A taxonomy of web search

Andrei Broder
IBM Research
broder@us.ibm.com

(Most of the work presented here was done while the author was with the AltaVista corporation.

Abstract:

Classic IR (information retrieval) is inherently predicated on users searching for information, the so-called "information need". But the need behind a web search is often not informational -- it might be navigational (give me the url of the site I want to reach) or transactional (show me sites where I can perform a certain transaction, e.g. shop, download a file, or find a map). We explore this taxonomy of web searches and discuss how global search engines evolved to deal with web-specific needs.

1 Introduction

A central tenet of classical information retrieval is that the user is driven by an information need. Schneiderman, Byrd, and Croft [SBC97] define information need as "the perceived need for information that leads to someone using an information retrieval system in the first place." But the intent behind a web search is often not informational -- it might be navigational (show me the url of the site I want to reach) or transactional (show me sites where I can perform a certain transaction, e.g., shop, download a file, or find a map). In fact as we show later, informational queries constitute less than 50% of web searches.

The main aim of this paper is to point out this difference and introduce and analyze a taxonomy of web searches. Secondly, we show how search engines evolved to deal with these web-specific needs.

The remainder of the paper is organized as follows: in section 2 we discuss the classic model for information retrieval; section 3 introduces a taxonomy of web searches; section 4 presents some statistics extracted from AltaVista surveys and logs regarding the prevalence of various types of searches; section 5 analyzes the evolution of search engines in light of this taxonomy; section 6 discusses some related work; finally, section 7 draws certain conclusions and points to some further directions for research.

2. The classic model for information retrieval

We start from the basic model used in many standard information retrieval references textbooks, for instance, van Rijsbergen [R79]. See also [BK94] and references therein for a detailed discussion.

Essentially, a user, driven by an information need, constructs a query in some query language. The query is submitted to a system that selects from a collection of documents (corpus), those documents that match the query as indicated by certain matching rules. A query refinement process might be used to create new queries and/or to refine the results. (Figure 1)

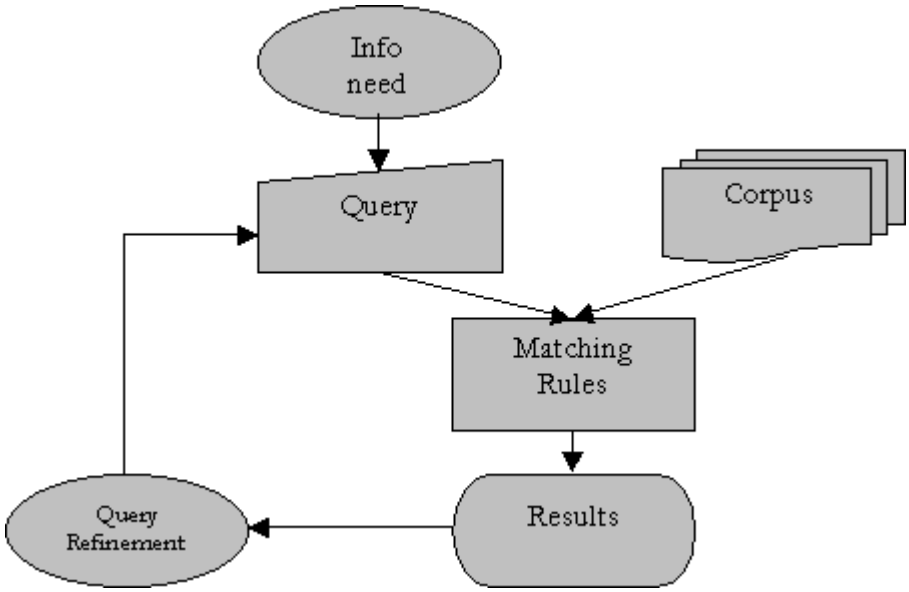

Figure 1. The classic model for IR.

Since in the web context the human-computer interaction factors and the cognitive aspects play a significant role, it is useful to detail this model further as in Figure 2.

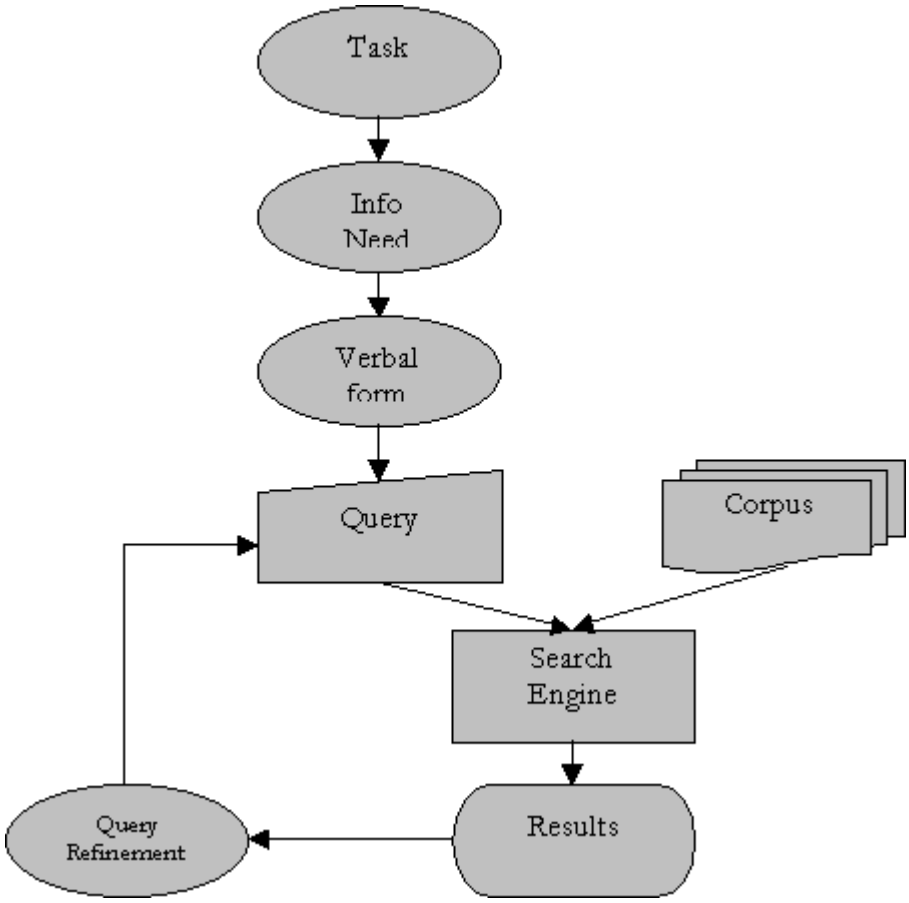

Figure 2. The classic model for IR, augmented for the web.

Thus we recognize that the information need is associated with some task. This need is verbalized (usually mentally, not loud) and translated into a query posed to a search engine. This process of deriving a query from an information need in the web context has received a great deal of attention: Holscher and Strube [HS00] point out that experienced and novice users construct searches

differently, Navarro-Pietro et al. [NSR99] derived a cognitive model for web search, Muramatu and Pratt [MR01] explore users' mental model of search engines, etc. See also [CDT99]. However, all this literature shares the assumption that web searches are motivated by an information need.

3. A taxonomy of web searches

In the web context the "need behind the query" is often not informational in nature. We classify web queries according to their intent into 3 classes:

1. **Navigational.** The immediate intent is to reach a particular site.
2. **Informational.** The intent is to acquire some information assumed to be present on one or more web pages.
3. **Transactional.** The intent is to perform some web-mediated activity.

Before we discuss these types in detail, we need to clarify that there is no assumption here that this intent can be inferred with any certitude from the query. The examples below might have alternative explanations.

Navigational queries. The purpose of such queries is to reach a particular site that the user has in mind, either because they visited it in the past or because they assume that such a site exists. Some examples are

- `Greyhound Bus`. Probable target http://www.greyhound.com
- `compaq`. Probable target http://www.compaq.com.
- `national car rental`. Probable target http://www.nationalcar.com
- `american airlines home`. Probable target http://www.aa.com
- `Don Knuth`. Probable target http://www-cs-faculty.stanford.edu/~knuth/

This type of search is sometimes referred as "known item" search in classical IR, but is mostly used in the evaluation of various systems. The TREC-2001 (Text Retrieval Conference) web track however has a "[web] Home Page Finding Task" based on 145 queries against the WT10g collection, distributed by CSIRO. (See http://www.ted.cmis.csiro.au/TRECWeb/guidelines_2001.html) These types of queries are essentially navigational queries.

With respect to evaluation, navigational queries have usually only one "right" result (up to syntactic/semantic aliases). For instance on the query `haaretz` (the name of an Israeli newspaper) the target results is probably one of

- www.haaretz.co.il (Hebrew edition)
- www2.haaretz.co.il/breaking-news/ (English edition served from Israel)
- www.haaretzdaily.com (US mirror of English edition)

"Hub" type results that are one click away from target may be acceptable but less desirable. Continuing with our example, on the query `haaretz`, a list of Israeli newspapers might be acceptable.

Informational queries. The purpose of such queries is to find information assumed to be available on the web in a *static form*. No further interaction is predicted, except reading. By *static form* we mean that the target document is not created in response to the user query. This distinction is somewhat blurred since the blending of results characteristic to the third generation search (see Section 5) engines might lead to dynamic pages.

In any case, informational queries are closest to classic IR and therefore need less attention here.

What is different on the web is that many informational queries are extremely wide, for instance `cars` or `San Francisco`, while some are narrow, for instance `normocytic anemia`, `Scoville heat units`. It is interesting to note, that in almost 15% of all searches the desired target is a good collection of links on the subject, rather than a good document. (A good hub, rather than a good authority, in the language of Kleinberg [K98]). See survey results below.

Transactional queries. The purpose of such queries is to reach a site where further interaction will happen. This interaction constitutes the transaction defining these queries. The main categories for such queries are shopping, finding various web-mediated services, downloading various type of file (images, songs, etc), accessing certain data-bases (e.g. Yellow Pages type data), finding servers (e.g. for gaming) etc.

The results of such queries are very hard to evaluate in terms of classic IR. Binary judgment might be all we have, say appropriate, non-appropriate. However most external factors important for users (e.g. price of goods, speed of service, quality of pictures, etc) are usually unavailable to generic search engines.

4. Statistics

We used two methods to determine the prevalence of various types of queries: a survey of AltaVista users, and an analysis of the query log at AltaVista.

User survey

For the user survey we used an interstitial survey window. (A "pop-up" window.)

The survey window was presented to random users and achieved a response ratio of about 10%. The data discussed here was collected between June 26 and November 3, 2001 and consisted of 3190 valid returns. The survey questions relevant to this paper were as follows.

2. Which of the following describes best what you are trying to do?

- ◯ I want to get to a specific website that I already have in mind
- ◯ I want a good site on this topic, but I don't have a specific site in mind

3. Which of the following best describes why you conducted this search?

- ◯ I am shopping for something to buy on the Internet
- ◯ I am shopping for something to buy elsewhere than on the Internet
- ◯ I want to download a file (e.g., music, images, programs, etc.)
- ◯ None of these reasons

4. Which of the following describes best what you are looking for?

- ◯ A site which is a collection of links to other sites regarding this topic
- ◯ The best site regarding this topic

Figure 3. Survey questions.

We note several salient aspects of the survey.

- The users are self selected. The self-selection might be influenced by many factors, among

them the users attitude towards interstitial windows. According to [JJ01] 83% of the users view such pop-ups as interference.

- As a consequence of self selection, the percentage of queries with sexual content is under 1%. In contrast, the percentage of such queries in the log is about 12%.
- Q2 is used to distinguish between navigational and non-navigational queries. The percentage of queries identified as navigational was 24.5%, non-navigational queries accounted for 68.4%, and 7.1% of the surveys did not answer Q1. Thus among respondents to Q2, the percentage of navigational queries was 26.4%.
- However, we could not find a simple question to distinguish between transactional and informational queries. Instead we identified some of the most popular transactional queries, namely shopping on the web (Q3.1) and downloading data (Q3.3). (Note that if the intent is shopping *off the web* the query is probably informational rather than transactional). Among people that answered both Q2 and Q3 these type of queries account for 32.3% of the non-navigational queries. (Shopping on the web 7.65% and download 24.65%.) Thus the total number of transactional queries is at least 23.8%.

  The survey had some additional questions, in particular Q7 was *In your own words, please describe the exact piece of information you are seeking*. Based on sample of 200 queries, the query text, and the explanation provided at Q7, we estimate however that the number of transactional queries among survey respondents is about 36%.
- Queries that are neither transactional, nor navigational, are assumed to be informational.
- The overall results of the survey are presented in figure 4.

2. Which of the following describes best what you are trying to do?

24.53% I want to get to a specific website that I already have in mind

68.41% I want a good site on this topic, but I don't have a specific site in mind

3. Which of the following best describes why you conducted this search?

8.16% I am shopping for something to buy on the Internet

5.46% I am shopping for something to buy elsewhere than on the Internet

22.55% I want to download a file (e.g., music, images, programs, etc.)

57.19% None of these reasons

4. Which of the following describes best what you are looking for?

14.83% A site which is a collection of links to other sites regarding this topic

76.62% The best site regarding this topic

Figure 4. Survey answers.

Log analysis

Since inferring the user intent from the query is at best an inexact science, but usually a wild guess, the data obtained from log analysis is very "soft". We selected at random a set of 1000 queries from the daily AltaVista log. From this set we removed non-English queries and sexually oriented queries. (The later being about 10% of the English queries). From the remaining set the first 400 queries were inspected. Queries that were neither transactional, nor navigational, were assumed to be

informational in intent.

Type of query	User Survey	Query Log Analysis
Navigational	24.5%	20%
Informational	?? (estimated 39%)	48%
Transactional	> 22% (estimated 36%)	30%

Figure 5. Query classification.

5. The evolution of search engines

In view of the taxonomy discussed so far we identify three stages in the evolution of web search engines:

- First generation -- uses mostly on-page data (text and formatting) and is very close to classic IR. Supports mostly informational queries. This was state-of-the art around 1995-1997 and was exemplified by AltaVista, Excite, WebCrawler, etc.
- Second generation -- use off-page, web-specific data such as link analysis, anchor-text, and click-through data. This generation supports both informational and navigational queries and started in 1998-1999. Google was the first engine to use link analysis as a primary ranking factor and DirectHit concentrated on click-through data. By now, all major engines use all these types of data. Link analysis and anchortext seems crucial for navigational queries.
- Third generation -- emerging now, attempts to blend data from multiple sources in order to try to answer ``the need behind the query". For instance on a query like `San Francisco` the engine might present direct links to a hotel reservation page for San Francisco, a map server, a weather server, etc. Thus third generation engines go beyond the limitation of a fixed corpus, via semantic analysis, context determination, dynamic data base selection, etc. The aim is to support informational, navigational, and transactional queries. This is a rapidly changing landscape.

6. Related work

There is an fairly rich literature dealing with navigation on the web, but it is mostly concerned with navigation via links on page, bookmarks, and the browser "back" button. See Tauscher and Greenberg [TG97], Cockburn and Jones [CJ96], and Catledge and Pitkow [CP95]. Navigation ("Go to page") is identified as on of the tasks in the "Taskonomy" (sic) of WWW use invented by Byrne at al. [BJWC99] However, search as a navigation tool has not been scrutinized much so far, maybe because search engines have become proficient at this only in recent years.

David Hawking and his colleagues have conducted a series of web search engines evaluations, in particular "Which search engine is best at finding airline home pages?" [CHG01] which is of course a representative navigational task, and "Which search engine is best at finding online services?" [HCG01] which is a transactional task. Hawking et al. also show that anchor text provides a very effective help for navigational queries [CHR01]

7. Conclusions and future directions

On the web "the need behind the query" might be

- Informational
- Navigational
- Transactional

Search engines need to deal with all three types although each type is best satisfied by very different results. An understanding of this taxonomy is essential to the development of successful web search. Current search engines deal well with informational and navigational queries, but transactional queries are satisfied only indirectly and hence a third generation in search engines is emerging: its main aim is to deal efficiently with transactional queries mostly via semantic analyses (understanding what the query is about) and blending of various external data bases. This is motivated both by the user needs and the fact that transactional queries are probably the easiest to monetize.

Acknowledgment

This paper has grown out of an invited talk at TREC in November 2000 on "Web search quality vs. informational relevance" and another talk at the Search Engine Meeting in April 2001 on "The Need behind the Query". Many colleagues have provided valuable input. I wish to thank them all and in particular Bob Travis and Danny Levinson who programmed the survey described here.

References

[**SBC97**] B. Schneiderman, D. Byrd, and W. B. Croft. Clarifying search: A user-interface framework for text searches. *D-Lib Magazine*, January 1997. Available at http://www.dlib.org/dlib/january97/retrieval/01shneiderman.html).

[**JJ01**] J. Jackson. Pop Goes the Interstitial. *eMarketeer*, 7 June 2001, Available at http://www.emarketer.com/analysis/eadvertising/20010607_ead.html

[**BJWC99**] Byrne, M. D., John, B. E., Wehrle, N. S., and Crow, D. C. The tangled web we wove: A taskonomy of WWW use. In *Human Factors in Computing Systems: Proceedings of CHI 99* (pp. 544-551), Addison Wesley, 1999.

[**R79**] van Rijsbergen, C. J. Information Retrieval. London: Butterworths, 1979. Available at http://www.dcs.gla.ac.uk/Keith/Preface.html.

[**BK94**] M. K. Buckland and C. Plaunt. On the construction of Selection Systems. In *Library Hi Tech*, 12(4), 1994.

[**HS00**] C. Holscher and G. Strube. Web search behaviour of Internet experts and Newbies. Proceedings of WWW9. 2000. Available at http://www9.org/w9cdrom/81/81.html.

[**NSR99**] Navarro-Prieto, R., Scaife, M., & Rogers, Y. Cognitive Strategies in Web Searching. *Proceedings of the 5th Conference on Human Factors & the Web*, 1999. Available at http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto/index.html.

[**MR01**] J. Muramatu and W. Pratt. Transparent queries: Investigating Users' Mental Models of Search Engines. Proceedings of SIGIR 2001.

[**CDT99**] Choo, C. W., Detlor, B., and Turnbull, D. . Information Seeking on the Web - An integrated model of browsing and searching. *Proceedings of the Annual Meeting of the American Society for Information Science (ASIS)*, 1999. Available at http://choo.fis.utoronto.ca/fis/respub/aisis99/. [**K98**] J. Kleinberg. Authoritative sources in a

hyperlinked environment. In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[**BP98**] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th International World Wide Web Conference*, 1998.

[**TG97**] Tauscher, L., and Greenberg, S. How people revisit web pages: Empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies, 47*, 1997.

[**CJ96**] Cockburn, A., & Jones, S. Which way now? Analysing and easing inadequacies in WWW navigation. *International Journal of Human-Computer Studies, 45*, 1996

[**CP95**] Catledge, L. D., & Pitkow, J. E. Characterizing Browsing Strategies in the World-Wide Web. *Proceedings of WWW3*, 1995.

[**HCG01**] D.Hawking, N. Craswell, and K. Griffiths. Which search engine is best at finding online services? WWW10 poster. Available at http://pigfish.vic.cmis.csiro.au/~nickc//pubs/www10actualposter.pdf

[**CHG01**] N. Craswell, D. Hawking and K. Griffiths. Which Search engine is best at finding airline site home pages?. CSIRO Mathematical and Information Sciences TR01/45, 2001. Available at http://pigfish.vic.cmis.csiro.au/~nickc//pubs/TR01-45.pdf

[**CHR01**] Nick Craswell, David Hawking and Stephen Robertson. Effective site finding using link anchor information. SIGIR'01. Available at http://pigfish.vic.cmis.csiro.au/~nickc//pubs/sigir01.pdf