Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1075–1098
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

The effectiveness of Web search engines for retrieving relevant ecommerce links

Bernard J. Jansen [a,*], Paulo R. Molina [b]

[a] School of Information Sciences and Technology, 329F IST Building, The Pennsylvania State University, University Park, PA 16801, USA
[b] Supply Chain and Information Systems, The Smeal College of Business, The Pennsylvania State University, University Park, PA 16801, USA

Received 11 April 2005
Available online 24 October 2005

Abstract

Ecommerce is developing into a fast-growing channel for new business, so a strong presence in this domain could prove essential to the success of numerous commercial organizations. However, there is little research examining ecommerce at the individual customer level, particularly on the success of everyday ecommerce searches. This is critical for the continued success of online commerce. The purpose of this research is to evaluate the effectiveness of search engines in the retrieval of relevant ecommerce links. The study examines the effectiveness of five different types of search engines in response to ecommerce queries by comparing the engines' quality of ecommerce links using topical relevancy ratings. This research employs 100 ecommerce queries, five major search engines, and more than 3540 Web links. The findings indicate that links retrieved using an ecommerce search engine are significantly better than those obtained from most other engines types but do not significantly differ from links obtained from a Web directory service. We discuss the implications for Web system design and ecommerce marketing campaigns.

Keywords: Web searching; Web search engines; Web search engine evaluation; Ecommerce searching; Paid searching; Sponsored results; Organic results; Non-sponsored links

1. Introduction

One of the measures of search engine effectiveness is the retrieval of relevant links to ecommerce queries. This has not been the focus of considerable past research. Prior studies have either mostly addressed issues pertaining to search engine effectiveness for general Web searches (Spink & Jansen, 2004) or have commented on trends in ecommerce-related Web searching (Spink, Jansen, Wolfram, & Saracevic, 2002).

While these studies present valuable information on *how* users search the Web to address ecommerce needs, they have ignored how *effective* these searches are. Web search tools are crucial to the further development of

* Corresponding author.
E-mail addresses: jjansen@acm.org, jjansen@ist.psu.edu (B.J. Jansen), prm139@psu.edu (P.R. Molina).

0306-4573/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2005.09.003

ecommerce, a sector that witnessed a 26% increase in retail sales in 2003 to $55 billion (The Economist Survey, 2004a). Although impressive, the growth accounts for only 1.6% of total worldwide retail sales, online and offline, indicating both huge potential in the further development of ecommerce Websites and the need for ecommerce friendly search engines (The Economist Survey, 2004a).

We report research measuring the effectiveness of five types of search engines in their retrieval of relevant links for ecommerce queries. Our study also evaluates user-determined relevancy for specified search results in response to particular ecommerce queries. The goal of our research is to determine whether or not an ecommerce-focused engine increases relevancy or if other types of search engines perform just as well for ecommerce searchers.

We begin with a review of literature by briefly addressing ecommerce Web searching and then presenting a comprehensive discussion of Web search engine evaluation research in order to establish the state of the field. We then discuss the study design and methodology used to measure the effectiveness of the search engines for ecommerce searching. Following, we present the results of the evaluation and discuss the implications of the results. We offer recommendations and implications for potential changes to system design, directions for future research, and conclusions.

2. Literature review

The Internet is still capable of producing surprises. New services can leave people wondering how they ever managed without them. The most notable of these is the search engine (The Economist Survey, 2004b, p. 1).

We have witnessed rapid growth in the search engine market since its inception. Search engines continue to attract large number of Web searchers and consistently rank as some of the heavily visited sites in the market in terms of number of visitors (Nielsen Netrating, 2002). On the consumer side, surveys indicate that search engines are the most important promotional method used by ecommerce sites and they represent the most common way new ecommerce sites market themselves. In addition, studies show that users spend a significant amount of time on search engines looking for relevant information to their ecommerce queries (Gandal, 2001). Ecommerce is the next big instigator the further development of the Web, and new tools to search for ecommerce information on the Web are needed (Wigand & Benjamin, 1995).

The buying and selling of products and services over the Web is now a part of everyday life for millions of people. Research shows that over the next decade online retailers may capture 10–15% of retail sales (The Economist Survey, 2004a), and that the Web is increasingly becoming the gateway to a company's brand, products and services. Ecommerce has several definitions within the literature, but these definitions usually contain the common elements of buying and selling over a network. For example, Turban states that ecommerce is the "process of buying, selling, transferring, or exchanging products, services, and/or information via computer networks" (Turban, King, Lee, & Viehland, 2004, p. 3). Rosenbaum defines ecommerce as "the use of networks to conduct business-to-consumer or business-to-business transactions involving the exchange of information, currency, and both digital and non-digital goods and services" (Rosenbaum, 2000).

There are numerous ongoing research centers on ecommerce, such as the Wharton Forum on Electronic Commerce (http://econ.Wharton.upenn.edu) and eBusiness Research Center (http://www.smeal.psu.edu/ebrc/). Ecommerce offers numerous advantages, such as allowing for lower search cost, speedy comparisons of products, and personalization (Turban et al., 2004). Web consumers shop online or use online services to save time (Bellman, Lohse, & Johnson, 1999). However, the success of online shopping sites is dependent upon customers actually finding relevant sites (Spiteri, 2000).

2.1. Web search engine evaluation

For locating ecommerce sites, search engines and directories are the most frequently used Web portals (Awad, 2004). Several sources report that more than 80% of Web visitors use a search engine as a starting point (Kehoe & Pitkow, 1996; Sullivan, 2003). How good are these search engines at locating relevant ecommerce Web sites? There have been an increasing number of studies evaluating search engines with many of these focused on establishing a standard set of metrics generally in line with those for general searching system evaluation. One can group these metrics into categories of relevance evaluations, ranking, and stability of links.

2.1.1. Web search engine relevance evaluation

Several researchers have examined elements relating to precision. Chu and Rosenthal (1996) used three Web search engines (AltaVista, Excite, and Lycos) and 10 queries to examine precision with a three-level relevance score (relevant, somewhat relevant, and irrelevant) for the top 10 links. The authors report that AltaVista outperformed the other search engines in search facilities and retrieval performance. The researchers also proposed a methodology for evaluating Web search engines.

Ding and Marchionini (1996) evaluated three search engines using five topics. Along with including links to other relevant documents the researchers employed a six-point scale and used it to calculate three different types of precision. They found no significant differences among the search engines. Nicholson (2000) replicated this study, using chaos theory to highlight possible patterns.

Tomaiuolo and Packer (1996) used 200 queries, the first 10 links, and three search engines (Lycos, InfoSeek, and AltaVista), along with two human evaluative search tools (Magellan and Point). Clarke and Willett (1997) examined recall of AltaVista, Excite, and Lycos by using a pooling technique. Leighton and Srivastava (1999) compared five search engines using 15 topics, reporting statistically superior performance among three of the search engines. The researchers determined relevance themselves, after cloaking the identity of the search engine that retrieved the links.

Gordon and Pathak (1999) evaluated eight search engines (AltaVista, Excite, Infoseek, Open Text, HotBot, Lycos, Magellan, and Yahoo!) using 33 queries and the top 200 links. The researchers used a four-level relevance judgment (highly relevant, somewhat relevant, somewhat irrelevant, and highly irrelevant) and one evaluator (i.e., the person that had requested the information search). Using an analysis of variance, AltaVista, Open Text, and Lycos had statistically higher precision and recall compared with the other five search engines. In an interesting twist for search engine evaluation, the researchers had expert searchers translate the information need into a series of queries, selecting the optimal query over a set time period. This approach of using expert searchers is not normally done when evaluating end user searching systems, such as Web search engines.

Hawking, Craswell, Bailey, and Griffihs (2001) evaluated the effectiveness of 20 search engines using Text REtrieval Conference (TREC)—inspired methods with 54 queries taken from real Web search logs, although the researchers limited the submissions to natural language queries. The researchers pooled the top 20 links from each search engine (although a bug in the script prevented retrieval of all links from all search engines) and presented them to one evaluator. The researchers examined various performance measures, including precision at various document cut-off values (DCV), mean reciprocal rank of first relevant document, and TREC-style average precision. Statistical testing revealed high inter-correlations between performance measures and significant differences between performances of search engines, with Northern Light the top ranked search engine.

Using a task-centered approach, Spink (2002) evaluated the Inquirus, a meta-Web search engine, with 22 volunteers searching their own information tasks. Each searcher rated the top 20 Web documents on relevance, using a four part scale. Along with usability, the searchers evaluated the precision of Inquirus (27%) which is lower than that reported for major Web search engines, which is approximately 50% (Eastman & Jansen, 2003).

Wildemuth and Carter (2002) also took a usability approach by looking at the perceived affordances of nine search engine (AltaVista, Ask Jeeves, Excite, Google, Hotbot, LookSmart, Lycos/Open Directory, Northern Light, and Yahoo). The analysis metrics included the text box, search button, search syntax, searching help, the number of links, query modification, directory structure, links displays, and user preferences.

Shang and Longzhuang (2002) evaluated six popular search engines, (AltaVista, Fast, Google, Go, iWon, and NorthernLight), with 3000 queries from two domains, using a generally automatic test design. The researchers computed relevance scores using three difference relevance algorithms and statistical comparisons of the ranking.

Su (2003a, 2003b) proposed and tested a model of user evaluation of Web search engines. Using data collected in 1998, the researcher evaluated four major search engines (AltaVista, Excite, Infoseek, and Lycos) using 36 undergraduates. Each subject selected his or her own topic on all search engines. The researcher used the top 20 links from the "best" session. The researcher proposed 16 performance measures within five evaluation criteria: relevance, efficiency, utility, user satisfaction, and connectivity. AltaVista had the statistically highest precision among the four search engines.

Veritest (2003) evaluated the relevancy of Google, Wisenut, Fast, Teoma and AltaVista, and compared the top 10 links using 100 randomly selected queries from the Inktomi engine transaction log. Three independent

evaluators judged relevancy. Inktomi and Google scored higher on all tests compared to the other search engines. The researchers conducted no statistical analysis comparing search engines.

Griesbaum (2004) investigated the retrieval effectiveness of three German Web search services (AltaVista.de, Google.de and Lycos.de), with 50 queries, using the top 50 links. One independent evaluator judged two links sets and made relevance assessments. Statistical analysis showed that Google performed significantly better than AltaVista, but there was no significant difference between Google and Lycos.

Using 12 queries, Wu and Li (2004) evaluated the effectiveness of four search engines (Google, AllTheWeb, HotBot, and AltaVista) and four meta-search engines (MetaCrawler, ProFusion, MetaFind, and Meta-EUREKA). Experimental results showed that on average the performances of selected meta-search engines and search engines are very close, although there were no statistical tests evaluating search engine differences performed.

Vaughan (2004) used three search engines (Google, AltaVista, and Teoma), four queries, and the first 10 links from each search engine. The researcher used 24 evaluators to rate the links. The researcher examined quality of ranking compared to human evaluations, ability to retrieve top-ranked pages, and stability. Google performed the best in rankings, in ability to retrieve top ranked pages, and in stability over a 10 week period.

2.1.2. Web search engine ranking evaluation

Ranking has been shown to improve user satisfaction with information retrieval systems (Witten, Moffat, & Bell, 1994). There have been limited research in this area (Courtois & Berry, 1999). Gwizdka and Chignell (1999) acknowledged the importance of result ranking in Web information retrieval and developed "differential precision" to measure the quality of ranking produced by search engines. Su, Chen, and Dong (1998) used a five-point relevancy scale and then correlated these rankings with the rankings returned by search engines.

Singhal and Kaszkiel (2001) compared the performance of a state-of-the-art keyword-based document ranking algorithm with four Web search engines (Excite, Google, Lycos, and AltaVista Raging) on the ability to retrieve relevant organizational or individual Web pages within the top 10 links. The researchers made the relevance judgments and reported that Web search engines are notably better than a state-of-the-art keyword algorithm.

Jansen (2000) also compared links among several search engines, noting that regardless of query formulation, there was an approximately 60% overlap on search engines. Eastman (2002) reported that links of searches using a variety of query formulations with several Internet search engines show that strategies intended to give narrower and more precise results may not give improvements in precision. Eastman and Jansen (2003) reported that query formulation has no significant effect on coverage, ranking, or relevance of links on the Google, MSN, or AOL search engines, although Google did have the statistically highest prevision.

One may need to evaluation of search engines often due to either the changing needs of users or the dynamic nature of search engines (e.g., their changing Web coverage and ranking technology); therefore, evaluation needs to be efficient. To automate the evaluation process, some researchers have explored software applications rather than human evaluators. Courtois and Berry (1999) studied the first 100 items retrieved by five search engines (AltaVista, HotBot, Excite, Lycos, and Infoseek) using 12 queries developed by the researchers. They evaluated the ranking of the search engines using three criteria: at one occurrence of all search terms, at least one occurrence of all search terms appearing as a contiguous phrase, and at least one occurrence of all search terms appearing within the title, headers, or metatags. By these metrics, Excite had the best ranking for the top 20 documents; however, there was no statistical comparison performed.

Soboroff, Nicholas, and Cahan (2001) suggested an automatic method that maps queries with a random set of documents. Chowdhury and Soboroff (2002) used five search engines (Lycos, Netscape, Fast, Google, HotBot), and compared their ranking performance automatically, finding that the performance for most search engines is are statistically equivalent. The researchers used 2000 queries taken from AOL transaction logs, pairing these queries with known Web documents.

Can, Nuray, and Sevdik (2004) proposed an automatic Web search engine evaluation method as an efficient and effective assessment tool of such systems. Using eight Web search engines (AllTheWeb, AltaVista, HotBot, InfoSeek, Lycos, MSN, Netscape, and Yahoo!), 25 queries, and the top 20 links, the researchers used binary user relevance judgments, determined by the information requester. Using Pearson's r correlation, the researchers report that their method provided results consistent with human-based evaluations.

2.1.3. Web search engine result stability

There is a growing interest in the stability of search engine results (Rousseau, 1998/1999; Vaughan & Thelwall, 2004), stability being the re-occurrence of Web documents within search engine results to the same query over time. Petersen (1997) examined the total number of documents returned by eight different search tools at three points in time over 10 months. Mettrop and Nieuwenhuysen (2001) tested the consistency of links from 13 search engines, concluding that content stability should be considered as a performance measure for Web search engines.

Bar-Ilan has conducted several studies investigating the content stability of search engines (Bar-Ilan, 1998/1999, 2000, 2002). Bar-Ilan (2002) defined several metrics for evaluating search engine functionality over time. Bar-Ilan (1998/1999) used six search engines (AltaVista, Excite, Hotbot, Infoseek, Lycos and Northern Light) and one query during a 6-month time period. The researcher reported that Excite was the least stable of the search engines.

Mowshowitz and Kawaguchi (2002) measured the performance of 12 search engines with 12 queries and the top 20 links, using the overlap of URLs of the matching pages. The researchers stated that Northern Light was consistently the most biased search engine in terms of stability of links.

These evaluations of Web search engines have rarely focused on ecommerce specific tasks and information needs. They seldom used multiple raters on Web links and usually did not conduct statistical analysis among search engines. More over, these evaluations typically employed general purpose Web search engines, rather than ecommerce search engines. These studies have not specifically examined ecommerce queries.

2.2. Ecommerce search engine evaluation

Ecommerce search engines cater specifically to customers rather than the broad range of Web users (Seda, 2004). The assumption is that these Web customers are looking for product information or to purchase a product online. People use ecommerce search engines because they can compare prices quickly, compare products, find stores, lower risk, locate the most popular product, narrow choices, and compare store reputations (ForeSee Results, 2002, 2004).

There are few analyses or evaluation of ecommerce search engines, and fewer evaluations of how well general search engines support ecommerce searching. Bakos (1997) addressed the search market theoretically, arguing that reduced searching cost would fuel intense competition. Spiteri (2000) analyzed 6 subject directories for their support of ecommerce searching. The researcher reported that the categories used to organize ecommerce sites were ambiguous and provided few opportunities for comparison shopping.

Mukhopadhyay, Rajan, and Telang (2004) modeled search engine competition, showing that lower quality search engines can survive as long as some searchers fail at the higher quality search engines. Gandal (2001) examined the effect of early entrance into the search engine market, noting that this provided some advantages but has declined over time, with the exception of Yahoo!, which has offered a superior product.

In summarizing the state the art of Web search engine evaluation, we see a series of studies examining both precision and usability of various search engines, using a variety of methods for determining relevance. In many studies, the number of queries was limited although varied in topic. There are also a growing number of studies examining the stability of Web search engine links. There are fewer studies that examine the effectiveness of search engine ranking.

Ultimately, we could not locate a study that dealt specifically with the evaluation of ecommerce queries and the effectiveness of ecommerce Web search engines in the retrieval of relevant links. Given the growing importance of ecommerce and online searching, this is a critical need, with implications for customers, Web search engines, and large and small commercial organizations.

3. Research questions

We address three research questions in this study: (1) Which type of search engine performs better for ecommerce searching? (2) Are the top ranked links the most relevant for ecommerce searching? and (3) Are sponsored links more relevant than organic links for ecommerce searching?

Our three hypotheses addressing these three research questions are:

Hypothesis 1. There will be a significant difference in the number of relevant links retrieved using an ecommerce search engine relative to non-ecommerce search engines.

We hypothesized that an ecommerce specific search engine would perform better than other types of search engines within the ecommerce domain. Niche search engines have been reported to improve the relevancy of results within a particular topical area (Glover, Lawrence, Gordon, Birmingham, & Giles, 2001), although we could locate no study supporting this assertion. For this, and the other hypothesis, we examined the relevancy of the search links retrieved by the various search engines, comparing these results to user relevancy evaluations. We compared an ecommerce search engine to four other search engine types. The findings could impact the development of future ecommerce search engines.

Hypothesis 2. Relevance of links will decrease as the rank of links increases (i.e., the more relevant links have the lowest rank).

Search engines rank the "most relevant" links at the top of the links listing, assigning the "most relevant" result the first rank, the next "most relevant" result the next lowest rank, etc. However, the search engine determines this relevancy algorithmically, and this may not conform to the searcher's evaluation. This is an important issue in the ecommerce area, given that the majority of searchers only view the first results page (Jansen, Spink, & Saracevic, 2000).

Hypothesis 3. Sponsored links will be more relevant than organic links.

Given that these are ecommerce queries, one would expect that the sponsored links would be especially relevant to these types of queries. Commercial organizations bid on key terms and phrases directed at the customer market for their business. Search engines also perform a measure of quality control, ensuring that the Web sites are related, in some way, to the key terms and phrases. Findings could impact pay-for-inclusion search engines and search engine optimization approaches.

Fig. 1 illustrates a typical placement of organic and sponsored links on a search engine results page (SERP), in this case from Google, in response to the query *bedding*.

4. Research methods

We designed and conducted a study to address our three research questions. We first selected our search engines for use in the study.

4.1. Web search engines

FirstPlace Software, Inc. (http://www.web-positiongold.com/) categorizes five types of search engines, which are meta-search, ecommerce niche, general purpose, pay-for-inclusion, and directory service (FirstPlace Software, 2004). Although each type of search engine may have different design criteria, all support ecommerce searching. The following describes each search engine type:

- Metacrawler search engines: These Web search engines do not have their own database. Instead, they search other search engines and rank the links according to a meta-ranking algorithm.
- Ecommerce search engines: A type of niche search engine, these engines focus on offering product comparisons, costs, and reviews.
- General purpose search engine: These Web search engines typically use crawlers to index Web documents. These search engines target the general population of Web searchers and the entire spectrum of information needs.
- Pay per click: Instead of ranking pages by their actual content, these Web search engines rank sites based on who pays the most per click to be listed for the keyword(s) being searched. One can place bids for top positions on these sites in an auction-like environment. One can secure top positions for as little as 5 cents per

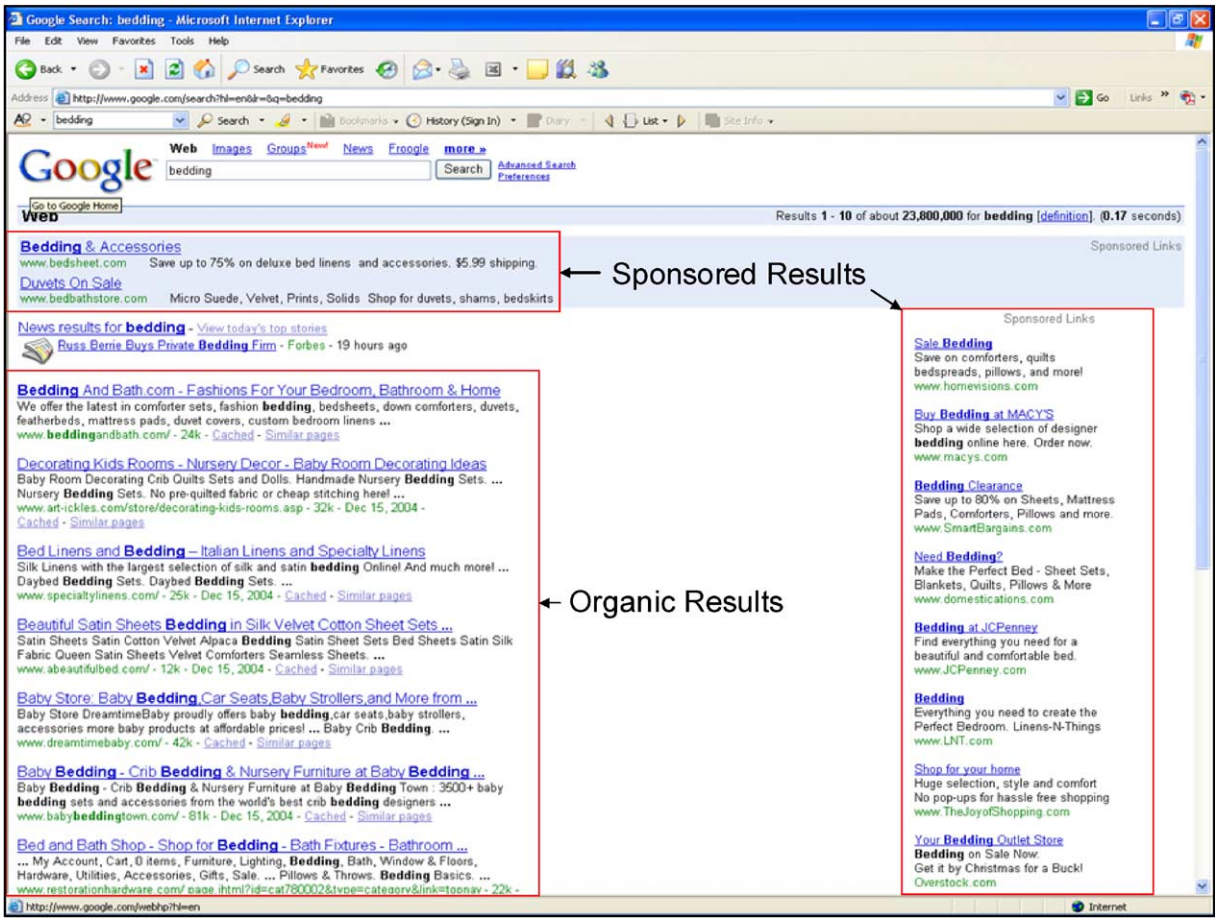

Fig. 1. Typical placement of organic and sponsored links on results page.

click (i.e., visitor to your site) at the major Pay-Per-Click services. Many major search engines usually include several pay per click listings in a sponsored portion of their search results (see Fig. 1).

- Directory engines: These are search services that manage a directory of categorized Web sites. These search services generally have human beings review Web documents and submissions. If these information objects meet with human reviewers' standards, the search engine will include the documents in the directory. Most directory-based search engines supplement their listings with crawler-based results.

We selected the most popular search engine within each of the five categories, as measured by Alexa Research (www.alexa.com). From 40 major search engines, the five search engines we selected for this study are: Excite (Meta search engine), Froogle (Ecommerce engine), Google (General engine), Overture (paid for inclusion engine), and Yahoo! Directory (Directory service engine). Table 1 displayed the descriptive information about each search engine.

The link popularity is the number of links on the Web pointing to that search engine as reported by Web Position Gold (FirstPlace Software, 2004) using the Web search engine, Lycos. The higher the number, the more popular the engine is in terms of links.

The Alexa.com ranking is an indicator of the popularity of an engine. Alexa Research (www.alexa.com) ranks Web sites starting at 1 for the most popular site on the Web. Therefore, the lower the Alexa number, the more popular the Web site domain, according to Alexa's estimate. As with link popularity, the Alexa rating does not account for the total picture of a search engine's visibility. For example, many major sites and

Table 1
Search engines used in study

Search engine[a]	Category[b]	Engines within category	Link popularity[c]	Alexa rating[d]	URL of submission
Excite	Meta-crawler	6	14,216	83	http://msyml.excite/com/info.sclck
Froogle	Shopping	1	35,325	5	http://froogle/google.com
Google	General purpose	18	9,468,625	3	http://www.google.com
Overture	Pay-for-inclusion	9	561,565	49	http://overture.com/d/home
Yahoo! Directories	Directory service	6	2,424,060	1	http://dir.yahoo.com
		40			

[a] Total of 40 search engines from all categories.
[b] All rankings as of 30 September 2004.
[c] Higher is better.
[d] Lower is better.

search engines on the Web serve Google's results; therefore, Google may be more popular than its link popularity or Alexa rating may indicate (FirstPlace Software, 2004).

However, while link popularity and Alexa ranking can be useful indicators to tell how popular one search engine is relative to another, they are not necessarily reflective of how much real traffic that search engine receives. However, they are reasonable indicators of how popular a search engine is and thereby how much traffic that search engine can direct to an ecommerce Web site (FirstPlace Software, 2004). The rankings are in line with survey data reported by online research companies (Nielsen Netrating, 2002; Sullivan, 2004).

The following is a description of the search engines selected:

Excite (http://www.excite.com)—Meta search engine—Excite retrieves the most relevant links from Google, Yahoo!, Ask Jeeves, About, Overture, and AltaVista and combines those links into a single list of what it algorithmically determines are the most relevant links. Users can then choose to rank the returned links by either overall "Relevancy" or by "Search Engine" (Goodman, 2002; The Excite Network, 2004).

Froogle (http://www.froogle.google.com)—Ecommerce search engine—is an ecommerce-focused engine that is operated by parent company, Google. It uses Google's PageRank algorithm (Brin, 1998) but to a lesser degree. Froogle claims to take advantage of technology designed to automatically scour the Internet for products on sale, and to rely on direct data feed from merchants in order to correctly catalog relevant search links (Froogle Merchants, 2004). Froogle also allows the user to search by price and price range (Subia, 2004).

Google (http://www.google.com)—General search engine—was originally founded as a Stanford University project by students Larry Page and Sergey Brin (1998). Google now operates as an independent company and is the world's most popular search engine, processing over 200,000,000 queries daily (Google, 2004).

Overture (http://www.overture.com)—Paid for inclusion search engine—was purchased by Yahoo! (Berkowitz, 2003). Overture utilizes a "Paid-for-Performance" method that allows businesses to essentially outbid each other in order to get a higher ranking when their keyword phrase is searched. These businesses fund accounts with Overture and are then charged for each person that clicks through from the search results to the businesses' Web sites (Marketing Tops, 2003).

Yahoo! Directories (http://dir.yahoo.com)—Directory Service Engine—is a human-compiled search service that returns both directory category links (called "Related Categories") and Directory Results, which are the top Website matches drawn from all categories of the Yahoo! Directories. Sites pay a fee to be included in the Yahoo Directory's commercial listings, though they must meet editor approval before being accepted. Non-commercial content is accepted for free (Sullivan, 2004).

4.2. Ecommerce queries

We next selected the ecommerce queries that we would use in the study. The queries selected originally came from an Excite transaction log that consisted of about 1 million queries. Note that the Excite search engine that provided the transaction log is not the Excite meta-search engine that we used in this research; therefore, the transaction log is unduly biased toward any of the search engines. Using commercial key terms (*buy*, *price*,

sale, *purchase*, etc.) as original seed terms and a modified snowball technique (Patton, 1990), we extracted from this large set of queries a focused ecommerce set of queries that expressed a commerce information need. From this list, the researchers selected 100 queries to use in the study that represented a broad range of ecommerce queries (i.e., we did not focus only on a single domain such as airlines, retailers, real estate, etc.). Examples include: "houses for sale in Lews county"; "region free DVD software"; "low AND cost AND flights AND Glasgow"; "1969 Camaro Convertible for sale in Florida"; "I want to buy things with my frequent flier miles", and "baseball card price guide". Appendix A contains the full list of 100 queries.

4.3. Ecommerce Web links

After the 100 ecommerce queries were selected, we used the WebPosition Gold software application to submit the 100 queries to the five search engines and to retrieve the links. Because studies show that approximately 80% of Web searchers never view more than the first 10 links in a results list (Jansen et al., 2000; Silverstein, Henzinger, Marais, & Moricz, 1999), we retrieved the first SERP for each query on each search engine. The SERP typically included 10 organic links and a number of sponsored links.

The entire process of submission and retrieval of results for four of the search engines took less than 5 min, with the entire processing taking approximate 54 min, primarily for Yahoo! Directories. This was because that Yahoo! Directories places a limit on the number of queries it will process from a single Internet Protocol (IP) address within a 30 min period.

This short period of retrieval ensured that the list was a consistent snapshot among all 5 search engines, which lessened the chance of experimental bias due to changes in search engine database algorithms that could have occurred with a list retrieved during a longer duration of time (Selberg & Etzioni, 2000). Table 2 shows the exact length of times that each engine took to retrieve the results of all 100 queries. Note that the periods run concurrently (i.e., the submission process for each search engine began at the same time).

The total number of search results returned was 3540. We imported all of the search results into a spreadsheet program, aligning the results into records containing the following categorical information in an easy-to-read format: Query, Search Engine, Rank, Title, Description, Uniform Resource Locator (URL), and Type of Result (whether or not the search result returned was a sponsored result). We also assigned each record a unique identifier. Fig. 2 shows the layout of the processed results.

4.4. Preparation for user evaluation

We then imported the processed data in a relational database to prepare the links for the user evaluation. Within the database, we removed the Search Engine, Sponsored, and Rank fields. We did not want the knowledge of the search engine or the result rank to bias the evaluator's rating of the result. With our unique identifier, we could later repopulate this data.

We also removed all duplicate links for each query, as we did not want the evaluators to review the same URL more than once for a particular query. For example, if all five search engines retrieved the same URL for the same query, four of the URL occurrences were removed. Similarly, if the same search engine retrieved a URL multiple times for the same query, the duplicate URLs were removed. Once we removed all duplicates, we had 3221 unique records containing the fields of Query, Title, Summary, and URL.

Table 2
Processing time for query submission and result retrieval

Search engine	Duration of process
Excite	4 min 28 s
Froogle	4 min 28 s
Google	5 min 39 s
Overture	4 min 30 s
Yahoo! Directories	53 min 42 s

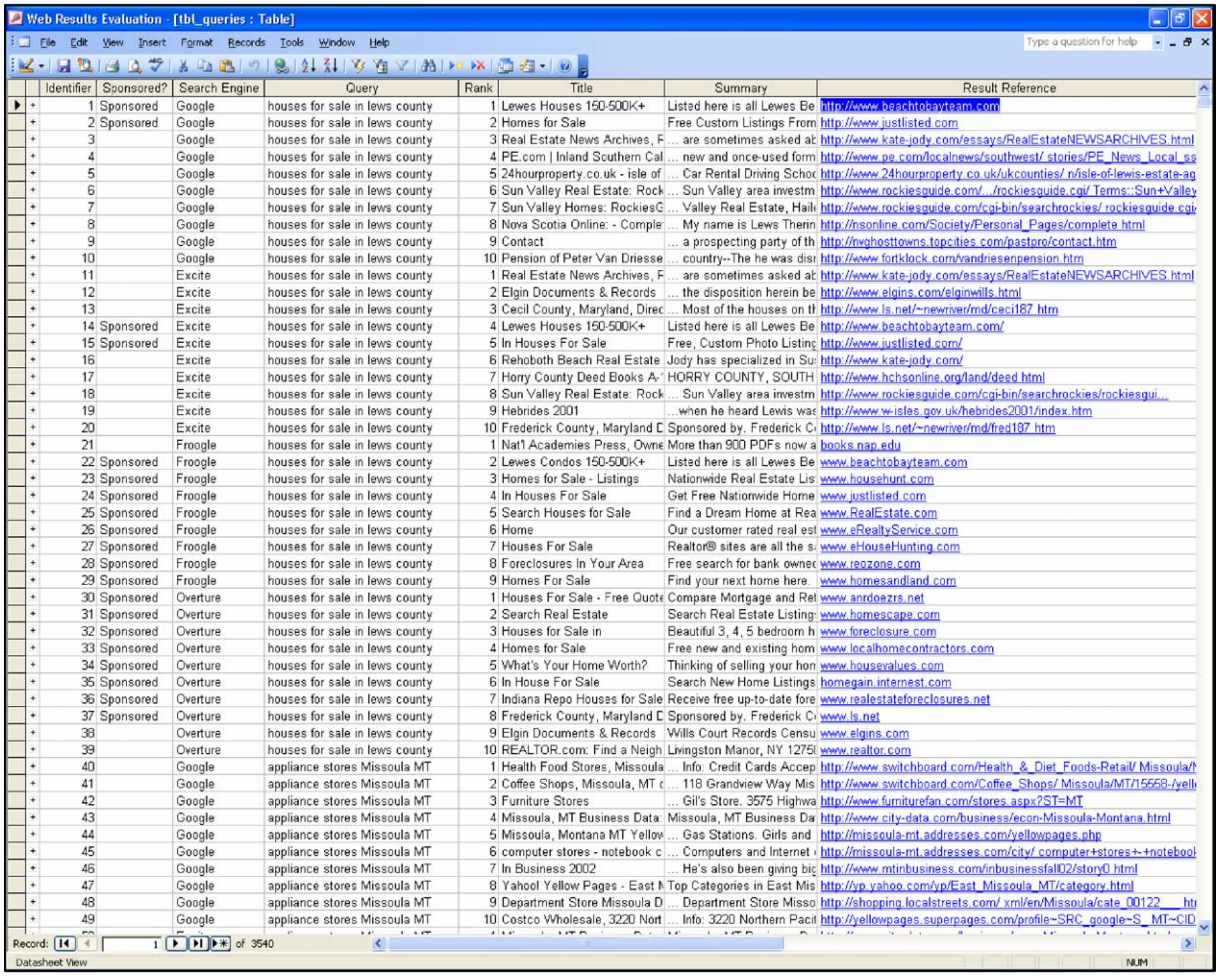

Fig. 2. Layout of Web search links after processing.

We then designed an interface to present the links to each evaluator one result at a time. Fig. 3 shows the application interface used by the evaluators.

As Fig. 3 shows, the application presented each result as a "stand alone" (i.e., not in a ranked result list) with the corresponding query. All URLs from a particular query were presented together, followed by the next query, etc. On start up, the application form would open and a dialog box would appear (see Fig. 3), setting the scenario for the session. The dialog box informed the evaluators that they had just entered the query in a search engine and were now evaluating the result returned. The evaluators were to base their evaluation on their interpretation of the query and the scenario.

Once the evaluators, clicked "OK", the evaluation form was now in front of them. The form displayed the Query, along with the result's Title, Summary, and URL. The form also had fields for the evaluators to enter their rating of the result (1 for relevant, 2 for somewhat relevant, and 3 for not relevant).

We desired to know the basis for the participants' evaluation. We selected a stratified sample of links, based on average relevancy scores, for our participants to determine the basis for their evaluation. To determine the number of ecommerce links for a representative sample, we calculated the sample size in links as: $ss = (z^2\sigma^2)/\beta^2$, where z is the confidence level, σ is the sample standard deviation, and β is the error rate. Setting the confidence level to 95% and the error rate to 5%, we required a sample of 15 links. We selected five links from each set of links where all participants agreed on the evaluation (i.e., average evaluation was 1, 2, and 3).

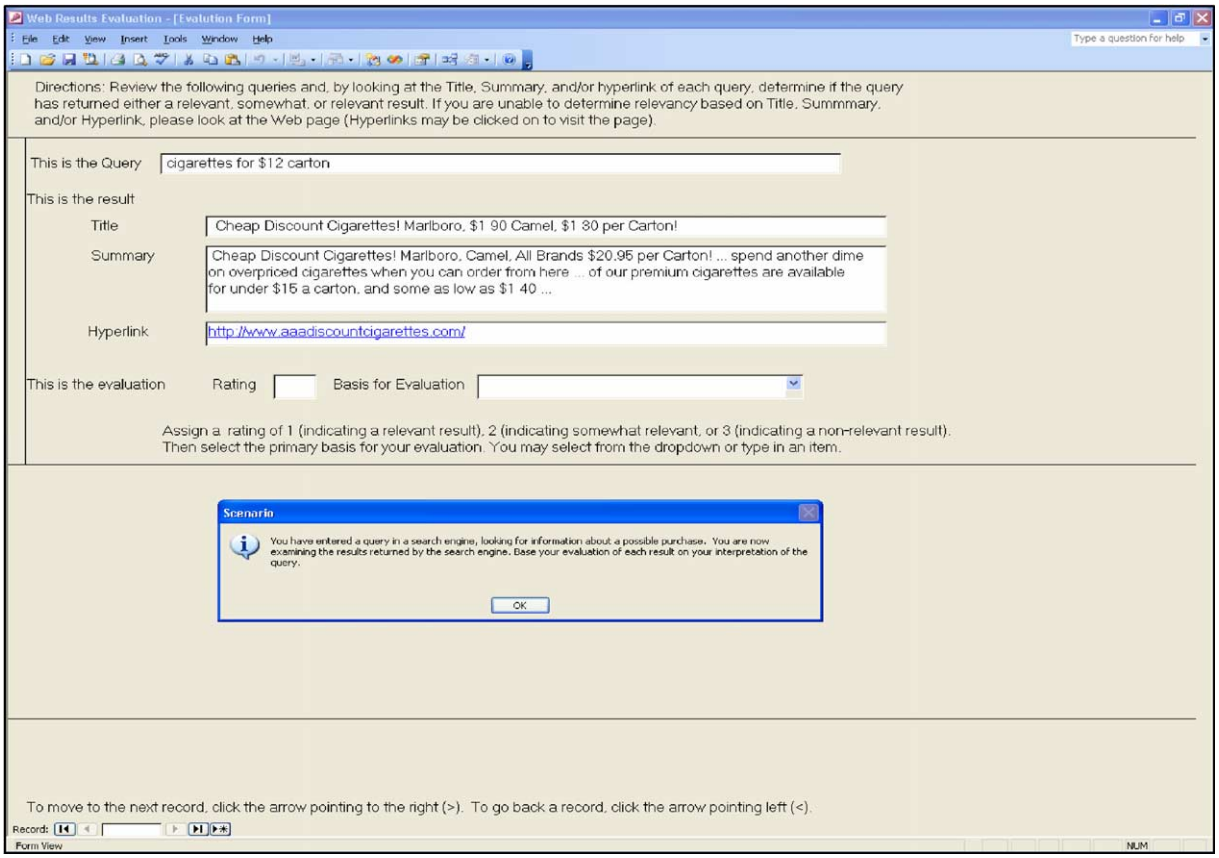

Fig. 3. Application interface used in evaluation of ecommerce links.

4.5. Evaluator judgment of links

Four evaluators participated in this study. The evaluations were not performed at a specified location under any controlled conditions. We presented the evaluators with their materials (Pre-surveys, Instructions, and the Evaluation Software Application) individually at different times and different locations. Each evaluator could complete the evaluations at an individual pace. We did this to ensure that the test reflected real-world conditions and to diminish the effects of fatigue on the users, given that they were being presented with more than 3200 search links to evaluate. We also desired the evaluators to make independent evaluations of the links.

All of the evaluators were traditional college students (i.e., 20–23 years old) and experienced using both computers and the Internet. Before beginning the evaluation, we gave each participant a pre-test survey to verify name, age, gender, as well as Internet and search engine experience. Table 3 shows the aggregate data retrieved from this survey.

Table 3
Pre-survey results for evaluators

Items	Participant 1	Participant 2	Participant 3	Participant 4
Age	21	23	22	20
Gender	M	M	M	F
Internet experience (years)	5 or more	3–5	5 or more	5 or more
Engine user most frequently	Google	Yahoo!	Google	Google
Engine used for ecommerce	Google	Yahoo!	Google	Google

After administering the pre-survey, we instructed the participants on how to complete the study. The instructions indicated to the evaluators that their objective was to judge Web search engine links in terms of their relevancy to the original ecommerce query. We also briefed the evaluators on the following scenario (which also appeared whenever the evaluators opened the software application):

You have entered a query in a search engine, looking for information on a possible purchase. You are now examining the results returned by the search engine. Base your evaluation of the result on your interpretation of the query.

We stressed to each participant the primary task was to assign a rating on a scale of 1, 2, or 3 based on whether, for the given query, the result is relevant (1), somewhat relevant (2), or non-relevant (3). The instructions stated the need for users to rate each result by looking at its Title, Summary, Hyperlink, or visiting the Web page. The evaluators could visit a result's corresponding Web page by simply clicking on its displayed hyperlink.

Due to the sheer number of Web search links included in the study, we told participants to take their time evaluating each search result. Participants did not have a deadline for completing the study and could come and go from the study if experiencing fatigue. We explicitly gave these latter instructions to the participants to make certain that each user carefully evaluated each individual search result.

After all evaluators had judged all links, we asked all of the users to perform an additional evaluation on our subset of queries to determine the basis for their evaluation. We randomly selected 15 queries from those queries that had received the same relevancy rating by each of the participants. Five queries selected had received a rating of 1. Five had received a rating of 2, and five had received a rating of 3. We chose the queries in this manner in order to assess any differences that could exist in the reasoning behind users' levels of evaluation. We showed evaluators each of the chosen search links with the corresponding rating that the user had given that result. We then asked the evaluator to give a basis for the ratings. Evaluators chose from the following reasons: Page Content, Load Time, Professionalism of Page, Summary, Title, Trust, and URL, or other, noted by typing in another basis.

When the four users had evaluated all of the search links, we combined and averaged the ratings for each of the search links. Inter-rater reliability calculated using Cronbach's α was 0.89, which indicates good correlation. For each record, we re-integrated the Search Engine, Rank, and Sponsored status. For the non-unique links, we reintroduced these records and automatically assigned the corresponding average evaluation. Once we had completed this, we had an evaluation between 1 and 3 for each of the 3540 links. We exported this tabulation from our database to a spreadsheet and then imported the data into SPSS 12.0 for the statistical analysis.

5. Results

We now address our three research questions: (1) Which type of search engine performs better for ecommerce searching? (2) Are the top ranked links the most relevant for ecommerce searching? and (3) Are sponsored links more relevant than organic links for ecommerce searching?

5.1. Comparison among search engines

We first examine if there is any difference in performance among the five types of search engines for ecommerce queries. Our hypothesis is: *There will be a significant difference in the number of relevant links retrieved using an ecommerce search engine relative to non-ecommerce search engines.*

In order to evaluate our hypothesis, we performed a three-step statistical evaluation to determine if there is a difference of means (relevancy means) among the five groups (search engines) tested.

The first step in the evaluation was to complete a one-way ANOVA statistical analysis to compare means and variance among the groups. The analysis tests the null hypothesis that group means do or do not differ. In the test, we are interested in the significant F value obtained at the experiment's critical value (α) of 0.05.

The results shown in Table 4 indicate that there is a significant difference among the groups ($F(4) = 25.078$, $p < 0.01$; the critical value of $F = 2.372$). This indicates significant effects among the group, and the group means differ more than would be expected by chance (experimental error) alone. However, this analysis only

Table 4
ANOVA—average relevancy comparisons among search engine types

	N	Mean	Std. deviation	Std. error	95% Confidence interval for mean	
					Lower bound	Upper bound
Excite	977	2.07	.65	.02	2.03	2.12
Froogle	719	2.34	.59	.02	2.30	2.39
Google	899	2.11	.62	.02	2.06	2.15
Overture	848	2.12	.68	.02	2.08	2.17
Yahoo! Directories	97	2.39	.54	.06	2.28	2.50
Total	3540	2.16	.64	.01	2.14	2.18

	Sum of squares	df	Mean square	*F*	Sig.
Between groups	**40.4**	**4**	**10.1**	**25.08**	**.000**
Within groups	1423.62	3535	.40		
Total	1464.02	3539			

tells us that there is a difference among the groups; it does not specify *which* groups. For that, we ran a Tukey's HSD (''honestly significantly different'' test).

Tukey's HSD post hoc test compares each control group to the other groups. The asterisks in the mean difference column in Table 5 identify the paired groups that show statistical difference at the 0.05 confidence level or lower. Looking at Froogle, we see that it is significantly different than all other engines except for Yahoo! Directories.

This is only indicative of a difference in the relevancy means among specified groups, not the actual means. In order to determine that, a secondary HSD analysis is performed in which the group means are evaluated per sub-sets. With the critical value still set at 0.05, the subset analysis, shown in Table 6, indicates that the higher evaluations are prevalent for Yahoo! Directory (mean = 2.4) and Froogle (mean = 2.3).

Table 5
Tukey's HSD—multiple comparisons

Search engine (*I*)	Search engine (*J*)	Mean difference (*I* − *J*)	Std. error	Sig.	95% Confidence interval	
					Lower bound	Upper bound
Excite	**Froogle**	**−.27***	**.03**	**.00**	**−.35**	**−.18**
	Google	−.03	.03	.83	−.11	.05
	Overture	−.04	.03	.49	−.13	.03
	Yahoo! Directories	**−.32***	**.07**	**.00**	**−.50**	**−.13**
Froogle	**Excite**	**.27***	**.03**	**.00**	**.18**	**.35**
	Google	**.24***	**.03**	**.00**	**.15**	**.33**
	Overture	**.22***	**.03**	**.00**	**.13**	**.31**
	Yahoo! Directories	−.05	.07	.96	−.24	.14
Google	Excite	.03	.03	.84	−.05	.11
	Froogle	**−.24***	**.03**	**.00**	**−.33**	**−.15**
	Overture	−.02	.03	.98	−.10	.07
	Yahoo! Directories	**−.29***	**.07**	**.00**	**−.47**	**−.10**
Overture	Excite	.05	.03	.49	−.03	.13
	Froogle	**−.22***	**.03**	**.00**	**−.31**	**−.13**
	Google	.02	.03	.98	−.07	.10
	Yahoo! Directories	**−.27***	**.07**	**.00**	**−.45**	**−.08**
Yahoo! Directories	**Excite**	**.32***	**.07**	**.00**	**.13**	**.50**
	Froogle	.05	.07	.96	−.14	.24
	Google	**.29***	**.07**	**.00**	**.10**	**.47**
	Overture	**.27***	**.08**	**.00**	**.08**	**.45**

* The mean difference is significant at the .05 level.

Table 6
Tukey's HSD—subset comparison of average evaluation

Search engine	*N*	Subset for $\alpha = .05$	
		1	2
Excite	977	2.0747	
Google	899	2.1051	
Overture	848	2.1226	
Froogle	719		2.3435
Yahoo! Directories	97		2.3918
Sig.		.867	.864

Note that *n* (the number of links for each search engine) differs, as each search engine sometimes returned less than 10 organic links for a particular query and a varying number of sponsored links. We did not penalize a search engine for not returning links to a number of the queries we tested. However, Yahoo! Directory only has 97 retrieved links while Froogle has the same statistically significant mean and a substantially higher number of retrieved links at 719. There is no statistical difference in either the mean or the retrieved links number among Excite (mean of 2.1, $n = 977$), Overture (2.1, $n = 848$), or Google (2.1, $n = 899$).

Therefore, Hypothesis 1 is partially supported. An ecommerce search engine performs better than a meta, general purpose, and pay-for-placement search engines but not better than a directory service for ecommerce queries.

In regard to the basis for evaluation, the results, shown in Table 7, indicate that the users based the majority of their relevancy ratings on the Title of the search result. Overall, Title was the basis for user relevancy rating for 85% of the search links evaluated.

We observe that the percentage of search links rated on a basis of Title increases with an increase in the relevancy ranking given by the user. For example, 90% of the search links given a rating of 1 (Relevant) by the users had Title as the basis of evaluation. This is in comparison to 85% of the search links given a rating of 2 (Somewhat relevant) and 80% of the search links given a rating of 1 (Non-relevant). We note that Summary and Trust are not deemed significant bases for rating; they accounted for only 13% and 1%, respectively, of the bases for rating for the search links evaluated.

5.2. Differences by rank

Our second research question investigates whether the top-ranked links are the most relevant for ecommerce searching. Our hypothesis is: *Relevance of links will decrease as the rank of links increases* (*i.e.*, *the more relevant links have the lowest rank*).

Table 7
Basis for evaluation of links

No.	Query	Ranking	Eval. 1	Eval. 2	Eval. 3	Eval. 4
1	"Beatles CDs for sale "Let it Be"	1	Title	Title	Title	Title
2	56k v9 pci modem driver free download	1	Title	Title	Title	Title
3	I want to buy things with my frequent flier miles	1	Title	Title	Summary	Title
4	Land for sale in Minnesota	1	Title	Title	Summary	Title
5	Region free DVD software	1	Title	Title	Title	Title
6	Cigarettes for $12 carton	2	Title	Title	Summary	Title
7	"Gibbs free energy" +review	2	Title	Title	Title	Title
8	Bauxite prices	2	Title	Summary	Trust	Title
9	Gothic boots online stores	2	Title	Title	Title	Title
10	Snakes for sale: colubrids	2	Title	Title	Title	Title
11	7th edition singles for sale magic the gathering	3	Title	Summary	Title	Title
12	Best AND price AND Xerox AND M950 AND M940	3	Title	Title	Title	Summary
13	Here to buy Essem kielbasa	3	Title	Title	Title	Summary
14	Houses for sale in Lews county	3	Title	Title	Summary	Title
15	Unlimited long distance $25	3	Title	Title	Title	Title

Table 8
Cross-tab evaluation of rank and average relevancy evaluation

Rank	Average evaluation									Total
	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	
1	67	17	5	16	155	20	3	17	79	379
2	57	17	1	11	149	21	3	28	87	374
3	50	18	0	11	159	20	2	26	83	369
4	43	13	3	9	165	19	2	20	86	360
5	44	11	0	11	152	19	0	25	96	358
6	35	16	0	12	152	17	5	24	87	348
7	44	9	1	10	145	20	2	24	90	345
8	33	15	0	11	148	19	0	24	92	342
9	30	8	1	11	144	22	3	27	92	338
10	33	6	1	6	135	27	7	27	85	327
Total	436	130	12	108	1504	204	27	242	877	3540
%	12.3	3.7	0.3	3.1	42.5	5.8	0.8	6.8	24.8	100.0

For the statistical analysis, we were originally planning to use regression analysis. However, there was no correlation between rank and average relevancy evaluation, which is the required assumption for the use of regression analysis.

Therefore, we conducted a cross-tabs analysis, which examines the relationship between two categorical variables to test for independence and measures of association and agreement for nominal and ordinal data. Table 8 shows the number of links at each average relevance evaluation and rank.

From Table 8, we see that there are 436 (12.3%) relevant documents which all four evaluators agreed. There were 877 (24.8%) documents which all four evaluators agreed were not relevant. From observation, there does appear to be a slight relationship between rank and relevancy evaluation, with 67 relevant documents at rank 1 and only 33 *relevant* documents at rank 10. This trend appears also at the *not relevant* level (79 documents at rank 1 and 85 at rank 10). We note also that the rank trend holds at the somewhat relevant evaluation, with 155 *somewhat relevant* documents at rank 1 and 135 at rank 10. However, the trend was not statistically significant.

Therefore, we reject Hypothesis 2 as there is no statistical difference in average relevance by rank for any search engine.

5.3. Difference between organic and sponsored links

For research question 3, we are interested in determining if sponsored links are more relevant than organic links for ecommerce searching. Our hypothesis is: *Sponsored links will be more relevant than organic links.*

For this evaluation, there is an issue on how to code the links from Overture, which is a pay-for-inclusion search engine. Since it is a pay-for-inclusion engine, one could say that all of these links should be coded as sponsored links. We address this issue by evaluating the hypothesis both ways; first coding the Overture links as identified from the search engine as either organic or sponsored. We then analyze the links as all sponsored.

There were 3079 organic links with an average relevant evaluation of 2.17 and a standard deviation of 0.65. There are 461 sponsored links with an average relevant evaluation of 2.07 with a standard deviation of 0.62. Yahoo! Directories had no links identified as sponsored links. The results of an ANOVA analysis indicate that there is a significant difference between the groups ($F(1) = 10.32$, $p < 0.0$). Mean relevance evaluation for organic links (mean = 2.2) is statistically higher than the mean evaluation for sponsored links (mean = 2.1).

We recalculated the statistical analysis coding each of the Overture links as sponsored. In this scenario, the organic links had an average mean relevancy evaluation of 2.2 with a standard deviation of 0.63. The sponsored links have an average mean relevancy evaluation of 2.1 with a standard deviation of 0.66. We conducted a two-way ANOVA to determine if there was any difference between organic and sponsored links by search engine, with descriptive results displayed in Table 9.

Table 9
Descriptive statistics of sponsored and organic links by search engine

Sponsored	Search engine	Mean	Std. deviation	N
Organic	Excite	2.0819	.65609	736
Organic	Froogle	2.3520	.59107	642
Organic	Google	2.1146	.62577	833
Organic	Overture	2.1368	.67786	771
Organic	Yahoo! Directories	2.3918	.54236	97
Organic	Total	2.1706	.64571	3079
Sponsored	Excite	2.0529	.65236	241
Sponsored	Froogle	2.2727	.53061	77
Sponsored	Google	1.9848	.51865	66
Sponsored	Overture	1.9805	.63965	77
Sponsored	Total	2.0678	.61933	461
Total	Excite	2.0747	.65495	977
Total	Froogle	2.3435	.58507	719
Total	Google	2.1051	.61921	899
Total	Overture	2.1226	.67562	848
Total	Yahoo! Directories	2.3918	.54236	97
Total	Total	2.1572	.64318	3540

The two-way ANOVA is a factorial method that allows us to study the interaction among the independent variables of type of result and search engine and their effect on average relevancy evaluation of the results. The results indicate that there was again a significant difference among the groups ($F(7) = 14.91$, $p < 0.01$). Mean relevance evaluation for organic links (mean = 2.2) is statistically higher than the mean evaluation for sponsored links (mean = 2.1). From Table 9, it appears that we see a decrease in average relevancy evaluations between organic and sponsored links among search engines. However, the results of a two-way ANOVA showed no significant interaction with Type of Links * Search Engine on average relevance evaluations.

Therefore, we reject Hypothesis 3. Sponsored links are not more relevant than organic links. In fact, the reverse is true.

5.4. An issue with duplicate links

Several of the search engines retrieved duplicate links with the first results query set. Table 10 displays the number of links from each search engine and the number of duplicate links.

In our initial analysis, we included these duplicate links. If these duplicate links were relevant, it could unduly bias the results by favoring a search engine that returned relevant links to ecommerce queries but that did not eliminate duplicates (i.e., these duplicate links would probably be of little value to the actual searcher). We conducted a one-way ANOVA statistical analysis to compare means and variance among the search engines with the duplicate links for each search engine removed. Table 11 displays the results of the analysis.

The results shown in Table 11 indicate that there is a significant difference among the groups ($F(4) = 25.42$, $p < 0.01$; the critical value of $F = 2.372$), just as there was when the duplicates were included. This indicates

Table 10
Number of Web search engine links and number of duplicate links

Search engine	Number of links	Number of duplicate links
Excite	977	26 (2.7%)
Google	899	115 (12.8%)
Overture	848	5 (0.6%)
Froogle	719	104 (14.5%)
Yahoo Directory	97	0 (0.0%)
	3540	250 (7.1%)

Table 11
ANOVA—group comparison

	Sum of squares	df	Mean square	*F*	Sig.
Between groups	**41.31**	**4**	**10.33**	**25.42**	**.000**
Within groups	1331.29	3276	.41		
Total	1372.60	3280			

significant effects among the group. The group means differ more than would be expected by chance (experimental error) alone. In fact, the difference is now stronger than when the duplicates were included.

We conducted a Tukey's HSD analysis to inform us when search engine mean relevance evaluations were significantly different, as shown in Table 12.

The asterisks in Table 12 identify the paired groups that show statistical difference at the 0.05 confidence level. Taking a look at Froogle, we see that it is again significantly different than all other engines except for Yahoo! Directories. A secondary HSD analysis, results shown in Table 13, indicates that the higher evaluations are prevalent for Yahoo! Directory (mean = 2.39) and Froogle (mean = 2.37).

Table 12
Tukey's HSD—multiple comparisons

Search engine (*I*)	Search engine (*J*)	Mean difference (*I* − *J*)	Std. error	Sig.	95% Confidence interval	
					Lower bound	Upper bound
Excite	**Froogle**	**−.30***	**.03**	**.00**	**−.39**	**−.20**
	Google	−.03	.03	.82	−.11	.05
	Overture	−.06	.03	.34	−.14	.03
	Yahoo! Directories	**−.32***	**.07**	**.00**	**−.50**	**−.13**
Froogle	**Excite**	**.29***	**.03**	**.00**	**.20**	**.39**
	Google	**.26***	**.03**	**.00**	**.17**	**.35**
	Overture	**.24***	**.04**	**.00**	**.14**	**.3320**
	Yahoo! Directories	−.02	.07	1.0	−.21	.17
Google	Excite	.03	.03	.82	−.05	.11
	Froogle	**−.26***	**.03**	**.00**	**−.35**	**−.17**
	Overture	−.03	.03	.92	−.11	.06
	Yahoo! Directories	**−.28***	**.07**	**.00**	**−.47**	**−1.0**
Overture	Excite	.06	.03	.34	−.03	.14
	Froogle	**−.24***	**.04**	**.00**	**−.33**	**−.14**
	Google	.03	.03	.92	−.06	.11
	Yahoo! Directories	−.26*	.07	.00	−.45	−.07
Yahoo! Directories	**Excite**	**.32***	**.07**	**.00**	**.13**	**.50**
	Froogle	.02	.07	1.0	−.17	.21
	Google	**.28***	**.07**	**.00**	**1.0**	**.47**
	Overture	**.26***	**.07**	**.00**	**.07**	**.45**

* The mean difference is significant at the .05 level.

Table 13
Tukey's HSD subset comparison

Search engine	*N*	Subset for $\alpha = .05$	
		1	2
Excite	950	2.08	
Google	893	2.11	
Overture	744	2.13	
Froogle	597		2.37
Yahoo! Directories	97		2.39
Sig.		.78	.99

The mean relevant evaluation scores are significantly higher for Froogle and Yahoo! Directories than to Excite, Google, and Overture. The mean evaluation for Yahoo! Directories is the same since it had no duplicates. Froogle's mean relevancy evaluation increased, implying many of its duplicates were *not* or *somewhat* relevant.

6. Discussion

We now discuss the implications of our results. Our evaluation found a statistical difference in the relevancy of the links retrieved among the five search engine types. Two of the specified search engines, Froogle and Yahoo! Directories, had statistically higher mean relevance than the other three search engines, Excite, Google, and Overture. No statistical difference was exposed among the relevancy means for Excite, Google, and Overture. There was also no statistical difference between Froogle and Yahoo! Directories.

In response to our first research question (i.e., *Which type of search engine performs better for ecommerce searching*?), it appears that an ecommerce search engine and a search directory best support ecommerce searching, relative to meta, general purpose, or pay-per-click search engines. Yahoo! Directories retrieved substantially fewer links than Froogle, however.

Although the results are statistically significant, one must always ask what the practical significance is. Basically, with mean relevance differences of about 0.25 of a result, in a group of 40 links, Froogle and Yahoo! Directories would provide the searcher one more relevant result than would Excite, Google, or Overture. This may not be a substantial increase for only one search. However, with people doing numerous searches a day, the aggregate performance increase over time may be considerable.

Concerning our second research question (i.e., *Are the top ranked links the most relevant for ecommerce searching*?), there appears to be no statistical difference in average relevance among ranks for any of the search engines. This is somewhat surprising, since one would expect some decrease in relevance as the rank increases; however, this research examined only the first SERP. The granularity of the search engine ranking algorithms may not be so fine. As more SERPs are included, one may begin to see a difference in relevance by rank.

Investigating our third research question (i.e., *Are sponsored links more relevant than organic links for ecommerce searching*?), our analysis showed that sponsored links were not more relevant than organic links. Surprisingly, organic links were more relevant than sponsored.

The findings have implications for Web search engines supporting ecommerce searching and those desiring to sell on the Web. First, it appears that ecommerce search engines (such as Froogle) or search directories (such as Yahoo! Directories) that provide targeted information content are helpful to ecommerce searchers. Major general purpose-search engines should "spin off" ecommerce niche search engines to support these Web searchers.

Second, it appears that there is little correlation among relevant links and result rank within the first SERP. Given this, search engines may want to investigate techniques such as clustering to provide searchers more choices on the first SERP page. Some search engines, such as Vivisimo (http://vivisimo.com/), already offer clustering.

Finally, it appears that organic links are more relevant to ecommerce searching than sponsored links. These results indicate that those bidding on keywords and phrases must more carefully refine their marketing campaigns to potential customers. Even with sponsored links, those desiring to sell products or services online should not ignore organic links. A dual strategy of pay-per-placement and search engine optimization may be the most beneficial market approach.

Given these findings, we believe that this study provides important insight into the effectiveness of five major types of search engines and their support of ecommerce searching. The strengths of the research are that we used five very popular search engines, ensuring that the results would have a broad impact. We used a large number of ecommerce queries from a Web search engine transaction log to ensure that the queries represented the actual type and range of ecommerce searching conducted by Web users. Finally, we employed a rigorous methodology (i.e., masked the underlying search engine, eliminated duplicate URLs, aggregated the links by query, presented the links individually to eliminate the ranking effect) to limit the possibility of bias in the evaluation.

Naturally, the study has limitations. First, the 100 ecommerce queries may not represent the complete range of ecommerce searching. The testing of a larger number of queries in the future should increase the confidence of the findings from this study. Second, the structure of the queries was mixed (i.e., plain text, Boolean, natural language), and certain search engines may favor certain types of queries. However, our take was that the queries are actual queries submitted by real users from the Web populations. Therefore, it is reasonable to assume that Web search engines servicing this population should handle these types of queries.

Third, the evaluators were not the originators of the queries and had to interpret the queries and scenarios. How to measure or even define relevance is a debated topic (Saracevic, 1975). Numerous researchers have written on the relationship between relevance and task (cf., Vakkari, 2000) and user (cf., Barry, 1994; Barry & Schamber, 1998; Greisdorf, 2003; Schamber, 1994; Wang & Soergel, 1998). Others have investigated the use of evaluators instead of actual users (cf., Chowdhury & Soboroff, 2002; Gordon & Pathak, 1999). So, certainly not having the actual users is a limitation. However, inter-rater agreement among evaluators was quite high (0.89), so this does not appear to be a major issue for this set of queries.

7. Conclusions and future research

We began with a review of the relationship between the growth of ecommerce and the continuing development of search engines. We then proceeded to examine some of the research studies that pertain to search engine evaluation and ecommerce Web searching. We found that few of these studies have investigated the effectiveness of search engines in the retrieval of relevant ecommerce results, and none have compared the success of an ecommerce-focused engine to that of other engines. We thus designed a study to test and evaluate the relevancy of search results retrieved from various engines, including an engine built specifically for ecommerce searches.

For future research, we are expanding on our post-test evaluation. The evaluation involved a limited number of queries and asked the participants a straightforward question: On what did you primarily base your rankings? All four of the participants overwhelmingly chose "Title" as their chief reason for the majority of their ratings. The post-test evaluation was limited in its scope, but it exposes an entirely new dimension to the study of search engine effectiveness. For ecommerce searching, it poses an attractive opportunity to investigate ecommerce search results from the user's perspective by further evaluating the reasoning behind their choosing of some links as more relevant than others.

The current study was limited in its interactivity. The links were taken from the first SERP, and we asked the evaluators to evaluate these pages. However, Web searchers sometimes follow links from these SERP to other pages, which sometimes take them to relevant results. We are designing a study to address this mode of Web searching.

Overall, this study supports the ongoing research into ecommerce searching technology. The concept of ecommerce has been around since the mid 1990s (Information Societal Portal, 1997), but it still represents only a fraction of the purchases made worldwide. The use of ecommerce searches is still relatively novel to both Internet users and search engine companies. There is plenty of room for growth. The fact that Froogle is currently (at the time of the study) the only major Web search engine designed for ecommerce searches is a testament to this notion. Research must continue in this area in order to both better understand the means by which users relate to results as well as how an increased number of relevant links can be achieved by all search engines.

Acknowledgements

We thank the four evaluators (Philip DePaola, James Trainor, Samantha Guss, and David Montanez) for their time, effort, and interest in this research. We also thank the three anonymous reviewers for their comments, suggestions, and recommendations, all which contributed to substantially improving the presentation of this research.

Appendix A. Ecommerce queries

1. Houses for sale in Lews county
2. Appliance stores; Missoula, MT
3. Kelly blur book rv prices
4. Where to buy Essem kielbasa
5. What department stores sell hypnotic poison by Christian Dior
6. Event horses for sale in New Jersey
7. Jumping horses for sale in New York
8. Spectra 400 prices
9. Music stores Lafayette Indiana
10. 7th edition singles for sale magic the gathering
11. Days of knights online store DE
12. +"San Francisco" +"real estate" +"for sale by owner"
13. Low price tennis racquets
14. Lowest price korg nxs-5r synthesizer module
15. Whole sale disposable cameras
16. $3.59 cameras
17. Kdur 20 MEQ prices
18. MOTELS FOR SALE JACKSONVILLE
19. For sale Yamaha xt600
20. Land for sale in Minnesota
21. Monmouth county + party stores
22. Homes for sale UK
23. Sailboats for sale + Utah
24. For AND sale AND 1988 AND Toyota AND Celica AND convertible
25. Order +win98 purchase cost price
26. Bauxite prices
27. Food under $5.00
28. 1975 mobile home prices
29. For sale 280 zx ground effects
30. Marshall swift residential cost handbook
31. Unlimited long distance $25
32. Jordan cattle prices
33. Cigarettes for $12 carton
34. v.a. homes for sale in San Diego, CA
35. "128 bit free encryption tools"
36. $500 AND basic AND Web AND presence AND design AND e-com
37. Seacat boat prices
38. 4 cylinder used cars for sale in thunder bay
39. +"used cars for sale in Thunder Bay"
40. New cars priced under $10000.00
41. Homes AND for AND sale AND in AND Queensland, Australia
42. 56k v9 pci modem driver free download
43. For sale ford f350 diesel
44. MITCHELL BROTHERS OFARREL CREDIT CARD
45. CASINO FOR SALE CRIPPLE CREEK COLORADO
46. Textronic printers—free offer
47. Toyota MR2 for sale in Calgary
48. "Castles for sale in England"
49. Beatles CDs for sale "Let it Be"

50. Gothic boots online stores
51. Dealer cost Pontiac
52. NJ puppies free to good home
53. Snakes for sale: colubrids
54. "Canada–Mexico freight prices"
55. Acid free envelopes
56. Nokia 2160 hands free kit
57. Restaurant for sale in Spain
58. Wooden pallets prices
59. Israel AND rj45 AND cable AND meter AND price AND il AND co
60. Land for sale in northern Idaho
61. Time share for Sale Lake Buena Vista
62. Mark mcgwire baseball cards prices 1989, 1991
63. Cars for sale in Washington state under $100.00
64. 55b Cornwall road brampton for sale information
65. "Ethernet 2 credit card" AND drivers AND "windows nt"
66. Easter bunny $2 bill
67. Oleanders for sale in Goliad Co. Texas
68. at&t $30 rebate
69. Volkswagon westfalia camper for sale Canada
70. Sitting Bull Wild West $1 souvenir playing card
71. Epil Stop Spray with Free Robe and life time supply of roll on
72. Girlactik $10.00 girLActik
73. Nice houses under the cost of $20,000 located in Chesapeake near the Hickory Middle School
74. Western digital 40gb retail boxed
75. OMRON AND RS55-50 AND retail AND systems
76. 20$ rolex
77. +"Gibbs free energy" +review
78. Best AND price AND xerox AND M950 AND M940
79. What is the selling price of Petroleos Mexicanos 14.5% ending in 2006?
80. $10 adult vcd Japanese exchange
81. Rollerskates to buy in the UK
82. For sale 1964 ford pickup
83. Dirt bike for sale in Pennsylvania
84. 1000 min for $20.00
85. Yamaha price wx5
86. Oldsmobile transmission prices
87. Region free DVD software
88. Bahamas AND restaurants AND prices AND ratings
89. Low AND cost AND flights AND Glasgow
90. 2001 Nissan Xterra Dealer Invoice
91. POINT OF SALE DEVICE
92. 1969 Camaro Convertible for sale in Florida
93. C AND Class AND Mercedes AND For AND sale AND London
94. Atlantic city free vacation giveaway
95. Acreage for sale Oregon Harrisburg
96. Cheap Go-karts for sale in PA
97. I want to buy things with my frequent flier miles
98. HUD homes for sale in Chino, CA
99. Baseball card price guide
100. Wig stores in New York

References

Awad, E. M. (2004). *Electronic commerce: From vision to fulfillment*. Upper Saddle River, New Jersey: Pearson-Prentice Hall.

Bakos, Y. (1997). Reducing buyer search costs implications for electronic marketplaces. *Management Science, 43*(2), 1676–1692.

Bar-Ilan, J. (1998/1999). Search engine results over time—a case study on search engine stability. *Cybermetrics, 2/3*(1), 1–10.

Bar-Ilan, J. (2000). Evaluating the stability of the search tools HotBot and Snap: a case study. *Online Information Review, 24*(6), 439–449.

Bar-Ilan, J. (2002). Methods for assessing search engine performance over time. *Journal of the American Society for Information Science and Technology, 53*(4), 308–319.

Barry, C. L. (1994). User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science, 45*(3), 149–159.

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: a cross-situational comparison. *Information Processing and Management, 34*(2–3), 219–236.

Bellman, S., Lohse, G. L., & Johnson, E. J. (1999). Predictors of online buying behavior. *Communications of the ACM, 42*(12), 32–38.

Berkowitz, B. (2003). Yahoo! to buy overture in $1.63 billion deal. *Reuter*. Available from http://www.reuters.com/newsArticle.jhtml?type=businessNews&storyID=3085485.

Brin, S. (1998). Extracting patterns and relations from the World Wide Web. In *Proceedings of the World Wide Web and databases*, Valencia, Spain, 27–28 March (pp. 172–183).

Can, F., Nuray, R., & Sevdik, A. B. (2004). Automatic performance evaluation of Web search engines pages. *Information Processing and Management, 40*(3), 495–514.

Chowdhury, A., & Soboroff, I. (2002). Automatic evaluation of World Wide Web search services. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*, Tampere, Finland (pp. 421–422).

Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology. In *Proceedings of the 59th annual meeting of the American Society for Information Science* (pp. 127–135).

Clarke, S., & Willett, P. (1997). Estimating the recall performance of Web search engines. *Aslib Proceedings, 49*(7), 184–189.

Courtois, M. P., & Berry, M. W. (1999). Results ranking in Web search engines. *Online, 23*(3), 39–46.

Ding, W., & Marchionini, G. (1996). A comparative study of Web Search service performance. In *Proceedings of the 59th annual meeting of the American Society for Information Science*, Baltimore, MD (pp. 136–142).

Eastman, C. M. (2002). 30,000 Hits may be better than 300: precision anomalies in Internet searches. *Journal of the American Society for Information Science and Technology, 53*(11), 879–882.

Eastman, C. M., & Jansen, B. J. (2003). Coverage, ranking, and relevance: a study of the impact of query operators on search engine results. *ACM Transactions on Information Systems, 21*(4), 383–411.

FirstPlace Software (2004). WebPosition Gold. Available from http://www.web-positiongold.com/.

ForeSee Results (2002). *ACSI annual E-business report 2002*. ForeSee Results. Retrieved, 1 January, 2005. Available from http://www.foreseeresults.com/.

ForeSee Results (2004). *ACSI annual E-business report 2004*. ForeSee Results. Retrieved, 1 January, 2005. Available from http://www.foreseeresults.com/.

Froogle Merchants (2004). *Froogle information for merchants* [Web site]. Froogle.com. Retrieved, 10 November, 2004. Available from http://www.froogle.google.com/merchants.

Gandal, N. (2001). The dynamics of competition in the Internet search engine market. *International Journal of Industrial Organization, 19*(7), 1103–1117.

Glover, E. J., Lawrence, S., Gordon, M. D., Birmingham, W. P., & Giles, C. L. (2001). Web search—your way: improving Web searching with user preferences. *Communications of the ACM, 44*(12), 97–102.

Goodman, A. (2002, 11 May). *Excite metasearch serves up equal doses of innovation and monetization* [Web site]. Traffick.com. Retrieved, 3 August, 2003. Available from http://www.traffick.com/article.asp?aID=83.

Google (2004). *Google Corporate Information: technology overview* [Web site]. Google.com. Retrieved, 12 October, 2004. Available from http://www.google.com/corporate/tech.html.

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management, 35*(2), 141–180.

Greisdorf, H. (2003). Relevance thresholds: a multi-stage predictive model of how users evaluate information. *Information Processing and Management, 39*(3), 403–423.

Griesbaum, J. (2004). Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. *Information Research: An International Electronic Journal, 9*(4).

Gwizdka, J., & Chignell, M. (1999). *Towards information retrieval measures for evaluation of Web search engines* [Web site]. Retrieved, 4 January, 2005. Available from http://www.imedia.mie.utoronto.ca/~jacekg/pubs/webIR_eval1_99.pdf.

Hawking, D., Craswell, N., Bailey, P., & Griffihs, K. (2001). Measuring search engine quality. *Information Retrieval, 4*(1), 33–59.

Information Societal Portal (1997). *History of ecommerce*. Europe's Information Society. Retrieved, 6 November, 2004. Available from http://europa.eu.int/information_society/index_en.htm.

Jansen, B. J. (2000). An investigation into the use of simple queries on Web IR systems. *Information Research: An Electronic Journal, 6*(1), 1–10.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing and Management, 36*(2), 207–227.

Kehoe, C. M., & Pitkow, J. (1996). Surveying the territory: GVU's five WWW user surveys. *The World Wide Web Journal, 1*(3), 77–84.

Leighton, H. V., & Srivastava, J. (1999). First 20 precision among World Wide Web search services (search engines). *Journal of the American Society for Information Science, 50*(July), 870–881.

Marketing Tops (2003). *How overture works* [Web site]. Internet Marketing Tops, Inc. Retrieved, 12 October, 2004. Available from http://www.work-at-home-profits.com/searchengine/myguide/i46.html.

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines: fluctuations in document accessibility. *Journal of Documentation, 57*(5), 623–651.

Mowshowitz, A., & Kawaguchi, A. (2002). Assessing bias in search engines. *Information Processing and Management, 35*(2), 141–156.

Mukhopadhyay, T., Rajan, U., & Telang, R. (2004). Competition between Internet search engines. In *Proceedings of the 37th annual Hawaii international conference on system sciences (HICSS'04)—track 8*, Big Island, Hawaii, 05–08 January (pp. 80216–80211).

Nicholson, S. (2000). Raising reliability of Web search tool research through replication and chaos theory. *Journal of the American Society for Information Science, 51*(8), 724–729.

Nielsen Netrating (2002). *Top Web properties March 2002* [website]. Nielsen/Netrating. Retrieved, 27 September, 2002. Available from http://www.nielsen-netratings.com.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage Publications, Inc.

Petersen, R. E. (1997). Eight Internet search engines compared. *First Monday, 2*(2).

Rosenbaum, H. (2000). The information environment of electronic commerce: information imperatives for the firm. *Journal of Information Science, 26*(3), 161–171.

Rousseau, R. (1998/1999). Daily time series of common single word searches in AltaVista and Northern Light. *Cybermetrics, 2/3*(1).

Saracevic, T. (1975). Relevance: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society of Information Science, 26*(6), 321–343.

Schamber, L. (1994). Relevance and information behavior. In M. E. Williams (Ed.). *Annual review of information science and technology* (vol. 29, pp. 3–48). ASIS.

Seda, C. (2004). *Search engine advertising: Buying your way to the top to increase sales*. Boston: New Riders.

Selberg, E.,& Etzioni, O. (2000). On the instability of Web search services. In *Proceedings of RIAO 2000: Computer-assisted information retrieval*, Paris, France, April (pp. 223–236).

Shang, Y., & Longzhuang, L. (2002). Precision evaluation of search engines. *World Wide Web, 5*(2), 159–179.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *SIGIR Forum, 33*(1), 6–12.

Singhal, A., & Kaszkiel, M. (2001). A case study in Web search using TREC algorithms. In *Proceedings of the 10th international World Wide Web conference*, Hong Kong, 1–5 May (pp. 708–716).

Soboroff, I., Nicholas, C., & Cahan, P. (2001). Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, New Orleans, LA, USA, 9–13 September (pp. 66–73).

Spink, A. (2002). A user centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing and Management, 38*(3), 410–426.

Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the Web*. New York: Kluwer.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-sex to E-commerce: Web search changes. *IEEE Computer, 35*(3), 107–111.

Spiteri, L. F. (2000). Access to Electronic Commerce Sites on the World Wide Web: an analysis of the effectiveness of six Internet search engines. *Journal of Information Science, 26*(3), 173–183.

Su, L. T. (2003a). A comprehensive and systematic model of user evaluation of Web search engines: I. Theory and background. *Journal of the American Society for Information Science and Technology, 54*(13), 1175–1192.

Su, L. T. (2003b). A comprehensive and systematic model of user evaluation of Web search engines: II. An evaluation by undergraduates. *Journal of the American Society of Information Science and Technology, 54*(13), 1193–1223.

Su, L. T., Chen, H., & Dong, X. (1998). Evaluation of Web-based search engines from the end-user's perspective: a pilot study. In *Proceedings of the 61st annual meeting of the American Society for Information Science*, Pittsburgh, PA, USA, 24–29 October (pp. 348–361).

Subia, T. (2004). *Inclusion methods into Froogle* [Web site]. Subia Search Engine Optimization. Retrieved, 12 November, 2004. Available from http://www.subiainteractive.com/froogle-incl.html.

Sullivan, D. (2003, February 23). *Nielsen/NetRatings search engine ratings* [website]. SearchEngineWatch.com. Retrieved, 21 October, 2002. Available from http://www.searchenginewatch.com/reports/netratings.html.

Sullivan, D. (2004). *Major search engines and directories* [Web site]. Search Engine Watch. Retrieved, 1 January 2005. Available from http://searchenginewatch.com/links/article.php/2156221.

The Economist Survey (2004a, 15 May). A perfect market: a survey of ecommerce. *The Economist*, 1–20. Available from http://www.economist.com/displaystory.cfm?story_id=2646107.

The Economist Survey (2004b, 13 May). Virtual fun. *The Economist*, 1–5. Available from http://www.economist.com/surveys/displayStory.cfm?Story_id=2646152.

The Excite Network (2004). *Excite search frequently asked questions* [Web Site]. The Excite Network. Retrieved, 3 January, 2005. Available from http://www.infospace.com/info.xcite/search/help/faq.htm#q1.

Tomaiuolo, N., & Packer, J. (1996). An analysis of Internet search engines: assessment of over 200 search queries. *Computers in Libraries, 16*(6), 58–62.

Turban, E., King, D., Lee, J., & Viehland, D. (2004). *Electronic commerce: A managerial perspective*. Upper Saddle River, New Jersey: Pearson-Prentice Hall.

Vakkari, P. (2000). Relevance and contributing information types of searched documents in task performance. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, Athens, Greece (pp. 2–9).

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management, 40*(4), 677–691.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing and Management, 40*(4), 693–707.

Veritest (2003). *Inktomi Corp. Web search relevance test* [Web site]. Veritest. Retrieved, 6 May, 2004. Available from http://www.veritest.com/clients/reports/inktomi/inktomi_Web_search_test.pdf.

Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project. Study I. document selection. *Journal of the American Society for Information Science, 49*(2), 115–133.

Wigand, R. J., & Benjamin, R. L. (1995). Electronic commerce: effects of electronic markets. *Journal of Computer Mediated Communication, 1*(2). Available from http://jcmc.indiana.edu/vol1/issue3/wigand.html.

Wildemuth, B. M., & Carter, A. R. (2002). *The perceived affordances of Web search engines: a comparative analysis: SILS technical report 2002-02* [Web Site]. University of North Carolina, School of Information and Library Science. Retrieved, 1 January, 2005. Available from http://ils.unc.edu/ils/research/reports/TR-2002-02.pdf.

Witten, I., Moffat, A., & Bell, T. C. (1994). *Managing gigabytes: Compressing and indexing documents and images*. New York, NY: Van Nostrand Reinhold.

Wu, S. L., & Li, J. Y. (2004). Effectiveness evaluation and comparison of Web search engines and meta-search engines. *Lecture Notes in Computer Science, 3129*, 303–314.