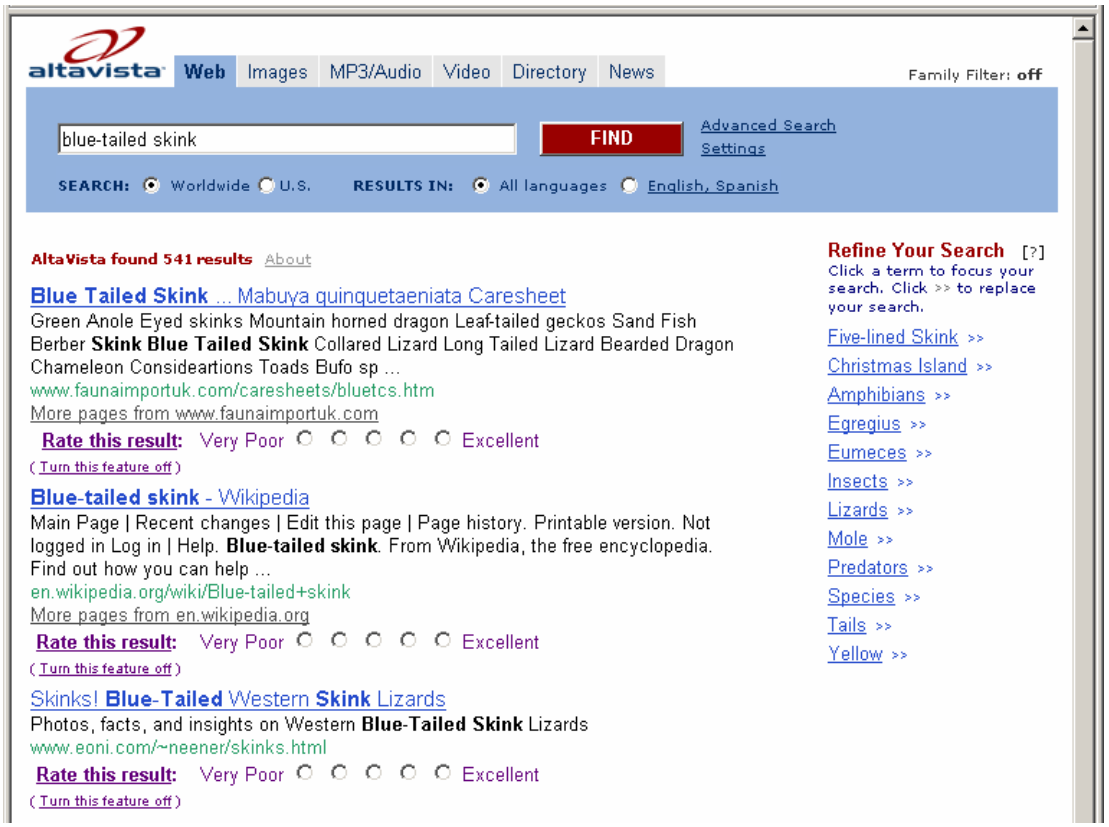

Figure 3: The AltaVista search result page showing an inline user rating form.

However, this approach has other problems. While users now have a motivation to work on the survey, they no longer have a motivation to give meaningful responses. In one study conducted by AltaVista, 55% of survey results had to be discarded because of bad data. For example, some users simply checked the same rating for every item or responded too fast to have read the choices. Others failed a simple effort test in which they were required to distinguish actual search results from random URLs. When the external financial incentive to complete surveys replaces the internal incentive of satisfying an information need, user feedback becomes relatively meaningless.

4. Conclusion

The characteristics of web search make evaluating its performance extremely difficult. Recall is ill-defined when there is an infinite supply of content; precision is difficult to measure without knowing the user's task context. Yet more user-oriented measures such as clickthrough, time on task, and self-reported satisfaction introduce problems of their own. Finding better evaluation methods – especially those that take user needs into account – remains a challenge for the research community. Until then, we will continue to rely on combinations of existing measures, however flawed, and hope that agreement between them is a reliable indicator of web search effectiveness.

Acknowledgements

The author wishes to thank Michelle Fulcher and Doug Young for their contributions to some of the ideas discussed in this paper, and Susan Gruber for her invaluable comments on earlier drafts.

References

1. Bates, M.J. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review* 13, October 1989, 407-424.
2. Broder, A. A Taxonomy of Web Search. *SIGIR Forum* 36(2), 2002.
3. Cho J. and Garcia-Molina H. The Evolution of the Web and Implications for an Incremental Crawler. *Proceedings of the 26th International Conference on Very Large Databases*, Sep. 2000.
4. Cuadra, C.A. and Katter, R.V. Opening the Black Box of 'Relevance.' *Journal of Documentation* 23:291-303, 1967.
5. Fetterly, D., Manasse, M., Najork, M., and Wiener, J.L. A Large-Scale Study of the Evolution of Web Pages. *Proceedings of WWW 2003*.
6. Flynn, L. Desperately Seeking Surfers: Web Programmers Try to Alter Search Engines' Results. *New York Times*, November 11, 1996, p. C5.
7. Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, 30(11):964-971, November 1987.
8. Harmon, D. Overview of the First Text REtrieval Conference. *Proceedings of the First Text REtrieval Conference (TREC-1)*, 1992.
9. Joachims, T. Optimizing Search Engines Using Clickthrough Data. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD),* 2002.
10. Kobayashi, M. and Takeda, K. Information Retrieval on the Web. *ACM Computing Surveys,* 32(2), June 2000.
11. Koehler, W. A Longitudinal Study of Web Pages Continued: A Report After Six Years. *Information Research*, 9(2) paper 174.
12. Liberatore, K. Getting to the Source: Is it Real or Spam, Ma'am? *MacWorld*, May 1997.
13. Mizzaro, S. Relevance, the Whole History. *Journal of the American Society for Information Science*, 48(9):810-832, 1997.
14. Nielsen, J. *Usability Engineering*. Morgan Kaufmann, 1994.
15. Ntoulos, A., Cho, J., and Olston, C. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. *Proceedings of WWW 2004*.
16. Rose, D.E. and Levinson, D. Understanding User Goals in Web Search. *Proceedings of WWW 2004*.
17. Rubin, J. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests.* Wiley, 1994.