INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the
text directly from the original or copy submitted. Thus, some thesis and
dissertation copies are in typewriter face, while others may be from any type of
computer printer.

The quality of this reproduction is dependent upon the quality of the copy
submitted. Broken or indistinct print, colored or poor quality illustrations and
photographs, print bleedthrough, substandard margins, and improper alignment
can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and
there are missing pages, these will be noted. Also, if unauthorized copyright
material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning
the original, beginning at the upper left-hand corner and continuing from left to
right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced
xerographically in this copy. Higher quality 6" x 9" black and white photographic
prints are available for any photographs or illustrations appearing in this copy for
an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zebed Road, Ann Arbor, MI 48106-1346 USA

UMI
800-521-0600
RICE UNIVERSITY

The Psychological Fidelity of Web Search Engines

by

Karin Quiñones Dudziak

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
MASTER OF ARTS

APPROVED, THESIS COMMITTEE:

Daniel N. Osherson, Chair
Autry Professor of Psychology

Kenneth R. Laughery
Henry R. Luce Professor of Engineering Psychology

Robert L. Dipboye
Professor of Psychology and Management
Chair of the Psychology Department

Houston, Texas

December, 1999
ABSTRACT

The Psychological Fidelity of Web Search Engines

by

Karin Quiñones Dudziak

Web search engines are the primary method for finding information in the vast World Wide Web, yet little empirical research of them has been conducted. This study examined how well search engines’ relevancy ratings of Web sites matched users’ ratings. Eighty-one people rated Web sites in three different topic areas that were delivered as top sites by Infoseek and/or HotBot. To ensure fair ratings, participants were unaware of the ratings assigned by the search engines, or even that search engines had generated the sites. Individual difference characteristics of the raters (e.g., browser-searcher, level of knowledge and interest in topic) were also collected. Agreement as to the top sites was quite low among search engines. However, people showed very high levels of agreement and reliability. Search engines’ ratings did not correspond well with the raters’ ratings. Nevertheless, Infoseek performed better than HotBot on every measure of agreement with user ratings.
Acknowledgments

I want to thank my incredible advisor, Dan Osherson, for seeing the value in this area of research and for his ideas and support from conception to completion. The members of my committee (Daniel Osherson, Ken Laugherly, and Bob Dipboye) have not only provided valuable input on this project, but throughout my graduate career. Many thanks to all of them.

I also want to thank my parents for setting such a wonderful example and always believing in me. I am honored that they are my friends as well as my parents. My entire family has been so supportive and has believed in me from day one, thank you! In addition, I want to thank Denise and Melissa for getting me through all the rough spots and reminding me to enjoy life.

I especially want to thank my husband, Mickey, for all his advice, enthusiasm, support, love, and Ben & Jerry's runs! It has been one incredible year.

In loving memory of Shelli and Louis XIV
# Table of Contents

Introduction .................................................................................................................. 1
Method .......................................................................................................................... 15
Results ......................................................................................................................... 20
Discussion ................................................................................................................... 34
Conclusion ................................................................................................................... 41
References ................................................................................................................. 43
Appendices .................................................................................................................. 46
List of Tables

Table 1. Inter-Participant Agreement and Reliability Values .................. 22
Table 2. Participant-Search Engine Agreement Measures ..................... 24
Table 3. Superior Engine Values .................................................. 27
Table 4. Results of Individual Difference ANOVAs .......................... 33
List of Figures

Figure 1. Search Engine by Topic Interaction............................................. 30
The Psychological Fidelity of Web Search Engines

Use of the World Wide Web has exploded over the past few years. In 1996, there were fewer than one million Web sites (Scientific American, 1997). Today, there are hundreds of millions of Web sites and the Web is growing at approximately one million pages per day (Scientific American, 1999). If all the Web sites in existence were viewed for five seconds each, it would take a person over four months of 24 hours a day, seven days a week surfing to see all the sites available on the Web (Shirky, 1999).

One great advantage of the Web is that it can support a multitude of mediums. Web sites can contain journal articles, entire books, poems, or any other pure text. What is fascinating, however, is that Web sites can also contain pictures, images, audio clips, video clips, transcripts of online conversations, advertisements, or raw data. People can buy things, make reservations, and even create personalized newspapers over the Web. This ability to display so many different types of information is partly accountable for the phenomenal growth of the Web.

Allowing access to so many different types of information has helped the Web appeal to a very broad audience. The Web is not limited to those in academia or big business; it is available to anyone with a computer and a modem. One American household in three has Web access and the numbers are increasing by about 25% per year. It is estimated that fifty percent of working Americans will be online by the end of 1999 and half of all Americans will be online by the end of the year 2000 (Scientific American, 1999). The number of
people who visited the Web at least once per month grew to approximately 62 million in May 1999 (NewsEdge Corporation, 1999).

People are viewing an increasing number of Web pages and spending more time on the Internet. The average number of distinct Web pages viewed per visitor per month increased nearly 50% from May 1998 to May 1999. The estimated total number of Web pages viewed in May 1999 exceeded 28 billion, 70% more than in May 1998. People spent 20% more days and 40% more time on the Internet in May 1999 than they did in May 1998. In May 1999, over 1.2 billion hours were spent online (NewsEdge Corporation, 1999).

The dramatic increase in the amount of information contained in these hundreds of millions of Web sites has created a situation of information overload for the users. With so many Web sites coming online each day, it is impossible to keep track of everything out there. The Web is sometimes referred to as the library for the digital age. Unfortunately, the variety of mediums and the explosion of individual sites do not allow it to be organized like a library. Not even a team of human indexers can keep up with the growth rate of the Web (Lynch, 1997). The Web has been referred to as "a global mess of previously unimagined proportions" (Scientific American, 1999). In an effort to allow users to retrieve information they desire, Web search engines have been developed.

Search engines are Web sites that allow users to access information on the Web by entering key words into a search field. Terms can be as simple as a single word (e.g., "hanggliding") or a set of words with Boolean operators (e.g., "hanggliding & equipment"). After conducting the search, engines typically
present the user with a list of sites which are deemed relevant to the search terms entered by the user. Most search engines rank these sites from most relevant to least relevant and some provide a numerical relevancy rating for each site.

Search engines have become the main “on-ramp” into the information superhighway. In fact, search engines represent 6 out of the 10 most commonly visited sites on the Web (NewsEdge Corporation, 1999). Accordingly, stocks of search engines, such as Yahoo! and Lycos, have become hot commodities on Wall Street.

An issue that emerges, which researchers have only just begun to address, is the extent to which these search engines are finding the information that users are looking for. Each search engine uses a unique algorithm to determine the relevancy of a Web site to a particular search query. As a consequence, the same search produces potentially different results depending on the particular search engine used (e.g., Ding & Marchionini, 1996; Venditto, 1996). An important issue concerns the extent to which the search engines use algorithms to judge relevance that match the criteria that users employ to judge relevance. In other words, what is the “psychological fidelity” of Web search engines?

The present study represents one of the first systematic efforts to address this issue. Its purpose is to determine the extent to which rankings and relevancy ratings generated by various search engines match each other and, more importantly, match the ratings of actual end-users.
This paper is organized as follows. First, a brief overview of search engines and their workings is presented. Second, a review of past research in the area of search engines will be presented. Third, the specific methodology used to address the research question will be described. Finally, the results of an experimental study will be discussed.

Search Engines

There are many ways a person can access a Web page. They can get the address of the site from a TV ad, a friend, a mailing, or from some other source. Web pages often have links, clickable words or icons that take the user to another site, which lead users to a Web site of similar or related content. However, this method of finding information on the Web can be quite inefficient.

Another method for finding information on a particular subject is through the use of portal sites, like Yahoo!, which offer high quality selective resources and browsable hierarchical subject menus. They are like catalogs or indices with subjects of common interest such as sports, news, and computers. However, if a user is looking for specific information these portals may overlook essential information.

Recently, Web search engines have been developed that actively search existing Web sites for a specific topic of interest to the user. Web search engines were first developed in 1994 (Chu & Rosenthal, 1996). A search engine is a service on the Web, usually provided free of charge, developed to provide fast and easy access to the enormous body of information on the Internet. Search
engines use three components: a robot (spider), a database, and an agent. Robots are automated software agents, which are usually referred to as spiders. These spiders crawl from Web site to Web site and record the addresses (URL's) of the sites in existence. Spiders extract some information about the site and send it back to the database. Each search engine uses a slightly different algorithm to catalog the Web sites, but they all use some common methods. The spider can extract the entire text of the document or just some key elements of the site. For example, a search engine may record the title and key words of the document, the size of the site, just headers, or any combination of elements. The number of links to that particular site may be recorded to determine the quality of the site, or its popularity (Venditto, 1996, Lynch, 1997, Su, 1997 and Glossbrenner & Glossbrenner, 1999).

All of the information collected by the spider is then stored in a database. When a user performs a search, the search engine's agent is used to extract key components from this huge database of information. The user goes to a Web Search engine site and types in the topic of interest, the query. A user may use Boolean symbols to make the search more precise, but 80% of searchers do not use these symbols (Search Engine Watch, 1999). Once the user has entered the topic into the search engine's interface, the agent determines which will be the most relevant sites about that topic. The logic of the search engine's agent is confidential, but they often use criteria such as the number of times the keyword appears in various sites, where in the document the word appears (title, header, footnote), and how popular the site is (how many other sites link to it). The
search engine then returns a list of Web sites, often ten sites per page, in rank order from most relevant to least relevant. For each Web site it recommends, the search engine provides a link to the site’s address and the title of the page. In addition, some search engines provide a brief summary of the contents and the size of the site. Several engines give a relevancy rating (e.g. 90%) in addition to rank ordering the sites. Often the search engine will produce a list of thousands of Web sites that may (or may not) be relevant to the user’s search.

This method of finding information on the Web is a popular method for users, in fact search engines service over a million queries per day (Lynch, 1997). Ideally, a search engine should allow a user to type in a question in plain English and return not only what the user asked for, but what s/he meant to ask for. This process should not require any knowledge of complex Boolean logic, nor should the user be required to perform several different queries to get the desired information (Feldman, 1996). Unfortunately, search engines are far from that ideal today.

Automated search engines do not work as well as human indexers at picking up the identifying characteristics of a document or the overall theme of the site. Simply counting the number of times a word appears in a site, even at the beginning of the document, may not be a very accurate measure of content. This example from Scientific American (1999) “Hypersearching the Web” makes the point very well. A search engine may determine that a Web page with an online version of Tom Wolfe’s book The Kandy-Kolored Tangerine-Flake Streamline Baby ranks very highly for searches concerning hernias. A human
indexer would realize that just because the book begins with the word hernia repeated dozens of times does not mean that the site is about hernias. Unfortunately, a search engine may not be able to tell the difference.

Web site designers will sometimes use this glitch in search engine technology to their advantage. They hope to attract popular search queries by putting those key words in invisible text hundreds of times in a document. This technique is called spamming. Search engines are getting more sophisticated at detecting spamming, but it is still used by designers because not all engines can detect it (Feldman, 1996, Lynch, 1997, Scientific American 1999, and Venditto, 1996).

Search engines also cannot identify the non-text information contained on the page unless it is labeled very clearly. Graphics and audio clips may be the key elements of a particular site, but a search engine has trouble coding that information. It is also difficult for search engines to determine what kind of information the site contains. Sometimes a search engine will produce everything from an advertisement to a play to a transcript of a meeting as sources relevant to a search query. Most search engines can only recognize text, but the ability to display other types of medium is one of the key advantages of the Web. (Feldman, 1996, Lynch, 1997, Scientific American 1999, and Venditto, 1996).

The result of these inadequacies in search engine technology can cause users a headache and waste valuable time. Often a simple search query will generate thousands of sites. Totally irrelevant sites can end up in the top of the
list while relevant sites may be buried in the bottom or even omitted completely. Are search engines producing the results users desire?

Past Research on Search Engines

Some researchers have begun to examine the adequacy of search engines. Early research was primarily descriptive in nature, discussing the differences among the various search engines available. While most users choose the search engine that they are most familiar with or the one that is easiest to access, the engines are inherently different so one may work better for a particular query. For example, Courtois, Baer, and Stark (1995) found that WebCrawler had the easiest interface to use, so it would be the best search engine for novice searchers. Notess (1995a) found that Lycos was best for single word searches, but for multiword searches including Boolean operators, WebCrawler was recommended (Notess, 1995b). In 1996, Alta Vista was shown to have the largest Web-search database so it could provide the most comprehensive results, but Excite offered the fullest range of services (Venditto, 1996). Scoville (1996) determined that Excite, Infoseek, and Lycos provided the most accurate results and easy to use interfaces. Amazingly, in a study of seven Web search engines, Internet World labs found that no two search engines yielded the same set of results on a search during their entire testing period (Venditto, 1996).

Several other researchers discovered the same phenomenon. When search engines were given the same input as a search query, the results varied dramatically (e.g., Chu & Rosenthal, 1996 & Ding & Marchionini, 1996). There
was often very little overlap at all between results of different search engines. So which engine is the best?

Researchers began to examine the output of the search engines and rate their quality. The studies reporting differences among engines do not agree about which is the best. For example, Leonard (1996) felt that out of the seven search engines he studied AltaVista was the best, while Kimmel (1996, as cited in Chu & Rosenthal, 1996) felt that Lycos was superior. Venditto (1996) determined that Infoseek provided the most relevant results. With such varying results, the criteria and methodology of these studies must be examined.

In all the aforementioned studies the author(s) of the articles rated the engines based on their personal expertise. Chu and Rosenthal (1996), for example, compared the search engines' search capabilities that related to Boolean logic and word/phrase searching. They also evaluated the response times of the servers. As other studies showed before them, the response times of the search engines were similar. In addition, Chu and Rosenthal (1996) examined the precision of results.

Precision is a common measure used to evaluate search engines, but the exact definition employed varies among studies. Roughly, it is a measure of the actual relevance to users of the top sites claimed to be relevant by the search engine. Intuitively, an engine is precise to the extent that it "tells the truth" about the sites it claims to be relevant to the search. Imprecise search engines falsely claim that a site will be relevant to the user.
The way in which Chu and Rosenthal (1996) performed the search differed depending on which engine they were testing. For example, to find information on the psychological analysis of contemporary British artist Francis Bacon, the searches were composed in two different ways. When searching in AltaVista the query was “psychological analysis” + “British artist” + “Francis Bacon.” In Excite and Lycos the same question was asked with this query: “British artist Francis Bacon”. For each of their ten search questions, Chu and Rosenthal (1996) created a unique search for each engine based on the capabilities of that engine. They then directly compared the results of the searches.

The results were evaluated by each author in terms of precision. Chu and Rosenthal’s (1996) precision measure was calculated by determining the percentage of the top ten sites that were relevant to the search (according to the authors). Precision scores were calculated for each individual query and for each engine across all ten queries. They only evaluated the top ten results of each query in each search engine. Citing time limitations, they based these judgements on the summary output of the search engine instead of following the links to the actual Web sites.

Using these evaluations, Chu and Rosenthal (1996) found that AltaVista produced the most precise results and Excite produced the least precise results among the three search engines. They also noted that among the 250 Web records that they viewed, there were very few duplicates. They inferred that the
spiders of the search engines were actually gathering sites from very different portions of the enormous World Wide Web.

While Chu and Rosenthal (1996) found one search engine to be significantly better, Ding and Marchionini (1996) found that none of the search engines they evaluated provided good output. Ding and Marchionini (1996) evaluated Infoseek, Lycos, and OpenText. They used a similar approach to Chu and Rosenthal (1996) by using specific query syntax for each particular engine and evaluating the top 20 sites each engine retrieved. Specifically, they measured the precision of the output (the number of the top twenty sites that were rated in the upper half of their 6 point scale), the amount of duplication within one engine (when an engine retrieves the same Web site more than once), the degree of overlap between engines (when two separate engines retrieve the same Web site), and they validated that the links were working (Ding and Marchionini, 1996).

Because the formats of the summary reports the search engines produced were so different, Ding and Marchionini (1996) actually viewed the Web sites themselves. They found no significant differences in precision scores between engines, but all of the engines were less than 55% precise according to their criteria. Invalid links were found in each search engine, but not necessarily in every query. While they did find overlapping sites between engines, they were surprised by the lack of overlap. They especially noticed that there were very few overlapping good Web sites.
But who determines what is a “good” Web site for a search? In all previous studies the authors of the study determined which sites were relevant and which were not. But do the authors of these studies have the same criteria as other users, especially more novice users? Su (1997) proposed that a broad spectrum of users should be asked to evaluate the results of searches and that the users’ characteristics (e.g., educational level and subject expertise) should be collected. Participants should vary in their information need and level of expertise. She also proposed some standard evaluations of search engines such as precision scores. One pilot study has been performed using some of her proposed methodology.

Su, Chen, and Dong (1998) followed the suggestions in Su (1997) with eleven end-users. The participants searched separately for their individual topics on four different search engines. After performing as many searches and as many search strategies as they needed to get the best results on each engine, the results were printed out. The next day the participants made relevancy judgements on the Web sites produced by the search engines. These relevancy judgements were based on the summary output from the search engine and were rated on a three-point scale. Su et al. (1998) also collected demographic characteristics of their participants. Eight of the eleven participants were either faculty or graduate students in information science or library science. The searches they performed were of their own choosing and ranged from simple one word searches (e.g., “literacy”) to more complicated searches (e.g., “the future of
technical services in libraries"). Were these searches typical of the searches performed on Web search engines?

To see what a “typical” search consisted of, 575 million real searches performed over 43 days were cataloged and analyzed by Compaq Systems Research Center. The results of this analysis showed that 80% of search queries make no use of search operators or Boolean symbols. In fact, 65.2% of queries consisted of only one or two words. Most users (77.6%) did not modify their searches and only performed one search per visit to a search engine. Out of the 575 million searches there were 154 million unique search queries, indicating that one out of four searches were for something that no one else had searched for (in the exact same way) over the six week period (Search Engine Watch, 1999).

The foregoing results highlight the atypical character of the searches performed in existing studies (e.g., Chu and Rosenthal, 1996 and Su, Chen, and Dong, 1998). Previous searches were often open ended and complicated, hence different than typical searches as noted above. Users in previous studies were allowed to modify their searches (as in Su et al., 1998) and different search strategies were used for particular engines (as in Chu and Rosenthal, 1996 and Ding and Marchionini, 1996). Neither of these techniques are representative of typical user behavior.

Another limitation to all the previous studies, except for Su et al. (1998), is that the relevancy of search results was rated by the authors, not by typical users. All of the previous studies, except for Ding and Marchionini (1996), rated
relevancy on the summary output of the search engine, not on the actual Web sites the engine recommended. Users are interested in the actual Web sites the search engine leads them to, so that end product is what should be judged. Also, in all previous studies the evaluators knew which search engine delivered the site being judged. This could potentially influence their ratings of the sites.

The current study builds on this line of research by taking into account the limitations of previous studies. A set of representative users evaluated the Web sites generated by a typical query into a search engine. Typical queries are short one to two word searches and make no use of Boolean logic. Consistent queries between search engines allowed fairer comparisons of the outputs. The actual Web site the search engine recommends, not just the summary produced by the search engine, was employed by the users to make their judgements. Finally, no information regarding which search engine produced the Web site, nor what ratings the search engine gave the site, were made available to the users making evaluations of the sites.

This purpose of this study is to address the following research questions:

- Question 1: Will the two search engines agree on the top Web sites for a particular search topic? Previous studies have shown low agreement among search engines.
- Question 2: Will participants (users) be in agreement as to the relevancy ratings assigned to Web sites?
- Question 3: Will participants be in agreement with the relevancy ratings assigned by search engines?
• Question 4: Will one search engine have higher levels of agreement with participants than the other?

• Question 5: Will the relationship between participants’ relevancy ratings and the search engine’s relevancy ratings be moderated by factors such as: interest in topic area, knowledge level in topic area, average amount of time spent on the Web, and browser vs. searcher classification?

• Question 6: Will the search topic affect the relationship between the participants’ relevancy ratings and the search engines’ relevancy ratings?

• Question 7: Will one engine show higher levels of precision than the other engine, when precision is defined as the percentage of the search engine’s top ten sites rated as a top ten site by the participants?

Method

Design

Three different search queries were created. The queries were designed to cover three different topic areas of potential interest to young adults and to conform to the criteria for a typical query as determined by Search Engine Watch, 1999. The queries used were “hanggliding,” “garlic health” (for the health benefits of garlic), and “acne.” The queries were entered into the two different search engines (HotBot and Infoseek) on the same day at approximately the same time. The first 20 unique sites delivered by each engine were recorded, and the sites’ rank order and numerical relevancy rating that each engine provided were recorded as well.
The union of the top 17 valid, unique sites from each of the two engines was determined. There was a maximum of 34 (17+17) sites in the union. This number of Web sites was determined through pilot studies to be a manageable amount of data for participants. In two of the search topics, the union resulted in only 31 sites to be judged by participants because there were three overlapping sites between the engines. In the end, three separate unions were formed, one for each topic. The titles of the Web sites in each union were then compiled into a Web page containing the titles of the Web sites, but excluding any information regarding which of the two engines produced the resulting site and what ratings the engine had calculated for the site. There were three Web pages created, one for the Hanggliding results, one for the Garlic Health results, and one for the Acne results.

This procedure can be illustrated with a toy example involving just a few Web sites. Imagine that the search “garlic health” was entered into HotBot and Infoseek. HotBot delivered Web sites A, B, C, and D while Infoseek delivered Web sites B, D, E, F. The relevancy ratings assigned to the sites by the search engines were recorded and set aside. A new Web page was designed that had the titles of Web sites A, B, C, D, E, and F (the union of the two search results) listed in random order. This new page of six Web site titles would be viewed by participants evaluating the relevancy of Web sites to a search on the health benefits of garlic. In the experiment actually performed, the top 17 rather than four sites were collected from each search engine, and the union constructed from these two larger sets yielded 31-34 sites in all.
Participants

Eighty-one students at Rice University participated in this study. Participants ranged in age from 17 to 37, with a mean age of 19.6 years. Approximately two-thirds of the participants were female, one-third were male. Participants spent an average of 9.3 hours per week using the World Wide Web and reported feeling very comfortable with Web search engines. The participants represented the full range in interest and knowledge about the three topic areas. Twenty-six participants evaluated Web sites relating to hanggliding, 28 participants evaluated sites related to the health benefits of garlic, and 27 participants evaluated sites about acne. No participant evaluated more than one topic area.

Equipment and Materials

Basic demographics were collected about the participants. Items included age, gender, year in school, hours per day and per week spent on the Web, comfort level with Web search engines, preferred search engine(s) (if any), knowledge level of the topic area, interest level in the topic area, and a measure of browser versus searcher behavior. See Appendix A for a sample demographic sheet.

The written instructions for the experiment were on a Web page displayed on the computers in the experiment room and were also repeated verbally to the participants before the experiment began. See Appendix B for the instructions. After verbal instructions were given about the experiment, information about their
individual topic areas was presented on the computer screen. See appendix C for a sample topic area explanation.

Internet Explorer 4.0 was used as the Web browser on computers with a minimum of 300Mhz processing speed.

Two search engines were tested, Infoseek and HotBot. Both search engines are rated in the top five by experts in the field, provide a list of Web sites in rank order of relevancy, and also provide a numerical relevancy rating on a percentage scale (0-100%).

As indicated above, the search queries for the study were “hanggliding”, “garlic health”, and “acne”. These search terms fit the profile of typical searches in that they are short and require no Boolean operators. They are also subject matters in which undergraduates have varying interest and knowledge levels.

Each participant was presented with a Web page of links related to the topic area, called their index page. These links were determined by the procedure described in the design section of this paper. The titles of the sites in the union for a topic were arranged in one of eight random orders on a Web page. The title of a Web page is written into the code of the page itself, so it is consistent regardless of which engine(s) retrieve it. On the page that the participant viewed, the title of each Web site was also a link to that site. Each participant in effect saw a Web page full of links to other Web pages about their topic area. No information about the search engines or any ratings were provided. See appendix D for a sample Hanggliding index page, appendix E for
a sample Garlic Health index page, and appendix F for a sample Acne index page.

Participants were given a sheet of paper to take notes on each site. To make it easier for the participants to evaluate each site, the sheet contained the Web site titles in the order that they appeared on their computer screen. They were also given a rating sheet that contained all the Web site titles they had evaluated, in order, with a place to write their relevancy rating for each site.

**Procedure**

Upon arrival at the experiment site, participants signed a consent form and read the instructions on the computer screen (Appendix B). They also filled out the demographic form (Appendix A). After all participants had read the instructions and filled out the demographic form, verbal instructions were given.

Participants were told that they would be evaluating the Web sites on their relevancy to a search topic. Participants then followed a link on their screen that randomly assigned them to one of the three topic areas. After reading the instructions about their topic area (Appendix C), they followed another link to their experiment page. This Web page, called the topic index page (Appendices D-F), consisted of a list of Web site titles about their topic area. The titles were in one of eight random orders, and each Web site title was a link to that Web site.

Participants visited every site in the list on their page, in any order they chose. They were encouraged to take notes on each site on the notes page provided (refer to appendix G for a sample notes page). These notes were to aid them in making evaluations of the sites and were strictly for their own use. They
were also allowed to visit each site as many times as they desired, but no official ratings were to be made during this stage. Only after visiting every site on their list, which was verified by the change in color of a visited link, were participants given a rating sheet. Participants then assigned each Web site a relevancy rating, from 0 to 100, on the rating sheet (see appendix H for a sample rating sheet). They were allowed to return to the Web sites as many times as they wanted to make these ratings. When the participants completed their evaluations of all of the Web sites, they were debriefed and dismissed.

Results

Engine-Engine Agreement

The number of Web sites that were identified by both engines in their top 17 sites was counted to measure the amount of agreement between search engines. For the Garlic Health search, no overlapping sites were found. For the Hanggliding and Acne queries, only three sites overlapped. These findings support the discoveries of previous research and address Question 1 regarding the agreement of search engines. The search engines were found to have low agreement as to the top Web sites relevant to a search query.

Inter-participant Agreement and Reliability

Three measures of inter-participant agreement and reliability were calculated. Kappa was calculated as a measure of overall agreement between participants in their relevancy ratings of the Web sites. The mean inter-participant Pearson correlation and the intraclass correlation were calculated as measures of inter-rater reliability. The intraclass correlation adjusts for the
number of raters making the judgements. All three measures found significant
\( (p < .001) \) agreement and reliability of participants' ratings. See Table 1 for a
summary of all of these values.

Overall (across all three topic areas), Kappa = 0.44, mean \( r = 0.64 \), and
Intraclass \( r = 0.97 \), all significant at \( p < .001 \). For Hanggliding, Kappa = 0.36,
mean \( r = 0.62 \), and Intraclass \( r = 0.97 \). For the Garlic Health search, Kappa =
0.55, mean \( r = 0.66 \), and Intraclass \( r = 0.98 \). For Acne, Kappa = 0.42, mean \( r =
0.64 \), and Intraclass \( r = 0.97 \). All three measures in all three topic areas were
significant at \( p < .001 \).

These finding provide an affirmative response to Question 2 that asked if
the participants (users) would have a high level of agreement with each other as
to the relevancy ratings of Web sites. Participants showed a high level of
agreement as to the actual relevancy ratings assigned to sites and were
consistent in their relative ratings of sites.
Table 1

Inter-Participant Agreement and Reliability Values

<table>
<thead>
<tr>
<th>Topic</th>
<th>Kappa</th>
<th>Z (Kappa)</th>
<th>p-value</th>
<th>Mean r (SD)</th>
<th>Intraclass r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hanggliding</td>
<td>0.36</td>
<td>13.80</td>
<td>0.0000</td>
<td>0.62 (.15)</td>
<td>0.97*</td>
</tr>
<tr>
<td>Garlic Health</td>
<td>0.55</td>
<td>19.06</td>
<td>0.0000</td>
<td>0.66 (.14)</td>
<td>0.98*</td>
</tr>
<tr>
<td>Acne</td>
<td>0.42</td>
<td>15.52</td>
<td>0.0000</td>
<td>0.64 (.13)</td>
<td>0.97*</td>
</tr>
<tr>
<td>Overall</td>
<td>0.44</td>
<td></td>
<td></td>
<td>0.64 (.14)</td>
<td>0.97*</td>
</tr>
</tbody>
</table>

* p < .01
Participant-Engine Agreement

Before conducting any tests of the relationship between search engines' ratings and participants' ratings, the data was plotted and checked for homoscedastic variance. It was found that the assumption of homoscedastic variance was violated. Therefore, no correlations were calculated between the engines' ratings and the participants' ratings because they would not be accurate measures of association. Two other measures of participant-engine agreement were calculated instead.

First, a median split was performed on each participant's relevancy ratings to categorize each Web site as a top site (high) or a bottom site (low). Similarly, if a search engine delivered a particular Web site, that site was classified as a top site (high) according to the engine. If the search engine failed to deliver the site in the top 17, the site was classified as a low rated site according to the engine. A phi coefficient was then calculated as a measure of agreement between the engines' top sites and the participants' top sites. Overall, the average phi = 0.20 with Infoseek and -0.28 with HotBot. For Hanggliding, the average phi = 0.05 with Infoseek and -0.12 with HotBot. For Garlic Health, the average phi = 0.40 with Infoseek and -0.40 with HotBot. For Acne, the average phi = 0.14 with Infoseek and -0.31 with HotBot. See Table 2 for a summary of these calculations.
Table 2

Participant-Search Engine Agreement Measures

<table>
<thead>
<tr>
<th>Topic</th>
<th>Phi (SD)</th>
<th>Precision Score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HotBot</td>
<td>Infoseek</td>
</tr>
<tr>
<td>Hanggliding</td>
<td>-0.12 (.17)</td>
<td>0.05 (.17)</td>
</tr>
<tr>
<td>Garlic Health</td>
<td>-0.40 (.11)</td>
<td>0.40 (.11)</td>
</tr>
<tr>
<td>Acne</td>
<td>-0.31 (.20)</td>
<td>0.14 (.21)</td>
</tr>
<tr>
<td>Overall</td>
<td>-0.28 (.16)</td>
<td>0.20 (.16)</td>
</tr>
</tbody>
</table>
As an additional measure of search engine and participant agreement, a precision score was calculated. In this study precision was defined as the percentage of search engines' top ten sites that were rated as top ten sites by participants. This precision measure was calculated for each participant and captures how well the search engines are determining the top ten sites to a search query.

Overall, Infoseek was 40.23% precise indicating that 40.23% of Infoseek's top ten sites were rated as top ten sites by participants. Overall, HotBot was only 16.00% precise. For the Hanggliding search, Infoseek was 32.69% precise while HotBot was 21.92% precise. For the Garlic Health search, Infoseek was 54.29% precise and while HotBot was only 6.07% precise. For the Acne search, Infoseek was 33.70% precise and HotBot 20.00% precise. Refer to Table 2 for a complete list of values.

These findings address Question 3 that asked if participants would agree with search engines' relevancy ratings. It was found that participants showed little agreement with search engines. The overall phi coefficients show very little relationship between the search engines' ratings of Web sites and participants ratings of Web sites and neither search engine was at least 50% precise in delivering the top ten sites for a search.

**Engine Superiority**

Additional calculations were performed to determine if one search engine was performing better than the other search engine. Two phi coefficients had
been calculated for each participant to measure the degree of agreement with each search engine. The number of times each search engine had a higher phi coefficient than the other was tabulated.

Overall (across all three topic areas), Infoseek had a higher phi coefficient for 68 out of 81 participants (83.95%). HotBot had a higher phi coefficient for 9 out of 81 participants (11.11%) and for four participants (4.94%) the engines were tied. A binomial test determined that Infoseek had a higher phi coefficient significantly more often ($p < .001$), even when the tied coefficients were assigned to HotBot. Within each topic, Infoseek had a significantly higher number of phi coefficients. In fact, in the Garlic Health search, Infoseek had a higher phi coefficient with every single participant (28/28). The Hanggliding search represented the poorest performance for Infoseek, yet Infoseek still had a higher phi coefficient than HotBot 65.38% of the time, reliably more often than chance ($p < .05$). See Table 3 for a complete list of values.

Recall that precision scores (the percentage of each search engine's top ten sites that were rated as top ten sites by participants) had been calculated. These precision scores were used to determine the superiority of one search engine in determining the top ten sites for a search. The number of times one search engine had a higher precision score than the other search engine was tabulated.
### Table 3

**Superior Engine Values**

<table>
<thead>
<tr>
<th>Topic</th>
<th>HotBot</th>
<th>Infoseek</th>
<th>Ties</th>
<th>Sig</th>
<th>HotBot</th>
<th>Infoseek</th>
<th>Ties</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hanggliding</td>
<td>6</td>
<td>17</td>
<td>3</td>
<td>*</td>
<td>7</td>
<td>18</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>(23.08%)</td>
<td>(65.38%)</td>
<td>(11.54%)</td>
<td></td>
<td>(26.92%)</td>
<td>(69.23%)</td>
<td>(3.84%)</td>
<td></td>
</tr>
<tr>
<td>Garlic</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td>***</td>
<td>0</td>
<td>28</td>
<td>0</td>
<td>***</td>
</tr>
<tr>
<td>Health</td>
<td>(0.00%)</td>
<td>(100%)</td>
<td>(0.00%)</td>
<td></td>
<td>(0.00%)</td>
<td>(100%)</td>
<td>(0.00%)</td>
<td></td>
</tr>
<tr>
<td>Acne</td>
<td>3</td>
<td>23</td>
<td>1</td>
<td>***</td>
<td>5</td>
<td>20</td>
<td>2</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>(11.11%)</td>
<td>(85.19%)</td>
<td>(3.70%)</td>
<td></td>
<td>(18.52%)</td>
<td>(74.07%)</td>
<td>(7.40%)</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>9</td>
<td>68</td>
<td>4</td>
<td>***</td>
<td>12</td>
<td>66</td>
<td>3</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>(11.11%)</td>
<td>(83.95%)</td>
<td>(4.94%)</td>
<td></td>
<td>(14.81%)</td>
<td>(81.48%)</td>
<td>(3.70%)</td>
<td></td>
</tr>
</tbody>
</table>

* p < .05. ** p < .01. *** p < .001
Overall, Infoseek had a higher precision score than HotBot for 66 out of 81 participants (81.48% of the participants). Participants agreed with HotBot's top ten more than Infoseek's only 12 out of 81 times (14.81%). For the remaining three participants, the two search engines returned an equal percentage of top ten sites. Even if the ties are assigned to HotBot, Infoseek was more precise in determining the top ten sites for a search significantly more often than HotBot ($p < .001$). In each individual topic area, a larger percentage of Infoseek's top ten sites were rated as top ten sites by participants a significant ($p < .05$) majority of the time (69.23% for Hanggliding, 100% for Garlic Health, and 74.07% for Acne).

These analyses show that Infoseek more closely matches participants' ratings of Web sites than does HotBot, thus providing an affirmative response to Question 4. Question 7 was also positively answered with these findings in that Infoseek was more precise in delivering the top ten sites than HotBot.

**Search Engine by Search Topic ANOVA**

In order to further examine the relationship between search engines' ratings and participants' ratings, and to see if this relationship was dependent on the search topic, a mixed (2) x 3 ANOVA was performed. Search engine was the within subject factor, search topic was the between subject factor, and the dependent variable was the phi coefficient capturing the agreement between users and search engines. A main effect of search engine was found indicating that Infoseek had a significantly higher phi coefficient (phi = .202) than HotBot (phi = -.280), $F(1,78) = 168.58$, $p < .001$. A main effect for topic was also found,
F(2,78) = 98.29, p < .001.Collapsed across engines, the Acne topic had the lowest phi coefficient (phi = -.09), followed by Hanggliding (phi = -.03), and then by Garlic Health (phi = .00).

There was also a significant topic by search engine interaction F(2,78) = 25.30, p < .001. See Figure 1 for a graph of this interaction. The phi coefficient between the participants' ratings and the search engines' ratings in the Garlic Health topic were Infoseek's highest (phi = .40) but HotBot's lowest (phi = -.40). In the Hanggliding topic condition, Infoseek performed its poorest and HotBot performed its best. However, as can be seen from Figure 1, Infoseek still performed better than HotBot in all three topic areas.

These findings provide an affirmative response to Question 6 that asked if the search topic area would moderate the relationship between participants' relevancy ratings and search engines' relevancy ratings. The search topic significantly influenced how well both search engines performed.
Figure 1

Search Engine by Topic Interaction
Participant Characteristics

Several individual difference variables were collected including age, gender, year in school, time spent on the Web, comfort level with search engines, browser-searcher classification, interest level in the search topic, and knowledge level of the search topic. The individual difference variables were all correlated with the search engines’ phi coefficients. It was found that for HotBot, knowledge in the search topic had a significant negative relationship with the correspondence between a participants’ ratings of Web sites and the search engines’ ratings. The correlation between knowledge level and HotBot’s phi coefficient with participants was -.26, p = .019. The number of hours per week spent on the Web was found to be positively correlated with HotBot’s phi coefficient, r =.23, p = .04. None of the other individual difference variables measured were significant predictors of the search engines’ phi coefficients.

Finally, a (2) x 2 x 3 mixed ANOVA was performed for each of the individual difference variables (knowledge of the topic area, interest in the topic area, browser-searcher classification, and amount of time spent on the Web). A median split was performed for each variable to examine if certain engines correspond to certain people on certain topics. Only one significant result was found in all of these ANOVAs (refer to Table 4 for a complete list of the F-values from each ANOVA). A significant interaction between search topic and knowledge of the search topic was found, F(2,75) = 3.32, p < .05. The interaction indicates that high knowledge users showed less agreement with the
search engines than low knowledge users in the acne topic, but in the
hanggliding topic high knowledge users showed more agreement with the search
ingines than low knowledge users.

These results do not lend much support to answer Question 5. Question 5
asked if the relationship between participants' relevancy ratings with the search
ingines' relevancy ratings would be moderated by several individual difference
factors. None of these factors were significant for both search engines, nor were
any of them robust predictors of the search engines' phi coefficients.
Table 4

Results of Individual Difference ANOVAs

<table>
<thead>
<tr>
<th>Participant Variables</th>
<th>Interest</th>
<th>Knowledge</th>
<th>Browser-Searcher</th>
<th>Hrs/week on Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic (T)</td>
<td>98.3**</td>
<td>77.85**</td>
<td>93.94**</td>
<td>94.86**</td>
</tr>
<tr>
<td>Engine (E)</td>
<td>171.94**</td>
<td>114.39**</td>
<td>162.07**</td>
<td>162.60**</td>
</tr>
<tr>
<td>Participant (P)</td>
<td>1.07</td>
<td>0.97</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>TxE</td>
<td>25.4**</td>
<td>20.64**</td>
<td>24.70**</td>
<td>24.59**</td>
</tr>
<tr>
<td>TxP</td>
<td>1.59</td>
<td>3.32*</td>
<td>0.12</td>
<td>0.78</td>
</tr>
<tr>
<td>PxE</td>
<td>1.93</td>
<td>0.93</td>
<td>0.02</td>
<td>2.47</td>
</tr>
<tr>
<td>TxExP</td>
<td>1.32</td>
<td>1.24</td>
<td>0.26</td>
<td>0.90</td>
</tr>
</tbody>
</table>

*p < .05.  **p < .001
Discussion

Agreement among Web Search Engines

Agreement among search engines is notably low. As Chu and Rosenthal (1996) pointed out, at times it seems as if search engines are capturing completely different parts of the World Wide Web. In a pilot study performed earlier this year (Dudziak, 1999), the same search query was entered into three separate search engines on the same day. The top 250 sites delivered by each of the three search engines (750 total sites) were recorded. Only 5 sites overlapped in all three engines. That is, five sites, out of the top 250 per engine, were all that the three search engines had in common.

In the present study, no overlapping sites were found in the Garlic Health search between Infoseek and HotBot. Only three overlapping sites existed in the Hanggliding and the Acne searches. This lack of overlap in search engines could mean a couple of things. First of all, the spiders could be collecting sites from completely different parts of the vast World Wide Web. Alternatively, the algorithms used by search engines to determine relevancy are totally unique for each search engine. Potentially, each search engine could be tailoring their search methodology for specific kinds of users or specific topic areas. Exactly why there is such a lack of overlap between search engines is still unknown, but users should not expect the results of searches performed in different engines to be the same.
Agreement among Users

Web search engines may not agree as to which sites are relevant to a search, but users certainly do. All three measures of agreement and reliability among participants were quite high, and significant. Kappa is a measure of absolute agreement among participants, corrected for chance agreement. In effect, what actual number participants assigned to a particular site is important, not just the relative ratings of sites. The average Kappa across all three search topics was 0.44, with significant agreement found in each topic area.

The fact that users have significant agreement in actual relevancy numbers assigned to Web sites is important. This indicates that users are in strong agreement as to what a numerical relevancy rating means, and how different sites should be rated. Many search engines provide numerical relevancy ratings for the sites they deliver. Although users may not agree with the relevancy numbers assigned by search engines, users do have some unified opinion on what those numbers should mean. If search engines can replicate the relevancy ratings assigned by users, people would be more satisfied with their searches.

The mean correlation and the intraclass correlation were computed as measures of inter-rater reliability. These correlations indicate how consistent the raters are in making relevancy ratings, a measure of inter-rater reliability. The mean correlation between participants was 0.64. Intraclass correlation is a better measure of reliability, however, because it takes into account the number of raters. There were not just two raters making these judgements, there were 26
to 28 people rating sites in each topic area. The average intraclass correlation was 0.97 and ranged from 0.97 to 0.98 in each topic area, indicating very strong reliability in ratings across participants. Both measures of reliability among participants were strong and significant.

These reliability indices indicate that even if users are not assigning the exact same relevancy ratings to Web sites, they are putting them in the same rank order. Users know which sites are poor, good, better, and best for a particular search topic and they are in strong agreement as to the rank order of sites. All search engines provide a list of Web sites in rank order from best to worst, or most relevant to the search to least relevant to the search. If users are in such strong agreement about the ordering of sites, search engines should be able to identify the characteristics of the sites that make them more or less relevant. As long as users are in agreement, a method of making search engines better suited for users is possible.

**Agreement between Users and Web Search Engines**

Overall, search engines are not delivering the most relevant sites according to users nor are they delivering sites in the correct order. HotBot even appeared to rate the sites in reverse order. The sites rated in the participants' top half were unlikely to be delivered by HotBot, and the sites that HotBot did deliver were often considered poor sites by participants. HotBot had a negative phi coefficient with participants in each topic area. Infoseek performed better than HotBot, as can be seen by the mere fact that all of Infoseek's phi coefficients were positive. Infoseek's performance was decent, it had a phi of
0.40 with participants in the Garlic Health condition indicating quite high levels of agreement. Overall, Infoseek had a phi equal to 0.20, while HotBot had an overall phi of -0.28.

The second measure of agreement between participants and search engines was the percentage of the search engine's top ten sites that were rated as top ten sites by participants, the precision score. This is a very simple, yet important measure. Search engines usually deliver the results to a search query with ten sites per page. The first thing a user sees as the result of a search query are the top ten sites according to the search engine. How much do people agree with the search engines' top ten? Overall, only 16% of HotBot's top ten sites were rated in participants' top ten. However, Infoseek was 40.23% precise in its' top ten overall. While it is clear that Infoseek performed better than HotBot, it should be kept in mind that these numbers are both artificially high because the participants' top ten was required to come from the approximately 34 sites evaluated.

The generally low phi coefficients indicate that search engines are still not very good at determining what it is that makes a Web site relevant to users. HotBot's performance was quite poor. With a negative phi coefficient in each topic area, HotBot was clearly unable to identify relevant sites for users. Even though Infoseek's performance was adequate in the Garlic Health search topic, it had phi coefficients close to zero (.05 and .14) for the other two searches. Although the criteria Infoseek uses to determine relevancy is more accurate than
HotBot's, it clearly could be better. The search engines are simply not sensitive to the criteria employed by users to judge relevancy.

*Is one of the Web Search Engines better than the other?*

This empirical test of two search engines has shown Infoseek to be better on every measure used to test the search engines. Even in determining the union of the top 17, Infoseek had half (8) the number of duplicate sites and dead links as did HotBot (17). Infoseek had a higher level of agreement with participants, as measured by the phi coefficient, significantly more often in each topic area and overall. Infoseek's top ten was more precise significantly more often in each topic area and overall. Even when the tied results are assigned to HotBot, Infoseek does significantly better on every measure. A main effect of search engine was found, indicating that Infoseek better matched actual users than did HotBot. The method used by Infoseek to determine which Web sites are relevant to users is clearly much better than the method used by HotBot.

Each search engine uses different methods to find Web sites and to determine the relevancy of those sites to a search. The results of these various methods have not been empirically tested by a set of representative users until now. It is clear from this study that the methodology employed by Infoseek is superior to that of HotBot. Now that one method is known to be better than another, it is time to determine what aspects of the methodology are making it better.

*Does the search query matter?*
There was a main effect of topic indicating that the topic of the search query influenced how well the search engines' sites and ratings matched up with participants' ratings of those sites. Overall, search engines did the best on the Garlic Health search, followed by the Hanggliding search, and the poorest on the Acne search.

This means that there is something about the search topic itself that influences how well search engines' ratings match up with participants' ratings. Perhaps finding out about the health benefits of garlic is a more common search than hanggliding or acne, and the engines have developed a better way of picking good sites for that topic area. It may also have to do with the number of sites the search engine has found related to that topic. If the number of sites available is very low, search engines cannot be choosy and finding good sites may be more difficult. More likely, however, is that the criteria used by search engines to determine relevant sites is constant, regardless of topic, and it should not be. Different topics may require slightly different algorithms to capture what is relevant about that topic. For example, in a hanggliding search pictures or diagrams may be important. However, in a search on the health benefits of garlic, pictures may be viewed as unnecessary and just slow down load times. Whatever the reasons, the search topic influences how well search engines perform.

Interestingly, there was also a topic by engine interaction. Infoseek performed its best on the Garlic Health search, while that search topic represented HotBot's poorest performance. HotBot did its best on the
Hanggliding search, while that was Infoseek's poorest performance. The nature of the search topic does influence the engines' performance, and different engines may be more suited to specific types of searches. The small number of different queries performed in the study prohibits making generalizations about what topic areas are better suited to specific search engines. Infoseek did perform better in all three topics in this study, however.

**Do characteristics of the participants matter?**

Several participant characteristics were measured, but none of them were strong positive indicators of the match between a search engine and a participant. Knowledge had a small negative correlation with the phi coefficient for HotBot, which could indicate that the more people know about a topic the less they agree with HotBot. Overall, none of the participant variables measured provided much insight into users' agreement (or lack thereof) with search engines.

It is possible that participant characteristics are important, and that this study just did not pick up on them. The characteristics that should be measured may be different than the ones measured here. It seems likely that search engines are not a "one-size fits all" type of mechanism. People are different and so are search engines. Perhaps certain search engines' methodology more closely matches a certain type of person's thinking, but that is still to be discovered.

**Future research**
Future research should examine more than two search engines and more than three search topics. If there are characteristics of participants or of search topics themselves that help predict what the participants want, search engines could be better designed to meet the needs of users. There are already search engines designed solely for finding people, so perhaps there should be separate search engines for every type of search topic imaginable. More simply, the user could tell the engine at the beginning of the query what type of search topic they are using. Is it a hobby, a sport, an illness, or a movie? With this knowledge the search engine could employ specific criteria when picking the relevant sites.

Perhaps instead of entering search type as part of the query, users should provide information about themselves. Users would log on to search engines and provide the engine with information about their individual characteristics. What information would be important for the engines is still unknown. The search engine could even be adaptive and learn what a particular user likes and dislikes about Web sites.

It is also important to start determining the characteristics of a Web page that make it relevant according to participants. Is it the content, design, ease of use, or even load time? Does this vary depending on the search topic or the specific user? How are people making relevancy judgements? What exactly do users want in return from a search query?

Conclusions

Practically everything is on the World Wide Web now. People shop for clothes, gifts, concert tickets, and even vacations on the Web. People get driving
directions, maps, and recipes off the Web. They search for information about diseases, hobbies, and career choices. People even locate lost family and friends over the Web. Search engines are the "on-ramp" for this information superhighway. The primary way to make using the Web easier and more productive for people is to improve Web search engines.

This study represents a step towards developing a methodology that can be used to empirically evaluate search engines, keeping in mind that everyday people are the users. Search engines have a long way to go before they reliably deliver what users want from a search. By determining exactly what users want, how users judge Web pages, and the characteristics of Web pages and search topics that influence relevancy judgements, large steps can be made in improving search engines.
References


NewEdge Corporation. (June, 1999). Media Metrix releases the top 50 at home and at work digital media and Web audiences ratings for May 1999 [WWW


[WWW page] URL
http://www2.pcworld.com/software/internet_www/articles/jan96/jan9635.html


Shirky, C. (1999). The hit parade: The World Wide Web is now as big as TV. Will it ever be the same? [WWW page]. URL


Appendix A1

Demographic Sheet

Please provide the following information:

Age: _______

Year in school (freshman, sophomore, etc..): ____________

Average number of hours per week spent on the World Wide Web: _________

Please rate your behavior when you use the Web on the following scale.

1 2 3 4 5 6 7 8 9 10
jump freely from link
very goal directed, never follow
to link, regardless of topic
a link off the topic

Please put an X next to the search engines you have used.

___Yahoo!  ___Northern Light
___HotBot   ___WebCrawler
___Excite   ___Snap
___Go/InfoSeek  ___MetaCrawler
___AltaVista  ___Netscape Search
___Ask Jeeves  ___MSN Search
___FAST Search  ___Lycos
___Other(s) Please list:

__________________________________________________________

Do you have a preferred search engine?  Y / N
   If Yes, which one(s)? _______

Please explain why you have this preference
How comfortable do you feel using Web search engines?

1 2 3 4 5 6 7 8 9 10
not comfortable at all completely comfortable

Please rate your interest level in the topic area “the health benefits of garlic” on the following scale:

1 2 3 4 5 6 7 8 9 10
No Interest moderately interested very interested

Please rate your knowledge level in the topic area “the health benefits of garlic” on the following scale:

1 2 3 4 5 6 7 8 9 10
No knowledge moderate knowledge very high knowledge

Please rate your interest level in the topic area “acne” on the following scale:

1 2 3 4 5 6 7 8 9 10
No Interest moderately interested very interested

Please rate your knowledge level in the topic area “acne” on the following scale:

1 2 3 4 5 6 7 8 9 10
No knowledge moderate knowledge very high knowledge
Appendix A1 (continued)

Demographic Sheet

Please rate your interest level in the topic area “hanggliding” on the following scale:

1  2  3  4  5  6  7  8  9  10

No  Interest  moderately interested  very interested

Please rate your knowledge level in the topic area “hanggliding” on the following scale:

1  2  3  4  5  6  7  8  9  10

No knowledge  moderate knowledge  very high knowledge
Appendix B

Written Instructions

Welcome to experiment #310b

Today, your task will be to judge a series of Web pages. You will be asked to pretend that you are writing a research paper on a topic area and are presented with this list of Web sites as your information source. You will decide how relevant each Web site is to your search for information on your paper topic. How well would you feel your time had been spent if a Web site was the only one you saw?

First, view all the Web sites on the list. You will be given a page on which you can take notes. You can go back to the Web sites as often as you like throughout the experiment.

When you have viewed all of the Web sites notify the experimenter and you will be given a rating sheet. Each rating will be on a scale of 0-100 with 0 indicating completely irrelevant and 100 indicating completely relevant (100% relevant to the search topic).

This rating sheet has a list of all the Web sites you will have seen. You will simply put your rating for the site next to the title of the Web site. When you finish your ratings, please inform the experimenter.

This experiment will last approximately the full hour, so please take your time and make your ratings thoughtfully and carefully.

Thank you.
Appendix C

Second topic area explanation, example is from the Hanggliding search

Imagine that you are asked to write a paper on "Hanggliding."

Your search results will be displayed on a page called "Hanggliding Index Page." Please view all of the sites on your index page, keeping in mind that you will be asked to rate the Web sites from 0 (completely irrelevant to the topic) to 100 (completely relevant to the topic).

Feel free to take any notes on the paper provided. When you have viewed all of the Web sites, notify the experimenter and you will be given a rating sheet. You may visit each Web site as many times as you would like, but do not follow any of the links on the Web sites.

If you have any questions please ask the experimenter now.

Thank you again for taking your time and using care in making these evaluations.
Appendix D
Sample Hanggliding Index Page

Hanggliding Index Page

Ottawa Hang Gliding & Paragliding Club (OHPC)
SkyFly - paragliding, hanggliding, ULLi and ULL
AviationNet Aviation Web Directory/Airlines
KO's Paragliding & Hanggliding
Landings: Hanggliding / Paragliding Links: Index - H
Hanggliding and Para-Gliding FAQ
Hang Gliding (4x4)
High Adventure - Hang gliding, Hanggliding, Gli
Dolomiti Delta Club - Cortina d'Ampezzo - Italy - H
HANG GLIDING resources from Nerd World Media
Zoeblen: Zoeblen, Tirol, alpine skiing, fishing, canyoning, hanggliding, bicycle riding
Hanggliding Pictures, hangglider flying, in the B
Anten Wisata Indonesia: Airsport-travels to Sum
Hang Gliding Digest HTML Archive
Hanggliding Pictures, paraglider flying, in the B
G&D's Hanggliding Page
http://www.univie.ac.at/ftp/archive/faq/hang-gliding.faq
ABC Airsports Hanggliding
UMass Hang Gliding Club
Crispi Boots. Made for paragliding, hanggliding
Albatros HFK
Kite Enterprises Hang Gliding and Paragliding H
Hanggliding Maui hang gliding and ultralight in
Skywings on the Web
www paragliding greece sitemap alexiptoto plagias parapente
Bob Mackey's Hanggliding Page
Golden Wings HangGliding / Paragliding
Hanggliding/Paragliding
1198website
AUSTRALIA HANG GLIDING and over 1000 activities
DamnHot Gear - Air Sports, Hanggliding, Paragl...
Appendix E
Sample Garlic Health Index Page

Garlic Health Index Page

RealAge: Commentaries
Healthy Ideas: Ask Chef Ney -- Roasting Garlic
Harris Garlic Farm
A garlic chemistry primer
Science Model Assessment Item 34
GH: Health Check -- What's So Good About Garlic?
Garlic Fights Ulcers by BodyAtlas
Mother Nature - Health and Nutrition
The Chemistry Of Garlic Health Benefits
CONFERENCETO FOCUS ON GARLIC HEALTH BENEFITS
HerbalShopper.com - Garlic - Odorless
Web Centers | Health | Hippocrates 1988 Jan-Fe
GarliPure
XINJIANG HEALTH GARLIC POWDER (FREEZE DRIED) -G-
Deodorized Garlic Tablets - Vitamin Power Nutritional Health Supplements
Sympatico
Epicurean's Delight: Health Center
Xinjiang Tianhe Health Products Co., Ltd.
http://www.uvm.edu/~rouellet/refer3.html
Tito & Vina Racho: Garlic Supreme
Geiger Elephant Garlic
Garlic - Research into health benefits continues
Garlic & Your Health | Natural Health | Health
Herb-of-the-Month: Garlic (Allium sativum) - De
Welcome to Fitness Online
The Garlic Page
GARLIC HEALTH BENEFITS - Medicinal Uses of Herb
Garlic at PlanetRx
For' Mor International [Health, Wealth, Service]
Lycos Health News
Health World - The Chemistry of Garlic Health B
WholeFoods.com : Market : Health & Body : Essen...
Garlic and its key constituents have been a s...
Health Benefits And Folklore Of Garlic
Appendix F
Sample Acne Index Page

Acne Index Page

The Health Network - Internet Directory - Acne
4Acne -- a guide to treating acne from 4anythin
FDA Consumer Reprint--On the Teen Scene: Acne A
Acne (healthgate)
Acne (doctorpage)
WebCrawler Health: Diseases & Conditions: Skin
Acne Scar Treatment - Dermatique Cell Renewal F
ACNE (amazing)
Acne (uiuc)
Treating acne whatever your age
The role of the physician
In Your Face! What Acne is and What You Can Do
New Study Links Mild Acne to Depression, Suicid
Skin Care Today: Acne: What Are The Psychologica
Patient Information Pamphlets: INDEX
Your No Longer Have to let Acne Affect Your Lif
Acne (bewell)
Yahoo! Health>Diseases and Conditions>Acne
Acne (Diseases and Disorders)
acne pamphlet page
Power Clear: acne treatment supplement
Acne Facts
Welcome to the Acne Support Group (ASG) Web site
acne (medstat)
Showbiz Confidential: The Agony and the Acne
Magellan Health: Professional Medicine: Special
Acne scars treatment
Acne, Pimples, Blackheads, Zits - Treatment and
Acne (omak)
Acne Products
Acne Pimples Complexion
<table>
<thead>
<tr>
<th>Web Site Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>RealAge: Commentaries</td>
</tr>
<tr>
<td>Healthy Ideas: Ask Chef Ney -- Roasting Garlic</td>
</tr>
<tr>
<td>Harris Garlic Farm</td>
</tr>
<tr>
<td>A garlic chemistry primer</td>
</tr>
<tr>
<td>Science Model Assessment Item 34</td>
</tr>
<tr>
<td>GH: Health Check -- What's So Good About Garlic?</td>
</tr>
<tr>
<td>Garlic Fights Ulcers by BodyAtlas</td>
</tr>
<tr>
<td>Mother Nature - Health and Nutrition</td>
</tr>
<tr>
<td>The Chemistry Of Garlic Health Benefits</td>
</tr>
<tr>
<td>CONFERENCE TO FOCUS ON GARLIC HEALTH BENEFITS</td>
</tr>
<tr>
<td>HerbalShopper.com - Garlic - Odorless</td>
</tr>
<tr>
<td>Web Centers</td>
</tr>
<tr>
<td>GarliPure</td>
</tr>
<tr>
<td>XINJIANG HEALTH GARLIC POWDER (FREEZE DRIED) -G-</td>
</tr>
<tr>
<td>Decodorized Garlic Tablets - Vitamin Power Nutritional Health Supplements</td>
</tr>
<tr>
<td>Sympatico</td>
</tr>
<tr>
<td>Epicurean's Delight: Health Center</td>
</tr>
<tr>
<td>Xinjiang Tianhe Health Products Co., Ltd.</td>
</tr>
<tr>
<td><a href="http://www.uvm.edu/~rouellet/refer3.html">http://www.uvm.edu/~rouellet/refer3.html</a></td>
</tr>
<tr>
<td>Tito &amp; Vina Racho: Garlic Supreme</td>
</tr>
<tr>
<td>Geiger Elephant Garlic</td>
</tr>
<tr>
<td>Garlic - Research into health benefits continues</td>
</tr>
<tr>
<td>Garlic &amp; Your Health</td>
</tr>
<tr>
<td>Herb-of-the-Month: Garlic (Allium sativum) - De</td>
</tr>
<tr>
<td>Welcome to Fitness Online</td>
</tr>
<tr>
<td>The Garlic Page</td>
</tr>
<tr>
<td>GARLIC HEALTH BENEFITS - Medicinal Uses of Herb</td>
</tr>
<tr>
<td>Garlic at PlanetRx</td>
</tr>
<tr>
<td>For'Mor International [Health, Wealth, Service]</td>
</tr>
<tr>
<td>Lycos Health News</td>
</tr>
<tr>
<td>Health World - The Chemistry of Garlic Health B</td>
</tr>
<tr>
<td>WholeFoods.com: Market: Health &amp; Body: Essen...</td>
</tr>
<tr>
<td>Garlic and its key constituents have been the s...</td>
</tr>
<tr>
<td>Health Benefits And Folklore Of Garlic</td>
</tr>
<tr>
<td>Rating (0-100)</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Sympatico</td>
</tr>
<tr>
<td>GARNIC HEALTH BENEFITS - Medicinal Uses of Herb</td>
</tr>
<tr>
<td>RealAge: Commentaries</td>
</tr>
<tr>
<td>The Garlic Page</td>
</tr>
<tr>
<td>Garlic - Research into health benefits continues</td>
</tr>
<tr>
<td>Garlic &amp; Your Health</td>
</tr>
<tr>
<td>Xinjiang Tianhe Health Products Co., Ltd.</td>
</tr>
<tr>
<td>Web Centers</td>
</tr>
<tr>
<td>Lycos Health News</td>
</tr>
<tr>
<td>GarliPure</td>
</tr>
<tr>
<td>Geiger Elephant Garlic</td>
</tr>
<tr>
<td>Health Benefits And Folklore Of Garlic</td>
</tr>
<tr>
<td>Herb-of-the-Month: Garlic (Allium sativum) - De</td>
</tr>
<tr>
<td>XINJIANG HEALTH GARLIC POWDER (FREEZE DRIED) - G-</td>
</tr>
<tr>
<td>Welcome to Fitness Online</td>
</tr>
<tr>
<td>Garlic Fights Ulcers by BodyAtlas</td>
</tr>
<tr>
<td>Science Model Assessment Item 34</td>
</tr>
<tr>
<td><a href="http://www.uvm.edu/~rouellet/refer3.html">http://www.uvm.edu/~rouellet/refer3.html</a></td>
</tr>
<tr>
<td>The Chemistry Of Garlic Health Benefits</td>
</tr>
<tr>
<td>Garlic and its key constituents have been the s...</td>
</tr>
<tr>
<td>Harris Garlic Farm</td>
</tr>
<tr>
<td>CONFERENCE TO FOCUS ON GARLIC HEALTH BENEFITS</td>
</tr>
<tr>
<td>Garlic at PlanetRx</td>
</tr>
<tr>
<td>GH: Health Check -- What's So Good About Garlic?</td>
</tr>
<tr>
<td>Mother Nature - Health and Nutrition</td>
</tr>
<tr>
<td>Tito &amp; Vina Racho: Garlic Supreme</td>
</tr>
<tr>
<td>Epicurean's Delight: Health Center</td>
</tr>
<tr>
<td>For' Mor International [Health, Wealth, Service]</td>
</tr>
<tr>
<td>Deodorized Garlic Tablets - Vitamin Power Nutritional Health Supplements</td>
</tr>
<tr>
<td>WholeFoods.com : Market : Health &amp; Body : Essen...</td>
</tr>
<tr>
<td>HerbalShopper.com - Garlic - Odorless</td>
</tr>
<tr>
<td>A garlic chemistry primer</td>
</tr>
<tr>
<td>Health World - The Chemistry of Garlic Health B</td>
</tr>
<tr>
<td>Healthy Ideas: Ask Chef Ney -- Roasting Garlic</td>
</tr>
</tbody>
</table>