ELSEVIER

Information Processing and Management 40 (2004) 291–317

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Measuring user perceptions of Web site reputation

Elaine G. Toms *, Adam R. Taves

Faculty of Information Studies, University of Toronto, 140 St. George Street, Toronto, Ont., Canada M5S 3G6

Received 20 August 2002; accepted 19 August 2003

Abstract

In this study, we compare a search tool, TOPIC, with three other widely used tools that retrieve information from the Web: AltaVista, Google, and Lycos. These tools use different techniques for outputting and ranking Web sites: external link structure (TOPIC and Google) and semantic content analysis (AltaVista and Lycos). TOPIC purports to output, and highly rank within its hit list, reputable Web sites for searched topics. In this study, 80 participants reviewed the output (i.e., highly ranked sites) from each tool and assessed the quality of retrieved sites. The 4800 individual assessments of 240 sites that represent 12 topics indicated that Google tends to identify and highly rank significantly more reputable Web sites than TOPIC, which, in turn, outputs more than AltaVista and Lycos, but this was not consistent from topic to topic. Metrics derived from reputation research were used in the assessment and a factor analysis was employed to identify a key factor, which we call 'repute'. The results of this research include insight into the factors that Web users consider in formulating perceptions of Web site reputation, and insight into which search tools are outputting reputable sites for Web users. Our findings, we believe, have implications for Web users and suggest the need for future research to assess the relationship between Web page characteristics and their perceived reputation.

Keywords: Web site reputation; Evaluation; Search output; Web site assessment; Google; AltaVista; Lycos; TOPIC

1. Introduction

Multiple search engines retrieve and rank Web sites based on their semantic content and/or on their positioning within the link structure of the Web. These engines output from a dozen to a million or more sites on almost any topic, but often leave the user to sort and sift through the results looking for quality information. The search engines do a reasonable job of ranking Web

* Corresponding author. Tel.: +1-416-978-7802; fax: +1-416-971-1399.
E-mail address: toms@fis.utoronto.ca (E.G. Toms).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2003.08.007

sites according to their semantic content, but they do not incorporate into search algorithms indicators of quality. Assessing the quality of a Web site's content is left to the individual, who often uses criteria other than topicality in his or her assessment (Rieh & Belkin, 1998).

One criterion used in evaluating Web sites is reputation. In general, reputation is an important personal asset in which to invest and protect (Weigelt & Camerer, 1988) and a strategic asset of public and private organizations (Fombrun, 2001). It can be considered an *expectation of quality* (Shapiro, 1982). In essence, one perceives through a variety of factors that a person or object has a good reputation and it raises expectations about the interactions one might have with that person or object. This situation is similar to interactions on the Web. For example, one has expectations about interactions with Amazon.com and those same expectations may not exist for a host of other online booksellers. On exposure to a Web site, however, a user may not know about the reputation or have perceptions about the reputation of that site and may mentally ask a number of questions: Is this "for real"? Can I trust it? Is it accurate? And so on. Users sometimes make split-second decisions about whether to stay and read, or move on quickly to other sites. This decision is based on how the individual perceives a whole series of visual and content cues emanating from the site. From this surface inspection, individuals derive an impression—they assess the quality of that site and thus, infer its reputation.

The volume of Web sites and the predicted growth of Web content preclude the ability to provide human indexing and ranking of Web sites for all given topics. Search engines must be able to automatically sift through this content and highly rank sites that are perceived as reputable to users, and not merely identify pages that are on a topic. At present, identifying good pages is left to users, who have to sift through results lists and the related Web sites, and make quick judgements about a page. In this research, we tested the ability of current search engines to output 'reputable' Web sites.

Unlike typical search tools, TOPIC (http://www.cs.toronto.edu/db/topic/search.html), developed by Rafiei and Mendelzon (2000), identifies the topics (or subjects) for which a Web site has a good reputation. It does so by mining both the link structure and the textual content of that site to indicate how well connected the site is to other well-connected sites on the Web. Thus, upon searching the Web for sites on a topic, TOPIC purports to output and highly rank sites that are highly perceived by the Web community for that topic.

To do this research, which is a follow-up to a pilot study (Keast, Toms, & Cherry, 2001), we develop a metric for testing the reputation of Web sites and compare the output from TOPIC with that of three other search engines (Google, AltaVista and Lycos) to ascertain if certain tools more often output and highly rank 'reputable' Web sites for given topics. Google (http://www.google.com) and TOPIC primarily rely on link structure analysis whereas AltaVista (http://www.altavista.com) and Lycos (http://www.lycos.com) depend on semantic content analysis when outputting and ranking Web sites for a search on a specific topic. In our pilot study, we included search engines and directories with human-selected and indexed content (Yahoo and Britannica.com). This third type was dropped in this larger study, as it performed no better than the automated tools in the pilot, and thus was deemed less essential to the intent of our study—the abilities of automatic indexing and ranking tools to output quality Web sites. Automated tools must be able to address the task of indexing and sifting through a large number of Web sites to allow users to access information they consider to be salient and of good quality. Thus, the work reported here assesses the ability of two types of indexing and ranking tools, represented by four

search engines, to output Web sites that are perceived by Web users to be highly reputable on a specific topic.

2. Previous work

2.1. What is reputation?

Reputation, according to the Oxford English Dictionary, is "the relative estimation or esteem in which a person or thing is held" (Simpson & Weiner, 1989, vol. XIII, p. 678). It is the external *perception* that humans have of another person or object, but it is not necessarily a true indication of that person's or object's qualities (Weigelt & Camerer, 1988). That said, a reputation might be positive or negative or simply neutral. It is used by people in their interactions with another person or object, reflecting an expectation about that person's or object's future behaviour. Unlike tangible attributes, such as size and income, reputation does not exist until a third party perceives it. Reputation is also not singular. A variety of types of reputation may be deemed to exist, for example a reputation for knowledge in a domain, for service, for fast delivery, for moral values, or for high quality content. In addition, it is multidimensional, as there is no single attribute that contributes to, or determines, reputation.

There are numerous ways in which reputation is acquired. For example, a reputation can be formed when people purchase a quality product and transfer that knowledge, word-of-mouth, to friends and colleagues (Allen, 1984); through continuous positive actions (Whitmeyer, 2000); from a mix of signals emitted by an organization, such as historical economic performance (Fombrun & Shanley, 1990); from an organization's directors who are perceived to have high personal reputations (Weigelt & Camerer, 1988); through the physical trappings of an organization, e.g., a prestigious address and expensive office furnishings (Weigelt & Camerer, 1988); or by the awards or prizes won by an individual (Whitmeyer, 2000), among other possibilities. In addition, sometimes the reputation is derived through a third party—by "renting another agent's reputation" (Chu & Chu, 1994). Chu and Chu use an example of a manufacturer acquiring a reputation by selling its product through a reputable retailer.

At the outset, an individual or organization does not have a reputation, or the person interacting with the individual or organization might not be aware of an existing reputation. Yet, some perception of reputation is derived by a person before the acquisition of a product, service or opinion. Using game theory, Weigelt and Camerer (1988) give the example of a player who does not know a second player's true type, but will perceive that player's reputation from cues emitted by the player. Kollock (1994) characterizes these situations as information asymmetries—an exchange relationship in which the partners have unequal information, and where one (or both) partner(s) can behave opportunistically. It can be argued that this situation mirrors the Web information exchange environment, an information exchange situation "in which deceit and opportunism are possible... and where actors can move into and out of different exchange relations..." (p. 314). Reputation, Kollock argues, is a major concern for partners in exchange relationships.

In situations such as those noted above, a retailer strives to retain its reputation by ensuring that a manufacturer's product has quality. This effect is evident in other types of interactions.

According to Wilson (1983), when a person has incomplete information accessible to her (i.e., information asymmetry), decisions become sensitive to personal beliefs and expectations. She realizes that her actions will affect the opinions and expectations of others, which may subsequently affect how those people interact with her. With this chain reaction of potential events in mind, a person will balance a current decision against the long-term effects. Short-term payoffs affect long-term benefits; earlier actions affect later reputation (Wilson, 1983). A manager, for example, may shift the blame for factors that are not under his control in order to protect his reputation (Sridhar, 1994). Reputation reporting systems that have emerged on the Web are based on similar principles (see Dellarocas (2001) for example).

2.2. Web site reputation

A Web site's reputation is much like that of an individual or organization (Fogg et al., 2001; Resnick, Zeckhauser, Friedman, & Kuwabara, 2000). It develops through positive previous exposure, through third-party assessments such as the rating services that have emerged on the Web, or indirectly through the linking of Web sites. But on exposure to a site, a visitor may not be aware of that reputation or, indeed, have any perception of the site's reputation, forcing the visitor to make inferences about its possible reputation (Jarvenpaa, Tractinsky, & Vitale, 2000) based on a surface inspection of the site (Bailey, Gurak, & Konstan, 2001).

Wilson (1983) claims that source authority to a large extent governs the perceived quality of information. There are two main components that contribute to credibility—competence (i.e., expertise) and trustworthiness. Hovland, Janis, and Kelley (1953) differentiate between expertness and trustworthiness in appraising how individuals assess communicator credibility. Rieh and Belkin (1998) note that there is no overall quality control on the Web, nor have users developed standard systems of authority recognition as they have in the print environment (where peer review, reputable names, etc., become well established and recognized). They suggest that in the relatively uncontrolled environment of the Web, users may use other sources of information filtering and that "people depend upon such judgments of source authority and credibility more in the Web environment that in the print environment" (p. 288). Their analysis revealed that 'source' was more often cited in their interviews with scholars regarding information quality on the Web than any other quality. Highly regarded sources were those deemed credible by the participants—university sites, library sites, or the status of the individual responsible for the site. However, as source can be difficult to ascertain in a Web environment, other characteristics are also used in making judgments about authority and credibility.

Various approaches have been taken in trying to elucidate which factors lead a person to decide that information, or the source from which it comes, is reputable. Often, these studies have both determined and looked at facets of reputation; in other words, they have examined various elements of information, format, or source that lend themselves to a user forming perceptions of information or source reputability. Most of this work has been done in the e-commerce sector, assessing the likelihood of a person purchasing a product or service. Assessing the reputations of Web sites outside of a purely e-commerce environment appears to have been skirted by researchers.

Fogg et al. (2001) examined the relationship between trustworthiness/expertise and credibility in Web sites. They define credibility as believability, arguing that credibility does not exist within

the information itself, but is rather a perceived quality. For information to be credible, it must be believable. From an overview of literature in communication theory, they deduce that the two key components of credibility are trustworthiness and expertise.

Using this conceptual foundation, Fogg et al. surveyed 1400 Web users. Participants in this study examined no actual sites. Rather, they indicated the perceived believability of a site based upon 51 Web site attributes that were grouped in seven dimensions, such as expertise, trustworthiness, commercial implications, etc. The attributes were presented in a seven-point Likert scale format (from much less believable to much more believable). Trustworthiness and expertise of a site were found to have significant effects on the believability, and thus the perceived credibility, of a Web site.

Jarvenpaa et al. (2000) took a different approach in examining the relationship between consumer attitude to an Internet store and intention to purchase. They hypothesized that trust is positively related to the store's perceived reputation. Participants performed four shopping tasks: two involving buying books and two involving making travel plans, using designated Web merchants (some with bricks-and-mortar entities, some completely virtual). Participants completed an experiential survey evaluating their impressions of the Web merchants' reputation, size, and trustworthiness, as well as their attitudes towards the store, willingness to buy, risk perception, and shopping enjoyment. Not surprisingly, they found that a consumer's trust in an e-store is positively related to the store's perceived reputation. In a followup study that included a cross-cultural comparison using participants in three different countries, Jarvenpaa, Tractinsky, Saarinen, and Vitale (1999) confirmed the results. Lynch, Kent, and Srinivasan (2001) similarly examined the shopping tasks completed by 299 participants in 12 countries. Like Jarvenpaa et al. (1999, 2000), they found that trust and site quality were key indicators in likelihood to purchase. Notably though, the effect on loyalty varied with the product and country.

In studies of Web sites to date, trustworthiness and expertise seem to be significant in interchanges between people and/or objects. However, we did not uncover studies that used reputation in the assessment of search engine output.

2.3. Assessing the reputation of a Web site

To date, a variety of methods have been developed on the Web to mediate and convey reputation. For example, eBay (2002) provides a system for ranking users, whereby buyers and sellers provide feedback on those with whom they buy and sell. Here, buyers and sellers develop reputations that each can use when considering whether to engage in transactions. Amazon (2002) and Epinions (2001) similarly provide a service whereby book (and other product) reviews can be scrutinized by other users, thereby providing a reputation system for reviewers.

Alexa (2001) ranks sites using a site's traffic statistics relative to other sites on the Web. As well, it denotes the number of links pointing to a site by other sites and provides reviews by users. These factors, arguably, provide some conferral of reputation on sites; however, traffic ranking and link statistics must be viewed within the framework of the popularity of a topic on the Web. Sites on popular topics will likely receive higher amounts of traffic, and *Alexa* cannot provide ranking statistics for a given topic, but can only compare a site to all other sites on the Web.

Media metrix (2001) and *Nielson Netratings* (Netratings, 2001) provide audience ratings for sites on the Web both generally, and within specific but broad topics. These ratings provide pure

traffic statistics and convey which sites receive the largest proportion of traffic share. The reputation of sites is conveyed to the extent that high traffic can be understood, within these systems, as being indicative of reputation.

In addition, numerous Web sites on information evaluation of Web resources emphasize the importance of determining the reputability of information provided. Generally, concepts such as relevance, reliability, authority (Science Academy, 2001), or quality of content, usability, and authority (Argus Associates, 2001) are used. The factors that users look at in determining these attributes are plentiful. This value lies both in the interplay that reputation has on perceptions of relevance (Barry & Schamber, 1998) and in the needs of users who depend on the Web for pertinent information (D'Esposito & Gardner, 1999).

In recent years, many rating services, such as *Global Information Infrastructure Award* and the *Argus Clearinghouse Seal of Approval*, have emerged to assess a variety of qualities of Web sites, such as authorship and disclosure (Jadad & Gagliardi, 1998). In addition, other services have been developed to handle corporate reputation management (Resnick et al., 2000). While the former tends to assess the quality of sites on a range of topics, the latter is limited to managing and assessing the reputation of a single site. In each case, the assessment is mostly manual or partially automated.

2.4. Assessing reputation using TOPIC

TOPIC, developed by Rafiei and Mendelzon (2000), uses a different approach. For any page on the Web, they define its reputation as the probability that a random visitor who is looking for a specific topic is likely to visit that page. Using the Hubs and Authorities model developed by Kleinberg (1999), they deem that a page is an authority on a topic if it is linked to by good hubs on that topic, and similarly, good authorities are pointed to by good hubs. The reputation of a page on a topic is thus proportional to the sum of reputational weights of all pages pointing to it that are also on that topic. This model is a variation on PageRank (Brin & Page, 1998) in that it identifies the value of a page for a specific topic. TOPIC thus demonstrates for a single Web site its value as an authority for a particular topic, which, in turn, may indicate its 'reputation' on the Web for that particular topic. This intimation of reputation follows the general theory that a reputation is developed from past positive performance. In this case, sites with good reputations are 'recommended' by hubs and authorities that have good reputations. Like people and organizations, a quality hub is unlikely to link to a poor quality site as it may detract from the reputation of the quality hub. TOPIC's ability to output reputable sites showed much promise in its initial test. But in a Web environment, when the sites may not be known, can TOPIC output sites on a particular topic such that an average user would discriminate among the reputability of sites? This question was a key component of this study.

2.5. Indicators of reputation

From previous work examined, we posit that reputation is composed of multiple dimensions or indicators. In this research, we do not assess directly the quality, or reputation, of a Web site; we assess whether ordinary people on surface inspection perceived that a site has quality, and whether it could be inferred that the site has a 'good reputation'. We do so by assessing a series of

characteristics that are either explicitly or implicitly associated with reputation. This strategy was developed because reputation is an external perception of quality. We opted not to directly ask about a site's reputation, as we surmised that the Hawthorne effect would likely develop. Instead, we employed an indirect approach to determine how reputable a Web site is perceived. We assessed the following criteria:

1. *Trust* is defined as "confidence in or reliance on some quality or attribute of a person or thing, or the truth of a statement" (Simpson & Weiner, 1989, vol. XVIII, p. 623). It is perhaps one of the most, if not the most, studied user perception dealing with information systems, having been the subject of numerous theses, reviews and studies (see, for example, Hosmer, 1995). Trust has been found to significantly affect purchaser behaviour (Jarvenpaa et al., 1999, 2000; Schurr & Ozanne, 1985) and is considered tightly coupled with reputation (e.g., Hill, 1990; Jarvenpaa et al., 1999, 2000; Reagle, 1996; Whitmeyer, 2000). Moreover, trust is considered one of the most important factors affecting exchange relationships (Klang, 2001) and human interactions with systems (Bailey et al., 2001).

2. *Authority* is synonymous with expertise. Expertise was identified as a property of reputation by Chen and Singh (2001). Authority indicates an element of knowledge and competence (Fogg et al., 2001; Fogg & Tseng, 1999). In an overview and analysis of work on credibility perceptions of information, Wathen and Burkell (2002) concluded that expertise and knowledge, in addition to trustworthiness, are qualities "that mark credible sources of information" in face-to-face interaction and that current research suggests that electronic information will be received by individuals in the same manner (p. 7). Thus, a good reputation cannot exist without a perception of authoritativeness in the person or object.

3. *Aboutness* indicates the extent to which a site is actually *about* the topic searched. The intent of this study is to examine reputation for a particular topic, but we did not assume that the search engines actually output sites that were on topic. There are numerous examples of the semantic differential between the topic searched and the site retrieved.

4. *Re-visit* is an indicator of loyalty. If a user is willing to re-visit a site for information on a designated topic, then it can be argued that the user is showing loyalty, as the concept is used in e-commerce (Lee, Kim, & Moon, 2000). Repeat visits to a Web site indicate loyal customers and may affect the value of the business; no repeat visits is synonymous with a zero business value. While loyalty is important to e-commerce, those repeat visits reflect an indicator of quality to all types of sites. Lee et al. (2000) validated a model of customer loyalty that included comprehensive information, shared values, specificity and trust. Whether a person would re-visit a site for a particular topic adds a refinement to the factor of aboutness. Users may return to a site for reasons other than to access information about the designated topic.

5. *Recommend* as an indicator of reputation could be construed in terms of how a person might interact with that site in the future. Doby and Caplan (1995) argue that a person's self-esteem, an important social need, is determined at least partially by how she is perceived by others. We extrapolate, thus, that people will recommend only sites that they perceive to be of quality because to do otherwise might tarnish their personal reputation. Moreover, recommendations form the core of Nielsen's (1999) interpretation of reputation managers on the Web, where reputation constitutes an aggregate of multiple users' professed satisfaction or dissatisfaction with a site. Again, we looked at whether the site would be recommended for information on a particular topic.

6. *Ranked:* the intent behind this indicator was to divide the set of sites into two groups: those that are of sufficient quality to be selected as the best sites, and those that are not. In making this decision, a person holistically decides that some sites are good and some are not so good. It provides a refinement and, for the researchers, a way of assessing the earlier criteria.

In summary, for a site to be perceived by users as reputable on a specific topic, we argue that the site must be perceived as trustworthy, authoritative and about the topic. Users will rank reputable sites among the best on the topic. Furthermore, users will protect personal reputations by only recommending reputable sites and will exhibit loyalty cues by only returning to reputable sites.

2.6. Research questions

The key questions that we addressed are:

(a) Do certain types of search tools yield sites that are perceived to be more reputable—authoritative and trustworthy—than others?
(b) Do search engines that use link structure analysis highly rank more reputable Web sites than those that rely primarily on semantic content analysis?
(c) Does TOPIC, a search engine designed to rank sites by their reputation for a particular topic, outperform the other search engines in highly ranking Web sites perceived as reputable for a particular topic?

Initially, we conducted a small pilot study using one topic (Keast et al., 2001). In this study, 22 participants used the criteria identified above to assess the reputations of 17 sites, all on Movie Reviews. The sites had been retrieved using the following: (1) Google and TOPIC, which use intra-site link structure analysis; (2) Altavista and Lycos, which use automatic indexing; and (3) Yahoo and Britiannica, which use human-assigned categories and assessments. Britannica sites were dropped from the assessment because too few could be accessed. Findings showed that TOPIC performed on average as well as Altavista and outperformed Google. Surprisingly, they matched or outperformed the human-selected-and-indexed Yahoo. In general, the two non-human techniques performed on par with the human-generated tool. On the basis of these results, we conducted a larger study that is the essence of this paper.

3. Methods

3.1. Overview

From the pilot study (Keast et al., 2001), we knew that assessing individual Web pages was a mentally taxing experience. Examining and evaluating a set of 20–30 sites in one sitting was about the maximum effort that a single person reasonably could expend on the task. In addition, we predicted that we would likely have attrition if participants were asked to return multiple times. In the design of our main study, we compromised by using a mixed between- and within-subjects design.

We selected 12 topics to use in our study and identified 20 Web sites per topic using the four different search tools. These numbers have no intrinsic value except that we wanted a sufficient range of topics with a sufficient number of Web sites assessed by a sufficient number of people. This breadth of data would allow us to report statistical significance about the differences among the tools and topics. To reduce the mental effort, we divided the topics into four groups, so that a single person would examine 20 sites in a single sitting with a maximum of 60 in a single session. The study was, in essence, replicated four times with a change of topics and participants. To avoid the attrition problem, we opted for longer sessions with a break at the change of each topic.

3.2. Participants

Eighty participants (40 female, 40 male) participated in the study. Overall, they were a relatively youthful group: 84% were under 36 years old. About 33% had completed a high school education or community college program, 41% had an undergraduate degree, and the remaining 26% had a graduate education. The participants were frequent Web users: 85% claimed to browse the Web more than 10 times a week. Only 8% browsed the Web less than five times a week, and all had previously browsed the Web. Participants were recruited by advertising on various electronic and physical billboards on the University of Toronto campus. No recruitment was made off-campus, although the urban and public nature of the campus suggests that some participants may not have been members of the university community and were responding to notices placed on physical billboards. For their participation, they were paid $40.00 in cash after completing the session.

3.3. Selecting the topics

To assess the reputations of a subset of Web sites, we first needed to identify a set of topics. The topics selected were deemed to be of 'general interest', which, for this study, was defined as those topics for which an average potential participant would likely be familiar and would likely possess some knowledge. The topics used in the study were selected from the *About—the human Internet* (About.com, 2001a) Web site. At approximately seven hundred topics, it provided a manageable but broad variety of topics. In addition, topics on this Web site reflect perceived or actual interests among World Wide Web users; the site claims that 1 in 5 online users visit *About.com* each month (About.com, 2001b).

Guidelines were developed to derive a master list of topics from which the test topics were sampled. In general, we avoided:

- faith-based topics (e.g., *Catholicism*) as they may require specialized knowledge
- technical topics (e.g., *network analysis*) and specific medical conditions (e.g., *anxiety disorders*) that require special expertise
- specific geographical locations as topics unto themselves due to their inherent broadness (e.g., *Canada*)
- specific television shows that may require specialized knowledge (e.g., *One Life to Live fans*)
- topics that may lead to complicated search strings due to lack of controlled vocabulary (e.g., *honeymoons* versus *romantic getaways*)

Table 1
Search topics by group

Group 1	Group 2	Group 3	Group 4
Movie Reviews	World News	Parenting	Parks Canada
Travel	Gardening	Hockey	Alternative Energy
Politics Canada	Antiques	Walking	Sailing

In addition, when forced to choose between a general topic (e.g., *travel*) and a facet of that topic (e.g., *travel—UK*), the more general topic was used. After filtering the *About—the human Internet* list, approximately 100 topics remained. From this list, eleven topics were randomly selected using a random number generator. The topic, *Movie Reviews*, which was used in the pilot study, was retained so as to have a basis for comparison between the two studies. The final list of topics is presented in Table 1.

Before selecting *About—the human Internet* as the source for topic suggestions, we explored other options. Both Media metrix (2001) and *Nielson Netratings* (Netratings, 2001) were consulted as they track sites that receive heavy traffic. However, their topical breadth proved too limited for the number of topics required for this study. Yahoo (2001) and *Open Directory* (Netscape, 2001) categories were too plentiful with which to derive a list, and the several levels of increasing specificity within single topics was too difficult to simplify in a systematic way. These options were not pursued any further.

3.4. Identifying the sites by topic

Each topic was searched in advance of the test using the standard search interface for each tool. No advanced search features were used on any tool, and all search tool features such as "featured listings" were ignored to avoid the results of human indexing or advertisement. Each of the topics listed in Table 1 was searched during a 2-h period, and the top five ranked hits from each tool were selected. Each of these Web sites was integrated into a single list for each topic, and all duplicates were removed.

In each case, only top-level domains were chosen; subordinate pages within listed sites were ignored. When two or more links led to the same site by different URLs, a common URL was established randomly from among the links. Sorting the results in this manner assured that participants would not visit a Web site twice. How this was presented to participants is described more fully in procedures.

3.5. Variables

In this study, we examined three independent variables:

(1) type of ranking technique: semantic content analysis or link structure analysis
(2) search tool: AltaVista, Google, Lycos, and TOPIC
(3) topic: 12 topics as listed in Table 1

The differences between levels of the following criteria were measured with a five-point Likert scale (for criteria a, b, and c), yes/no (for criteria d and e), or a checkmark (for criterion f) using the following, which indicate the quality of a site on a particular topic:

(a) trustworthiness of the information (from *not at all trustworthy* to *very trustworthy*)
(b) authoritativeness of the information (from *not at all authoritative* to *very authoritative*)
(c) aboutness of the information on the designated topic (from *not at all about "designated topic"* to *very much about "designated topic"*)
(d) willingness to return to the site for further information on the designated topic, i.e., loyalty (*would you return to site for information on "designated topic"?*)
(e) willingness to recommend the site as a good source of information on that topic to a friend or colleague, i.e., risk personal reputation (*would you recommend site for information on "designated topic"?*)
(f) rank a site as the best. In the pilot study we asked participants to assign rank order to this list, but we found that participants, in many cases, merely ticked the sites and provided no ordering. Thus, this particular task was modified for this study

In addition, two covariates were assessed:

(i) interest level in the topic: None, Minor, Major, Main
(ii) previous visit to a site: yes/no

In essence, we wanted to ascertain if prior exposure to a site or interest in a topic affected the ratings.

We were interested in the user's perceptions of Web sites that were highly ranked by different types of search tools. Users assessed each Web site presented by topic using the criteria a–f discussed above. In addition, we wondered if this perception was affected by participant's interest level in the topic and by whether they had previously visited the site.

3.6. Instruments

Participants used three instruments over the course of the study to assess the criteria, as well as to allow us to collect a profile of the participant group. The following instruments were used:

(a) *Assessment Questionnaire (see Fig. 1):* One of these was used with each site. Each identified the Title (or Name) of the Web site at the top of the page and asked the following: whether participants had previously visited the site; perceptions of trustworthiness, authoritativeness, and aboutness of the Web site; and whether the participants would return to and recommend the site.

(b) *Ranking Questionnaire:* This page listed in order of visiting all of the sites visited on the topic. Participants were asked to: identify the five sites that they consider to be the best for that topic; indicate the URLs for sites that did not appear in this exercise, but that they use; and indicate their interest level in the topic (Main/Major/Minor/None).

(c) *Demographics Profile:* This is a typical demographic survey that asked participants about their age, gender, educational level, and how often they use the Internet.

University of Toronto
Faculty of Information Studies

Participant ID_____

Assessing Outputs from World wide Web Search Tools

Questionnaire for Site **[Title or Name of Site]**

Click on the link for Site _____ , and answer the following questions:

1. Have you previously visited or heard of this site? No ☐ Yes ☐

Browse through the site. Spend no more than two minutes doing this. Please answer the following questions.

2. This site is (*mark the appropriate number with a circle*):

not at all trustworthy	1	2	3	4	5	very trustworthy
not at all authoritative	1	2	3	4	5	very authoritative
not at all about **[topic]**	1	2	3	4	5	very much about **[topic]**

3. Would you return to this site in the future for **[topic to be inserted]** information? No ☐ Yes ☐

4. Would you recommend this site to a friend or colleague? No ☐ Yes ☐

Fig. 1. Assessment Questionnaire.

Participants received one evaluation package for each topic to be assessed. Within the evaluation package, the order of the sites was randomised by participant to control for a learning effect. The package contained, in the following order:

(i) Cover page that identified the topic, the participant by unique identifier, and the list of Web sites to be visited
(ii) Copies of the Assessment Questionnaire (Fig. 1)—one for each Web site
(iii) One copy of the Ranking Questionnaire (described above)
(iv) One copy of the Demographics Profile

3.7. Procedure

Testing took place in a computer lab over a three-week period. Participants used Pentium 3 computers running Windows NT with 17-in. monitors equipped with Internet Explorer and Netscape that were connected to the university's high speed Internet connection. A special "Home" page was created for each topic for each participant duplicating the "Cover page" described above. This page was set as "Home" on the Web browsers so that participants could always find their way back after viewing each Web site. When participants began assessing their next assigned topic, "Home" on the browsers was set to reflect the next list of Web sites.

Before the arrival of participants, a consent form and one evaluation package (corresponding to that participant's first designated topic) was placed at each workstation. Note that in addition to randomising the order of the sites, the order of topics was also randomised for each participant. At the start, participants signed consent forms, had the task and survey explained to them, and had any questions answered. Participants were told they could use either Netscape Navigator or Internet Explorer browsers. Participants started with the "Home" page described above and visited each site, in turn, by following the hyperlinked titles to each site. As was predicted, not all sites were available during the course of the study. When participants had trouble loading sites, they were asked to try reloading the site three times, and to try again when they had assessed the last site on a particular topic. When one topic was finished, participants were asked to take a 15-min break after which their next designated topic was assigned. When the sections for all their assigned topics were completed, they were asked to fill out the demographics profile. Refreshments were provided. The entire process took about 2–2.5 hours per participant.

These sessions were repeated with new participants until all 12 topics had been assessed. The size of each session varied. Sometimes a session contained only a single person and sometimes a dozen. The study proceeded until four groups of 20 participants, with each group assessing Web sites for three topics, had completed the study (4 groups $\times$ 3 topics = 12 topics assessed). Participants were assigned to groups according to the topics as identified in Table 1.

3.8. Data analysis

The series of four tests resulted in 4800 individual assessments (80 participants $\times$ 12 topics $\times$ 20 sites). Approximately 4% of the cases had missing data because some sites were not available at the time of testing and, in the case of some user perception questions, the participant did not answer all questions. All data were first analyzed using primarily SPSS's GLM (univariate or multivariate analysis of variance) and chi-square, depending on the type of variable. The Bonferroni-adjusted post hoc test was used for pairwise comparison between the different tools. Because of the mixed results obtained from these analyses, factor analysis was used to derive a measure of reputation. The Kaiser–Meyer–Olkin measure and the Bartlett's Test of Sphericity indicated that the sampling was adequate and the variables were likely to be related. The resulting factor was considered to have internal reliability according to Cronbach's alpha. In addition, after we had tested reputation by tool, we used Tukey's honestly significant difference test to look at pairwise comparisons among topics.

4. Results

4.1. Overview

Eighty participants rated the 240 Web pages representing 12 topics, on average, 3.7, 3.9 and 3.8 for *authoritativeness*, *trustworthiness* and *aboutness*, respectively (using a five-point scale). They indicated that they would *recommend* 52%, and would *return* to 53% of the sites for more information on the topic. Overall, they placed about 31% of the sites in the "top five" sites for information on the topic.

Table 2
Percentage of participants with interest in topics

Search topic	Interest (% of participants)		
	None	Minor	Major
Alternative Energy	10	85	5
Antiques	55	40	5
Gardening	40	40	20
Hockey	35	50	15
Movie Reviews	5	60	35
Parenting	35	45	20
Parks Canada	5	40	55
Politics Canada	21	58	21
Sailing	50	45	5
Travel	10	20	70
Walking	21	42	37
World News	0	26	74
Mean (%)	24	46	30

Most of the sites were new to participants before the start of the study. Only 10% of the sites had been previously visited or heard of by participants. Interest in the topics varied among the group. Approximately 45% of participants indicated that the topics were of minor interest while 25% of the participants expressed having no interest at all in the topics. The remainder (30%) had a major or main interest in the topics.

Among the 12 topics, only a few were of major interest to many of the participants as illustrated in Table 2. Major and Main were collapsed into a single attribute. Only three: World News, Parks Canada, and Travel were of Major or Main interest to more than half the participants. Two topics, Antiques and Sailing, were of No interest to half the participants.

4.2. Effect of prior knowledge/exposure and interest on assessment ratings

We were interested in determining whether prior knowledge of a site and/or specific interest in the topic affected *authoritativeness*, *trustworthiness* and *aboutness* ratings. There was an interaction of previous visit and interest level ($F(6, 4419) = 5.140$, $p < 0.0005$). In an analysis of simple effects, the results were significant only for *aboutness* ($F(2, 4419) = 16.092$, $p < 0.0001$). As illustrated in Table 3, those with a major interest in the topic tended to rate sites that they had not previously visited as being less on topic than those for which they had prior knowledge.

There was a main effect of previously visited; those who had previously visited a site tended to rate its *trustworthiness* ($F(1, 4419) = 150.11$, $p < 0.001$), *authoritativeness* ($F(1, 4419) = 160.269$, $p < 0.001$) and *aboutness* ($F(1, 4419) = 19.265$, $p < 0.001$) as significantly higher than those who had not previously visited the site, regardless of the interest level in the topic. In summary, it seems that those with an interest in the topic who had previously visited a site were more conservative and/or critical in their assessment of the topicality of a site. Overall, the three ratings for sites were affected by whether the site had been previously visited.

Table 3
Effect of previous visit and interest on ratings of authoritativeness, trustworthiness and aboutness

Previously visited	Interest level	*N*	Trustworthiness	Authoritativeness	Aboutness
No	None	1003	3.4	3.2	3.6
	Minor	1890	3.5	3.4	3.8
	Major	1118	3.5	3.4	3.5
	Total	4011	3.5	3.3	3.6
Yes	None	54	4.2	4.2	3.8
	Minor	129	4.2	4.1	3.9
	Major	226	4.3	4.2	4.2
	Total	409	4.3	4.2	4.0
Total	None	1057	3.4	3.2	3.7
	Minor	2019	3.5	3.4	3.8
	Major	1344	3.7	3.5	3.6
	Total	4420	3.6	3.4	3.7

Previous visit and interest level were associated with three additional variables: *will return*, *will recommend*, and *highly rated* (i.e., ranked as a top five site on the ranking questionnaire). In general, those that had previously visited the site were equally likely to *return*, *recommend* or *highly rate* a site according to their interest level as illustrated in Tables 4–6. When participants had not visited the site, the associations were much less clear for *will return* ($\chi^2 = 23.317$, df = 2, $p < 0.001$) and *will recommend* ($\chi^2 = 18.804$, df = 2, $p < 0.001$). The same was not true for *highly rated*—whether a site had previously been visited ($\chi^2 = 12.119$, df = 2, $p = 0.002$) or not ($\chi^2 = 6.282$, df = 2, $p = 0.043$). Thus, the ratings of *will recommend*, *will return*, and *highly rated* can be predicted if the site had been previously visited, but the same is not true for those which had never been seen.

Because of these significant albeit mixed results, the variables previously visited and interest level were included in the rest of the analysis.

Table 4
Effect of previous visit and interest level on will return

Previous visit	Interest level	Will return to site		Total
		No	Yes	
No	None	625 (28.0%)	382 (21.4%)	1007
	Minor	1021 (45.7%)	879 (49.2%)	1900
	Major	589 (26.4%)	527 (29.5%)	1116
	Total	2235	1788	4023
Yes	None	15 (14.6%)	39 (12.8%)	54
	Minor	35 (34.0%)	93 (30.6%)	128
	Major	53 (51.5%)	172 (56.6%)	225
	Total	103	304	407

Table 5
Effect of previous visit and interest level on will recommend

Previous visit	Interest level	Will recommend site		Total
		No	Yes	
No	None	602 (27.5%)	405 (22.1%)	1007
	Minor	976 (44.6%)	924 (50.3%)	1900
	Major	609 (27.8%)	507 (27.6%)	1116
	Total	2187	1836	4023
Yes	None	14 (13.2%)	40 (13.2%)	54
	Minor	36 (34.0%)	93 (30.8%)	129
	Major	56 (52.8%)	169 (56.0%)	225
	Total	106	302	408

Table 6
Effect of previous visit and interest level on highly rated

Previous visit	Interest level	Highly rated		Total
		No	Yes	
No	None	729 (25.5%)	278 (23.8%)	1007
	Minor	1314 (45.9%)	586 (50.3%)	1900
	Major	817 (28.6%)	302 (25.9%)	1119
	Total	2860	1166	4026
Yes	None	13 (7.4%)	41 (17.6%)	54
	Minor	67 (38.1%)	62 (26.6%)	129
	Major	96 (54.5%)	130 (55.8%)	226
	Total	176	233	409

4.3. Assessment by type of ranking technique

As previously mentioned, two types of indexing and ranking techniques were assessed in this study: one that relied primarily on the Link structure of the Web (Links technique) and one that relied primarily on the Semantic content of sites (Semantics technique). The six variables used to assess the perceived quality of a site indicated significant differences between the two types of indexing techniques. Output from the tools based on the Links technique were perceived as more *authoritative* ($F(1,4597) = 52.835$, $p < 0.001$), more *trustworthy* ($F(1,4597) = 30.563$, $p < 0.001$) and, in general, ranked higher on *aboutness* ($F(1,4597) = 40.007$, $p < 0.001$) than those that relied on the Semantics technique (see Fig. 2).

Participants indicated that they would *return* to 50% of the sites generated by the Links technique compared with 44% of the sites generated by the Semantics technique ($\chi^2 = 16.963$, $df = 1$, $p < 0.0005$). They indicated that they would *recommend* about 51% of sites generated by the Links technique to friends for information on the defined topic, while 46% of sites from the Semantics technique were *recommended* ($\chi^2 = 11.025$, $df = 1$, $p = 0.001$). Overall, about 34% of the sites generated by Links technique were *highly rated*, while 30% of those found using the Semantics technique were *highly rated* ($\chi^2 = 11.201$, $df = 1$, $p = 0.001$). Thus, there appears to be a

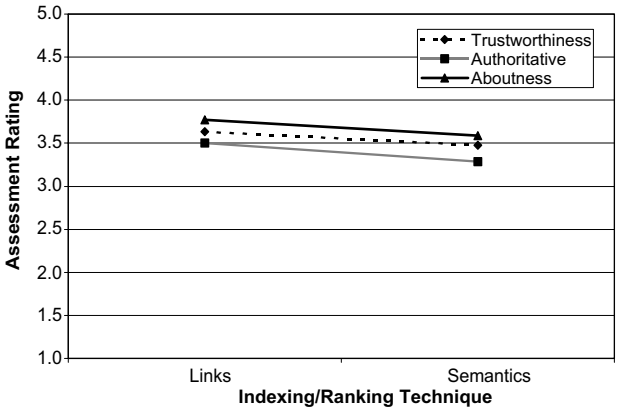

Fig. 2. Average assessment rating of trustworthiness, authoritativeness and aboutness by type of indexing technique.

strong association between the output generated by the two search tools using the Links structure technique for ranking sites and participants' willingness to *return to*, *recommend* and *highly rate* the site.

Overall, the tools that used a Links technique tended to receive higher ratings or be more closely allied with the user intentions of returning to or recommending a site.

4.4. Effect of previous exposure

How were the *authoritativeness*, *aboutness* and *trustworthiness* ratings affected by interest in topic and prior exposure to the site? There was no interaction of ranking technique with previous visit ($F(2,4540) = 1.309$, ns) or interest level ($F(2,4540) = 0.876$, ns). In addition, we examined the association between the type of ranking technique and the will *recommend*, will *return* and *highly rated* variables. While there is no strong association between those variables and sites previously visited, this was not the case with those not previously visited. Among the sites not visited, more sites retrieved using the Links technique would be *returned* to in the future than those retrieved by Semantics technique ($\chi^2 = 17.629$, $\mathrm{df} = 1$, $p < 0.00051$). Similar results were found for both willingness to *recommend* ($\chi^2 = 12.639$, $\mathrm{df} = 1$, $p < 0.0005$) and *highly rated* ($\chi^2 = 9.985$, $\mathrm{df} = 1$, $p < 0.002$).

Interest level also affected the ratings by type of indexing technique. Those with a major ($\chi^2 = 9.943$, $\mathrm{df} = 1$, $p = 0.002$) or minor ($\chi^2 = 7.003$, $\mathrm{df} = 1$, $p = 0.008$) interest in the topic were more likely to *return* to sites retrieved by Links technique than by the Semantics technique. Only those with a minor interest ($\chi^2 = 6.338$, $\mathrm{df} = 1$, $p < 0.012$) in the topic were more likely to *recommend* sites and *highly rate* sites ($\chi^2 = 11.113$, $\mathrm{df} = 1$, $p = 0.001$) that were retrieved by the Links technique; this was not the case for those who had a major interest in the topic or no interest at all.

In summary, previous exposure and interest in the topic had no affect on the *authoritativeness*, *trustworthiness* and *aboutness* ratings for sites retrieved using a specific ranking technique used to retrieve the site. Both, however, did have a strong but mixed association with participants' willingness to *recommend*, *return* to and *highly rate* a site, although it was not consistent across all the interest levels. Those with a minor interest in the topic were more likely to *recommend*, *return* to and

highly rate sites derived from the Links technique than those with a major or no interest in the topic.

4.5. Assessment by search tool

The two types of ranking techniques were represented in this study by four search tools. Since the precise algorithms on which each is based are not publicly known, the data was analyzed at a lower level of granularity to determine if the differences observed by type of ranking technique existed primarily within a particular tool. In essence, did one of the tools in these two groups contribute more to its positive (or negative) performance? Google and TOPIC use the Links technique while AltaVista and Lycos primarily use Semantics and to the best of our knowledge, link analysis was not an important criterion in the ranking within these latter two tools at the time the data was collected.

There were significant differences in the assessment ratings by type of tool ($F(9, 4598) = 10.607$, $p < 0.001$). In an analysis of Bonferroni-adjusted post hoc pairwise comparisons, Google sites were perceived to be more *authoritative* than the other three, but both Google and TOPIC sites were considered more *trustworthy* than AltaVista and Lycos sites. While Google sites were considered more *about* the topic than the other three, TOPIC sites exhibited more *aboutness* than Lycos and AltaVista. In general, Lycos was rated significantly lower on all measures, while Google was rated significantly higher on all measures.

Participants expressed a willingness to *return* to about 52% of those sites retrieved by Google to a low of 44% of the sites retrieved by AltaVista ($\chi^2 = 18.746$, df = 3, $p < 0.0001$) as illustrated in Table 7.

Participants indicated a willingness to *recommend* about 53% of those sites retrieved by Google to a low of 45% of the sites retrieved by AltaVista ($\chi^2 = 15.198$, df = 3, $p = 0.002$) as illustrated in Table 8.

Participants identified the best sites on the topic, which we have called highly rated sites, as illustrated in Table 9. Overall, participants identified 36% of the sites retrieved by Google as *highly rated*. Twenty-eight percent of Lycos sites were considered as such. In general the type of tool affected the *highly rated* status of sites ($\chi^2 = 17.537$, df = 3, $p = 0.001$).

4.6. Effect of previous exposure

On average, participants had previously visited about 10% of the sites. These sites were fairly evenly distributed across the sites output from the four search tools, and no further analysis was done with previous visits by search tool.

4.7. Effect of tool by topic

Analyses to date have dealt with the topics of sites as an aggregate. Yet, as illustrated in Table 1, the topics were quite diverse—from sports to current events to hobbies and general interest. We wondered if different tools dealt more or less favourably with different topics. A multivariate analysis of variance identified an interaction of tool and topic. The output by tool differed according to the site's *trustworthiness* ($F(33, 4597) = 8.030$, $p < 0.0001$), *authoritativeness*

Table 7
Indication of will return to site, by search tool

Will return	Search tool				Total
	TOPIC	Google	Lycos	AltaVista	
No	597 (51%)	563 (48%)	617 (56%)	652 (56%)	2429
Yes	572 (49%)	601 (52%)	495 (44%)	511 (44%)	2179
Total	1169	1164	1112	1163	4608

Table 8
Indication of will recommend a site, by search tool

Will recommend	Search tool				Total
	TOPIC	Google	Lycos	AltaVista	
No	603 (52%)	552 (47%)	602 (54%)	635 (55%)	2392
Yes	566 (48%)	613 (53%)	510 (46%)	528 (45%)	2217
Total	1169	1165	1112	1163	4609

Table 9
Highly rated sites for the topic, by search tool

Ranked	Search tool				Total
	TOPIC	Google	Lycos	AltaVista	
No	802 (69%)	748 (64%)	801 (72%)	814 (70%)	3165
Yes	368 (31%)	418 (36%)	312 (28%)	350 (30%)	1448
Total	1170	1166	1113	1164	4613

($F(33, 4593) = 10.253$, $p < 0.0001$) and *aboutness* ($F(33, 4593) = 10.896$, $p < 0.0001$). An analysis of simple main effects indicated significant differences between the tools by topic (see Table 10).

Results on all variables were somewhat mixed across all tools. Hockey was the only topic in which all four tools performed about the same on all measures. For example, for Movie Reviews, 33% indicated that they would *return* to TOPIC generated sites, compared with 28%, 15% and 24%, respectively, for Google, Lycos and AltaVista, a response not unlike the one obtained in the pilot study (Keast et al., 2001). This is markedly different for Politics Canada, in which 34% of the participants would *return* to Lycos sites, compared with 21%, 24% and 28% for AltaVista, TOPIC and Google. Similar and somewhat inconsistent patterns existed for different measures and for different topics. Although, in general, sites retrieved from Google were more highly perceived than those from the other three tools, this was not consistent across all measures. Selectedly, Google and TOPIC were equally matched, and TOPIC and AltaVista were equally matched while Lycos rarely moved from the bottom position.

4.8. Defining a measure of reputation

Because of the mixed results on all measures, factor analysis was used to ascertain if there were any potential underlying factors of reputation. Eleven measures were loaded initially:

Table 10
Differences in search tool ratings among search topics

Search topic	Rating					
	Trust	Authority	Aboutness	Will return	Will recommend	Highly rated
Alternative Energy	ns	ns	$p < 0.0001$	ns	ns	$p = 0.055$
Antiques	ns	ns	$p < 0.0001$	ns	$p = 0.055$	ns
Gardening	$p = 0.005$	ns	$p < 0.0001$	ns	ns	$p = 0.053$
Hockey	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$
Movie Reviews	ns	ns	$p < 0.0001$	$p < 0.0001$	$p = 0.001$	$p = 0.007$
World News	$p = 0.014$	$p = 0.031$	$p < 0.0001$	ns	ns	ns
Parenting	$p = 0.001$	$p = 0.005$	$p < 0.0001$	ns	ns	ns
Parks Canada	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$	ns	ns	ns
Politics Canada	ns	$p = 0.026$	$p < 0.0001$	$p = 0.004$	$p = 0.002$	ns
Sailing	$p = 0.021$	$p = 0.036$	ns	$p < 0.0001$	$p < 0.0001$	$p = 0.001$
Travel	ns	$p = 0.006$	$p < 0.0001$	ns	ns	ns
Walking	ns	ns	ns	$p = 0.003$	$p = 0.003$	$p = 0.010$

trustworthiness, *authoritativeness*, *aboutness*, willingness to *return*, will *recommend*, previous visit, interest level in topic, and *highly rated* in relation to other sites assessed on the same topic. In addition, age, education, and Web browsing experience of the participant were loaded because Fogg et al. (2001) found some differences in ratings by these measures. After the first run, age, education, and browse experience were removed because of their low correlation (<0.05) with all other variables.

The Kaiser–Meyer–Olkin measure (0.828) and the Bartlett's Test of Sphericity ($\chi^2 = 15745.4$, $df = 28$, $p < 0.0001$) indicated that the sampling was adequate and that the measures were likely to be related. Factor analysis was conducted on the remaining eight measures using principal components analysis as the method of extraction and direct oblim as the method of rotation. Communality values tended to be high—all greater than 0.55; all variables loaded on 2 factors that together accounted for 62% of the variance.

Both factors had eigenvalues greater than one; the factor loadings are illustrated in Table 11 and the relationship between the measures and the factors are illustrated in Fig. 3. Factor 1 accounts for nearly 50% of the variance. All measures but the last two (interest and previous visit) correlated at greater than 0.71 with the first factor. This factor includes measures that are indicators of reputation, as established previously, and this variable is identified as a measure of repute. The second factor contains two highly correlated measures that are indicators of previous exposure—knowledge of site or interest in the topic. This factor accounts for only 12% of the variance. Six variables loaded heavily on factor 1 as illustrated in Fig. 3.

The level of interest in the topic and the fact that the site was previously visited were not tightly correlated with the other variables. The remaining six variables appear to be indicators of the perceived value of a Web site's content on a particular topic. Their internal reliability was measured using Cronbach's alpha, which, at 0.83, indicates that the variables, as a set, represent a unified construct.

Personal interest in a topic and the fact that a site was previously visited seem not to be related to user perceptions of the quality of that site. Thus, while previous exposure and interest had an

Table 11
Factor loadings

Variable	Factors	
	Repute	Prior knowledge
Recommend	0.844	−0.117
Will return	0.837	−0.074
Trustworthiness	0.800	0.066
Authoritative	0.796	0.097
Rated	0.737	−0.102
Aboutness	0.714	−0.231
Interest	0.104	0.790
Previous visit	0.294	0.644

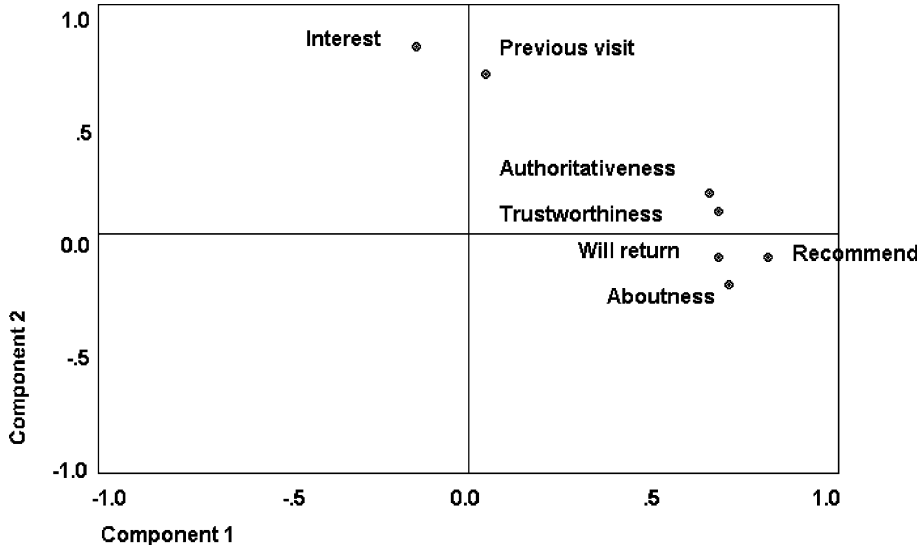

Fig. 3. Factors resulting from factor analysis of criteria.

effect on selected assessment ratings as discussed earlier, both appear not to be highly correlated with the *set* of assessments. From previous work, we believe that trustworthiness and authoritativeness (i.e., expertise) are good indicators of reputation (Fogg & Tseng, 1999; Jarvenpaa et al., 1999, 2000).

Selected user behaviours that represent key loadings in the analysis were a surprising result. Recommending a site and returning to a site seem to be major elements in people's perceived reputation of a site. From factor 1, a new variable, repute, was created using the regression quotient for the factor.

4.9. Perceived reputation

Using the metric, repute, as defined above, an analysis of variance assessed the differences among the tools. There was an interaction of tool and topic ($F(33, 4592) = 7.720$, $p < 0.001$) and

Table 12
Repute by search tool

	TOPIC	Google	Lycos	AltaVista	Significance
Alternative Energy	0.01	0.03	−0.20	−0.08	ns
Antiques	0.10	−0.22	−0.19	−0.15	ns
Gardening	−0.16	0.27	0.06	0.05	$p = 0.019$
Hockey	0.54	0.59	−0.70	0.46	$p < 0.0001$
Movie Reviews	0.14	−0.09	−0.48	−0.21	$p < 0.0001$
World News	−0.26	0.24	−0.01	−0.06	$p = 0.002$
Parenting	0.23	0.20	−0.11	0.12	$p = 0.037$
Parks Canada	0.42	0.47	0.42	0.23	ns
Politics Canada	−0.20	−0.12	0.36	−0.26	$p < 0.0001$
Sailing	−0.29	0.31	0.01	−0.11	$p < 0.0001$
Travel	0.02	−0.07	−0.39	−0.08	$p = 0.012$
Walking	−0.22	−0.01	0.01	−0.48	$p < 0.0001$
Mean	0.03	0.13	−0.10	−0.05	

a main effect of both tool ($F(3, 4592) = 15.366$, $p < 0.001$) and topic ($F(3, 4592) = 12.790$, $p < 0.001$). As shown in Table 12, there were significant differences among the tools. In pairwise comparisons, Google outputs sites that are perceived to be more reputable than the other three, and, in turn, TOPIC is also significantly different from AltaVista and Lycos, both of which do not differ from each other.

Table 12 also illustrates the differences between the tools by topic. In all but three topics, there were significant differences among the ability of the tools to output sites of repute for a topic. Not all tools performed at the same level for all topics. Google, the apparent key performer was not able to consistently output reputable sites. For some topics, such as *Movie Reviews* and *Parenting*, TOPIC outperformed Google, as was witnessed in the pilot study. Similarly Lycos, the apparent under-performer, was able to output sites that were perceived as more reputable than the other

Table 13
Subsets of topics with similar repute measures

Rank	Search topic	*N*	Subsets				
			1	2	3	4	5
12	Walking	383	x				
11	Movie Reviews	397	x	x			
10	Travel	397	x	x			
9	Antiques	393	x	x			
8	Alternative Energy	348	x	x	x		
7	Politics Canada	399	x	x	x		
6	World News	389	x	x	x		
5	Sailing	393	x	x	x		
4	Gardening	341		x	x	x	
3	Parenting	373			x	x	
2	Hockey	380				x	x
1	Parks Canada	400					x
Significance			0.548	0.063	0.185	0.384	0.24

three tools. Thus, overall, it appears that Google outputs the most reputable sites. However, this does not happen for each topic. Different tools perform differently according to the subject matter.

What were the differences by topic? The results of these simple comparisons appear in Table 13, which contains the topics ordered from lowest repute ranking overall to highest. In general, sites on the *Parks Canada* topic were perceived to be more reputable than those that dealt with *Walking*. The sites representing the topics *Hockey* and *Parks Canada* tended to be perceived as more reputable than most other topics in this set. In addition, Tukey's honestly significant difference test was used to assess pairwise comparisons between topics. Tukey was used because of the large number of comparisons that had to be made. Tukey's multiple range procedure identified five subsets of topics (see Table 13). Within each subset, repute does not differ significantly from topic to topic. Ideally, these subsets would have clustered according to some higher nameable category, but this was not the case with this data set.

5. Discussion and analysis

In this study, ordinary people assessed the quality of the output from four search tools, two of which were based on link structure analysis, and two on semantic content analysis. One of the link-based tools, TOPIC, purports to output reputable Web sites. The Web sites output by the four search engines were assessed using six metrics that were derived from quality assessments in other domains and were also considered to be indicators of reputation. In general, search tools based on link structure analysis were more successful than those that use primarily a semantic analysis technique. However, these differences were not as clear at the tool and topic levels. Assessment at the tool level was needed because the tools are created differently and for the most part, knowledge of their internal mechanisms is confidential.

In this study, tools using link structure analysis outperformed the tools using semantic content analysis on all measures. This observation was consistent with the pilot study conducted earlier. Because we do not know the details of how these individual tools work, we analyzed also at the tool level. Google and TOPIC consistently outperformed AltaVista and Lycos. Google also tended to outperform TOPIC while AltaVista tended to outperform Lycos. This finding was not the same in the pilot study, where TOPIC outperformed all the others on some measures. Differences, however, were only evident with aggregate data. When the data were analyzed by search topic, more discriminating differences showed up among the four tools. Google, for example, did not consistently output the most reputable Web sites and Lycos was not consistently the poorest performer. The range of differences by topic suggests that no single tool consistently retrieves the most reputable Web sites.

Reputation, as stated earlier, is elusive and non-explicit, but is an implication of quality. Unlike monetary value and information value, the value of a reputation cannot merely be acquired; a person or object does not hold a reputation until it is ascribed by a third party. It is a perception of value. Surprisingly, different tools and different sites received different results with different measures. There was little correlation among the results and, indeed, no clear patterns were discerned across the set of measures. We did not anticipate this at the outset. Because of the variation from measure to measure by tool, we used factor analysis to determine if one or more components underlay the metrics. In essence, could some of these metrics form a reputation

measure? The six metrics identified from the literature as having a contributory role in reputation were combined with selected user and use attributes. Factor analysis identified the original six metrics as a single component which we call "repute". Thus, this study has also shown that authoritativeness, trustworthiness, aboutness, recommend, re-visit and rank, together, form the basis of a measure of quality—a measure of reputation. Two of these metrics, trustworthiness and authoritativeness, also appear in Fogg et al.'s (2001) results. Furthermore, our findings also suggest reputation might be useful as an indirect measure for a user-centred assessment of search results—a quantitative approach to Wilson's (1983) information asymmetry.

Overall, the scores for Web sites were relatively low. In general, 3.5 was the average rating assigned on almost all variables. Only about half of the sites were considered worth recommending or re-visiting. Perhaps these assessments are indicative of the quality of Web sites in general. Our method examined only the top five Web sites in the hit list retrieved from each search tool, which one would surmise were the best sites for a topic. Surprising, also, is the average aboutness—3.7 on a five-point scale. One would hope that the aboutness of Web sites at these ranks would rate a higher degree of topicality. Did these four tools in fact uncover the best quality sites for these topics, or are the spiders simply not accessing the highest quality sites in all cases, or the indexers not ranking the best quality sites? This we cannot assess from our data.

6. Future work

In this study, people were asked to judge the quality of each Web site they were asked to examine. While the intention was to assess content, we do not know the extent to which the "window dressing" contributed to the evaluation. Weigelt and Camerer (1988) argue that game players who do not know a player's true type will perceive a reputation from cues that the player emits. In our case, did some of those cues come from design elements, and not content? Design quality varies substantially from Web site to Web site. To date, there are no known standards dictating design, but ever-increasing lists of guidelines are being produced. In a series of studies, Ivory, Sinha, and Hearst (2000, 2001) assessed the characteristics of Web sites that have received awards. They have identified over 100 quantitative characteristics and have developed a tool that can predict (around 90% successful) sites that likely would win an award based on their "window dressing". We believe that the physical manifestations of Web sites also likely influence the perceived reputation of a Web site. The degree to which participants used these external characteristics in their assessment is unknown, and the degree to which these characteristics factor into how Web sites are interlinked is also unknown. Furthermore, we wonder about the extent to which reputation is correlated with digital genre (Dillon & Vaughan, 1997). Are there culturally laden visual cues that genre emit (Toms & Campbell, 1999), that might, in turn, induce a perception of quality? It has been suggested elsewhere that information has shape (Dillon & Vaughan, 1997), and that shape is determined by genre (Toms, Campbell, & Blades, 1999). Does genre and shape contribute to a Web page's reputation?

Our future work will entail examining those visible cues present on a Web page. We are creating an inventory of the features of a Web page and plan to correlate the presence of those cues with the metric of reputation developed in this paper. In essence, how much do physical features contribute to reputation? In much the same way that physical appearance creates a first im-

pression of an individual, how much do the visible features on a page contribute to how a person perceives that page? If visual elements make a difference, then we wonder how these aspects might be factored into a search engine's algorithm so that search engines highly rank pages that people will perceive as reputable. This, of course, has caveats: like a person, a page may acquire an undeserving reputation.

7. Conclusions

We assessed the ability of four search tools to output reputable Web sites and know of no other work in which this particular attribute has been assessed. In general, output was perceived at just above neutral. Sites retrieved from Google were perceived to be of higher quality than the other three tools. TOPIC, which purports to output reputable Web sites, was a close second. In general, the link-based tools performed better than the semantic-based tools. But at the specific subject-matter level, no tool outperformed the others. Thus it would appear that no search engine can yet lay claim to outputting consistently high quality Web sites. In addition, in the process of doing this study, we created a metric to measure user perceptions of reputation, one that could be used for other purposes.

Predicting the reputation of a Web site is a complex problem. TOPIC's approach to measuring reputation is novel and clearly needs some tweaking. The findings from this work are promising, but suggest a more comprehensive study that adds another dimension to the analysis: the physical characteristics of Web sites and how these impact user perceptions. Adding this factor into TOPIC might indeed enable it to outperform Google.

Acknowledgements

We thank the Faculty of Information Studies for the use of its facilities; the 80 participants; Alberto Mendelzon for providing funding for this project; and Joan Cherry and Greg Keast who worked on the pilot study. We also thank the anonymous reviewers for their very helpful comments and suggestions.

References

About.com. (2001a). *About—the human Internet*. Available: http://www.about.com.

About.com. (2001b). *About—our story*. Available: http://ourstory.about.com/index.htm?PM = 59_1100_T.

Alexa. (2001). *The Web Information Company*. Available: http://www.alexa.com.

Allen, F. (1984). Reputation and product quality. *The RAND Journal of Economics, 15*(3), 311–327.

Amazon. (2002). *Amazon.com—earth's biggest selection*. Available: http://www.amazon.com.

Argus Associates. (2001). *The Argus clearinghouse*. Available: http://www.clearinghouse.net.

Bailey, B. P., Gurak, L. J., & Konstan, J. A. (2001). An examination of trust production in computer-mediated exchange. In *Proceedings of the 7th conference on human factors and the Web*. Available: http://www.optavia.com/hfweb/7thconferenceproceedings.zip/Bailey.pdf.

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

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *7th international World Wide Web conference, Brisbane, Australia*. Available: http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm.

Chen, M., & Singh, J. P. (2001). Computing and using reputations for Internet ratings. In M. P. Wellman & Y. Shoham (Eds.), *Proceedings of the third ACM conference on electronic commerce* (pp. 154–162). New York: ACM.

Chu, W., & Chu, W. (1994). Signalling quality by selling through a reputable retailer: an example of renting the reputation of another agent. *Marketing Science, 13*(2), 177–189.

Dellarocas, C. (2001). Analyzing the economic efficiency of ebay-like online reputation reporting mechanisms. In M. P. Wellman & Y. Shoham (Eds.), *Proceedings of the third ACM conference on electronic commerce* (pp. 171–179). New York: ACM.

D'Esposito, J. E., & Gardner, R. M. (1999). University students' perceptions of the Internet: an exploratory study. *The Journal of Academic Librarianship, 25*(6), 456–461.

Dillon, A., & Vaughan, M. W. (1997). It's the journey and the destination: shape and the emergent property of genre in digital documents. *New Review of Multimedia and Hypermedia, 3*, 91–106.

Doby, V. J., & Caplan, R. D. (1995). Organizational stress as threat to reputation: effects on anxiety at work and at home. *The Academy of Management Journal, 38*(4), 1105–1123.

eBay. (2002). *eBay—the world's online marketplace*. Available: http://www.ebay.com.

Epinions. (2001). *Epinions.com—welcome*. Available: http://www.epinions.com/.

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Trangnekar, A., Shon, J., Preeti, S., & Treinen, M. (2001). What makes Web sites credible? A report on a large quantitative study. In J. Jacko & A. Sears (Eds.), *Proceedings of the ACM CHI 2001 conference on human factors in computing systems* (pp. 61–68). New York: ACM.

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In M. Williams & M. Altom (Eds.), *Proceedings of ACM CHI 99 conference on human factors in computing systems* (pp. 80–87). New York: ACM.

Fombrun, C., & Shanley, M. (1990). What's in a name? Reputation building and corporate strategy. *The Academy of Management Journal, 33*(2), 233–258.

Fombrun, C. J. (2001). *Reputations: measurable, valuable, and manageable*. American Banker Online, May 25, 2001. Available: http://www.americanbanker.com/.

Hill, C. W. L. (1990). Cooperation, opportunism, and the invisible hand: implications for transaction cost theory. *The Academy of Management Review, 15*(3), 500–513.

Hosmer, L. T. (1995). Trust: the connecting link between organizational theory and philosophical ethics. *The Academy of Management Review, 20*(2), 379–403.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. New Haven: Yale University Press.

Ivory, M. Y., Sinha, R. R., & Hearst, M. A. (2000). Preliminary findings on quantitative measures for distinguishing highly rated information-centric Web pages. In *Proceedings of the 6th conference on human factors and the Web*. Available: http://www.tri.sbc.com/hfweb/ivory/paper.html.

Ivory, M. Y., Sinha, R. R., & Hearst, M. A. (2001). Empirically validated Web page design metrics. In J. Jacko & A. Sears (Eds.), *Proceedings of the ACM CHI 2001 conference on human factors in computing systems* (pp. 53–60). New York: ACM.

Jadad, A. R., & Gagliardi, A. (1998). Rating health information on the Internet. *Journal of the American Medical Association, 279*(8), 611–614.

Jarvenpaa, S. L., Tractinsky, N., Saarinen, L., & Vitale, M. (1999). Consumer trust in an Internet store: a cross-cultural validation. *The Journal of Computer-Mediated Communication, 5*(2). Available: http://www.ascusc.org/jcmc/vol5/issue2/jarvenpaa.html.

Jarvenpaa, S. L., Tractinsky, N., & Vitale, M. (2000). Consumer trust in an Internet store. *Information Technology Management, 1*(1–2), 45–71.

Keast, G., Toms, E. G., & Cherry, J. (2001). Measuring the reputation of Web sites: a preliminary exploration. In E. A. Fox & C. L. Borgman (Eds.), *Proceedings of first ACM-IEEE-CS joint conference on digital libraries* (pp. 77–78). New York: ACM.

Klang, M. (2001). Who do you trust? Beyond encryption, secure e-business. *Decision Support Systems, 31*(3), 293–301.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery, 46*(5), 604–632.

Kollock, P. (1994). The emergence of exchange structures: an experimental study of uncertainty, commitment, and trust. *American Journal of Sociology, 100*(2), 313–345.

Lee, J., Kim, J., & Moon, J. Y. (2000). What makes Internet users visit cyber stores again? Key design factors for customer loyalty. In T. Turner & G. Szwillus (Eds.), *Proceedings of the SIG CHI conference on human factors in computing systems* (pp. 305–312). New York: ACM.

Lynch, P. D., Kent, R. J., & Srinivasan, S. S. (2001). The global Internet shopper: evidence from shopping tasks in twelve countries. *Journal of Advertising Research, 41*(3), 15–23.

Media metrix. (2001). *Media metrix global landing*. Available: http://www.mediametrix.com.

Nielsen, J. (1999). *Reputation managers are happening*. Available: http://www.useit.com/alertbox/990905.html.

Netratings. (2001). *Nielsen//Netratings*. Available: http://www.netratings.com.

Netscape. (2001). *ODP—open directory project*. Available: http://www.dmoz.org.

Rafiei, D., & Mendelzon, A. O. (2000). What is this page known for? Computing Web page reputation. *Computer Networks, 33*(6), 823–835.

Reagle, J. M. (1996). Trust in electronic markets: the convergence of cryptographers and economists. *First Monday, 1*(2). Available: http://www.firstmonday.dk/issues/issue2/markets/index.html.

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000). Reputation systems. *Communications of the Association for Computing Machinery, 43*(12), 45–48.

Rieh, S. Y., & Belkin, N. J. (1998). Understanding judgment of information quality and cognitive authority in the WWW. In C. M. Preston (Ed.), *Proceedings of the 61st annual meeting of the american society for information science, 35* (pp. 279–289). Medford, NJ: Information Today.

Schurr, P. H., & Ozanne, J. L. (1985). Influences on exchange processes—buyers perceptions of a sellers trustworthiness and bargaining toughness. *Journal of Consumer Research, 11*(4), 939–953.

Science Academy. (2001). *How to find information on the Web*. Available: http://www.scienceacademy.com/find.html.

Shapiro, C. (1982). Consumer information, product quality, and seller reputation. *The Bell Journal of Economics, 13*(1), 20–35.

Simpson, J. A., & Weiner, E. S. C. (1989). *Oxford English dictionary* (2nd ed.). Oxford: Clarendon Press.

Sridhar, S. S. (1994). Managerial reputation and internal reporting. *The Accounting Review, 69*(2), 343–363.

Toms, E. G., & Campbell, D. G. (1999). Genre as interface metaphor: exploiting form and function in digital environments. In *Proceedings of the Hawaii international conference on system sciences, January 5–8, 1999, Maui (HICSS-32)*. IEEE (on CDROM).

Toms, E. G., Campbell, D. G., & Blades, R. (1999). Does genre define the shape of information: the role of form and function in user interaction with digital documents. In L. Woods (Ed.), *Proceedings of the 62nd annual meeting of the American society for information science* (pp. 693–704). Medford, NJ: Information Today.

Wathen, C. N., & Burkell, J. (2002). Believe it or not: factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology, 53*(2), 134–144.

Weigelt, K., & Camerer, C. (1988). Reputation and corporate strategy: a review of recent theory and applications. *Strategic Management Journal, 9*(5), 443–454.

Whitmeyer, J. M. (2000). Effects of positive reputation systems. *Social Science Research, 29*(2), 188–207.

Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Westport, CT: Greenwood Press.

Yahoo. (2001). *Yahoo!* Available: http://www.yahoo.com.