Are Web Based Informational Queries Changing?

Chadwyn Tann
Department of Information Studies,
University of Sheffield,
Regent Court, 211 Portobello St,
Sheffield, S1 4DP, UK
Tel: +44 114 22 22648
Fax: +44 114 27 80300
Email: m.sanderson@shef.ac.uk

Mark Sanderson
Department of Information Studies,
University of Sheffield,
Regent Court, 211 Portobello St,
Sheffield, S1 4DP, UK
Tel: +44 114 222 2630
Fax: +44 114 278 0300

Abstract

This brief communication describes the results of a questionnaire examining certain aspects of the web-based information seeking practises of University students. The results are contrasted with past work showing that queries to web search engines can be assigned to one of a series of categories: namely navigational, informational and transactional. The survey results suggest that a large group of queries, which in the past, would have been classified as informational have become at least partially navigational. We contend that this change has occurred because of the rise of large web sites holding particular types of information, such as Wikipedia and IMDB.

Introduction

In information retrieval research prior to the rise of the web, queries were by and large assumed to be of one type: namely a query intended to seek information. Researchers such as Salton (Salton, 1968) and Van Rijsbergen (van Rijsbergen, 1979) described how to rank documents relative to queries, but said nothing about the potential for different types of queries being submitted to an IR system and the impact such different forms might have on the ranking algorithms they described.

This consistent view of search persisted until web search engines emerged and the searches that people issued to the engines were studied in query logs. It was then that distinct forms of search were identified. The pioneering work in this area was conducted by Broder who wrote about it several years later (Broder, 2002), Broder stated that queries to search engines could be classified into three forms.

- **Navigational** – queries where users wish to locate home pages. Broder gave examples, such as *Greyhound Bus, national car rental, american airlines, Don Knuth* as queries where users wished to locate a home page of those organisations or people.
- **Informational** – queries where users was seek web pages containing information on a particular topic. Broder described these queries as being "closest to classic IR".
- **Transactional** – queries where users seek web sites where they could conduct a transaction: e.g. music downing, online shopping, or other databases of information not indexed by the search engine.

To confirm his hypothesis that queries fitted into the three categories, Broder conducted a questionnaire of search engine users and analysed query logs. He also determined the balance between the different query types. He found that between 39% and 49% of queries were informational; 20%-24.5% were navigational and 30%-36% were transactional. Broder pointed out that in order to serve these queries accurately, it was necessary to apply different search algorithms to each. Methods such as matching on anchor text (Craswell, Hawking, & Robertson, 2001), Broder stated, were "*a very effective help for navigational queries*" (p. 8), for example. Authority measures such as PageRank (Brin & Page, 1998) are also commonly used.

Broder's analysis is one of the most discussed in the literature, with his categories further adapted, altered and sub-divided. Rose & Levinson (2004), for example, sub-divided informational queries into five categories (directed, undirected, advise, locate, list); re-named transactional queries to *Resource* queries and broke them into four categories (download, entertainment, interact, obtain). The researchers also stated that based on their analysis, navigational queries were less frequent than was previously thought, numbering between 12% and 15%. Jansen, Booth & Spink (2008) surveyed past work in this area and developed a method for automatically detecting the three types of query. They broadly agreed with Rose and Levinson in finding that around 10% of queries were navigational, 10% resource and 80% informational. Jansen et al characterized informational queries as follows:

"*The intent of informational searching is to locate content concerning a particular topic in order to address an information need of the searcher. The content can be in a variety of forms, including data, text, documents, and multimedia. The need can be along a spectrum from very precise to very vague.*"

Many have looked to categorize the majority informational queries into a number of subject areas: Jansen et al (Jansen, Spink, & Saracevic, 2000) initially identified broad areas (e.g. business, entertainment, adult) and later examined a number in more detail (Spink, Jansen, Wolfram, & Saracevic, 2002). Other sub-areas that have received much attention include queries with a spatial component (Sanderson & Kohler, 2004); (Zhou, Xie, Wang, Gong, & Ma, 2005). In these various cited studies, little was written about users' desire for informational queries to be served by authoritative web pages. However, (Bharat & Henzinger, 1998) showed for a set of informational queries (they referred to as *topic distillation*) that users prefer documents that are both authoritative and relevant.

With the rise of certain Web 2.0 sites (O'Reilly, 2005), there has been a trend towards information of certain forms coalescing into a small number of large web sites. Sites such as Wikipedia for overviews of general knowledge and the Internet Movie Database (IMBD) for film and television information are well known and well used. Given the rise of such sites, it was hypothesized that users of search engines, when seeking a particular piece of information may not only expect to find information from an authoritative source, but also expect to be shown their sought information from one of the large web sites.

For example, a user entering the query "George W. Bush" into a search engine, might not only want to find information on the 43rd president of the USA, but may explicitly wish see the Wikipedia version of that information, not for example the version on the whitehouse.gov web site. A user searching for information about the film "Jaws", might be expecting to see that information on the IMDB web site as opposed to the film company that produced the film. Such preferences might extend to a broad range of topics. If such preferences were real, a large number of what are generally regarded as informational queries could in fact be viewed as a form of navigational query.

An examination of retrieval results from any number of informational queries reveals that search engines, such as Google and Yahoo![1], commonly place pages from well known sites, such as Wikipedia and IMDB at or near the top of search rankings. However, examining ranks on their own is not a sufficient indication of the strength of users feelings about wishing their information served from particular sources. Therefore, we present a short set of results that were part of a wider survey of users to understand their searching preferences. The specific research question to be answered by this study was

"*Do users have a specific web site in mind when they are searching for particular types of informational queries?*"

The rest of this brief communication details our work. It starts with a short introduction to the questionnaire used to address this question and the population the questionnaire was sent to. Next, the results of the part of the questionnaire that addresses the research question is described, which is followed by conclusions and future work.

Questionnaire and population

The questionnaire was part of a wider study into information seeking behavior of students conducted at the University of Sheffield. The full study is described by (Tann, 2007). In designing this part of the questionnaire, the aim was to ascertain if users had a specific web site in mind when searching for particular forms of information. To achieve the aim, users were asked what site they started their search on for a number of query types. Because a large number of users were expected to answer with the name of their favorite generic web search engine, users were also asked what web site they expected to end up using as the source for their information need. Three domains of information searching were addressed: generic factual information, entertainment information and technological information.

The questionnaire was put online and students at the university were invited, via an email list, to fill it in. In total 4,482 students were contacted, 220 (5.0%) filled in valid responses. Of the respondents, 126 were female, 94 were male; 147 were 18-21 years old, 67 were 22-25, 5 were 26-30, 1 was over 40. All but one of the respondents were students taking a bachelors degree. The other respondent was studying for a PhD.

[1] Search engines examined in January 2009

The students were asked about their search engine experience: 152 used search engines everyday; 48 used them 6-4 times a week and 20 used search engines 1-3 times a week. When asked if there was a particular search engine they preferred, 211 (96%) choose Google.

Do users seek a specific site when information searching?

The students were asked "which is the first source you use when searching for **factual** information" (emboldening added). The majority (138, 63%) responded that they used Google, 74 (34%) said they chose Wikipedia as their primary source, 8 (3%) stated another source. In a follow up question, users were asked "Which source would you expect to end up using [to serve their information need]?". Of the 213 who answered (97% of the 220), 106 (50%) said they would use Wikipedia, 107 (50%) stated they would use another source.

Users were also asked "which is the first source you use when searching for **entertainment** information": 127 (58%) indicated Google, 62 (28%) indicated IMDB, 14 (6%) indicated Wikipedia and 17 (8%) indicated others. When asked, "Which source would you expect to end up using?" out of the 220 respondents, 100 (45%) chose IMDB, 65 (30%) chose Wikipedia and 55 (25%) chose other sources.

Finally, users were asked, "which is the first source you use when searching for **technology** information", here 169 (76%) chose Google, 27 (12%) chose Wikipedia, 18 (8%) chose "others", 6 (3%) chose the technology site CNET. In the question on which site they expected to end up using 121 (55%) of the respondents chose Wikipedia, 69 (31%) chose "other" and 30 (14%) chose CNET.

The results of this part of the survey are summarized in Table 1 and Table 2. A common pattern that can be seen across all three questions is that the majority of users said they would start their search on Google, but regardless of how they initiated their search, for all three types of information, many users had one of a limited number of web sties in mind as their eventual source. For technology and factual information, around half of the users expected to use Wikipedia; for entertainment, nearly half expected to use IMDB, many of the others expected to use Wikipedia. The responses would indeed seem to offer strong support for the research question of this study: regardless of what other content there is on the web, users have developed an expectation of where they will find the information they seek for a large number of information types.

A second striking result from these questions was that although the majority of users often started their search with Google, a large number of users started with other sources. For factual information, 34% said that they started their search on Wikipedia, 29% said they started at IMDB for entertainment information. Note, that when asked which was their common search engine for searching in general, 96% of respondents replied with Google. However, when the same respondents were asked about preferences for searching specific types of information, the answers were more nuanced.

Type	Google	Wikipedia	Other	IMDB	CNET
Factual	62	34	4	-	-
Entertainment	58	7	5	29	-
Technology	76	13	8	-	3

Table 1: Which is the first source used when searching? Percentages shown.

Type	Wikipedia	Other	IMDB	CNET
Factual	50	50	-	-
Entertainment	30	25	45	-
Technology	55	31	-	14

Table 2: Which source do you expect to end up using? Percentages shown.

When the users who selected Wikipedia were asked from a list of pre-defined responses why they chose this source, a common response was that the information in Wikipedia was "in-depth enough" for factual

information queries (58% of respondents), for entertainment queries (46%) and for technology queries (52%). Another reason, "Finding information on Wikipedia is more convenient", was chosen by 52% of the respondents for factual, 41% for entertainment, 45% for entertainment.

Conclusions and future work

This brief communication surveyed past research on categorizing web queries. Broder was the first to describe such work and over the ensuing years his three broad categories continued to be used. It can be generally observed that search engines like Google, rank highly the matching pages of web sites, such as Wikipedia or IMDB, for a wide range of informational queries. It would appear that this is being done because for those queries, many users have a clear intention to obtain their information from these particular sites; so much so, that a substantial minority of respondents choose to start their searches directly from those sites. (Choo, Detlor, & Turnbull, 2000, p. 12) showed that in the information seeking literature, *accessibility* was identified as a key factor in users' preferences for selecting sources, citing (Gerstberger & Allen, 1968) work on this topic. The questionnaire respondents' stated that Wikipedia and IMDB were convenient and held a sufficient level of information to address their information need. It would appear that to these users, the sites were expected to be accessible.

Jansen et al stated that the "*…intent of navigational searching is to locate a particular Website. … It can be a particular Web page, site or a hub site. The searcher may have a particular Website in mind, or the searcher may just 'think' a particular Website exists."* Similarly Rose and Levinson declared that "*To be considered navigational, the query must have a single authoritative web site that the user already has in mind. For this reason, most queries consisting of names of companies, universities, or well-known organizations are considered navigational. Also for this reason, most queries for people – including celebrities – are not."*

Based on such definitions and the results from the survey of user intentions, one should consider if many of the informational queries submitted to search engines should be thought of as at least partially navigational, where searchers expect their need served by a specific large web site, which they know holds the type of information they seek. Although (Bharat & Henzinger, 1998) showed the importance of ensuring the results of informational searches were authoritative, they did not suggest that a limited number of web sites would become authoritative for broad classes of queries, which the results of the questionnaire seem to suggest. However, the survey only points to indications of this type of navigational information searching. Users were asked about their *intentions* when searching, their actual *querying behavior* was not observed.

What we can conclude is that the results of this survey support the need for a follow up study to examine user searching in the identified domains. It would also be valuable to expand the study to a wider population and examine in more detail the types of informational queries where users have a common predefined source in mind and understand further the reasons for these stated preferences.

Acknowledgements

We wish to thank Paul Clough for his valuable comments and suggested references. We also wish to thank the reviewers for their thoughts and comments also.

References

Bharat, K., & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 104-111). ACM New York, NY, USA.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1-7), 107-117.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3-10. doi: 10.1145/792550.792552.

Choo, C. W., Detlor, B., & Turnbull, D. (2000). *Web Work: Information Seeking and Knowledge Work on the World Wide Web*. Kluwer Academic Publishers.

Craswell, N., Hawking, D., & Robertson, S. (2001). Effective Site Finding using Link Anchor Information. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 250-257). New Orleans, Louisiana, USA: Assn for Computing Machinery.

Gerstberger, P. G., & Allen, T. J. (1968). Criteria used by research and development engineers in the selection of an information source. *J Appl Psychol*, *52*(4), 272-9.

Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing and Management*, *44*(3), 1251-1266.

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.

O'Reilly, T. (2005). What is Web 2.0. *Design Patterns and Business Models for the Next Generation of Software*, *30*, 2005.

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed., p. 224). Butterworth-Heinemann Ltd.

Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web* (pp. 13-19). ACM Press New York, NY, USA.

Salton, G. (1968). *Automatic Information Organization and Retrieval.* McGraw Hill Text.

Sanderson, M., & Kohler, J. (2004). Analyzing geographic queries. In *SIGIR Workshop on Geographic Information Retrieval, held at the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 1-2).

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *Computer*, *35*(3), 107-109.

Tann, C. (2007). *Users' searching behaviour patterns of Wikipedia and Google*. Undergraduate dissertation, Department of Information Studies, University of Sheffield, Sheffield, S10 2TN, UK.

Zhou, Y., Xie, X., Wang, C., Gong, Y., & Ma, W. Y. (2005). Hybrid index structures for location-based web search. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 155-162). ACM New York, NY, USA.