Mining Query Logs: Turning Search Usage Data into Knowledge

Mining Query Logs: Turning Search Usage Data into Knowledge

Fabrizio Silvestri

ISTI — CNR
via G. Moruzzi, 1
56124 Pisa
Italy
fabrizio.silvestri@isti.cnr.it

now
the essence of knowledge
Boston – Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is F. Silvestri, Mining Query Logs: Turning Search Usage Data into Knowledge, Foundation and Trends® in Information Retrieval, vol 4, nos 1–2, pp 1–174, 2010

ISBN: 978-1-60198-282-7

Foundations and Trends® in Information Retrieval

Volume 4 Issues 1–2, 2010

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2010, Volume 4, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in
Information Retrieval
Vol. 4, Nos. 1–2 (2010) 1–174
© 2010 F. Silvestri
DOI: 10.1561/1500000013

Mining Query Logs: Turning Search Usage Data into Knowledge

Fabrizio Silvestri

ISTI — CNR, via G. Moruzzi, 1, 56124 Pisa, Italy,
fabrizio.silvestri@isti.cnr.it

Abstract

Web search engines have stored in their logs information about users since they started to operate. This information often serves many purposes. The primary focus of this survey is on introducing to the discipline of query mining by showing its foundations and by analyzing the basic algorithms and techniques that are used to extract useful knowledge from this (potentially) infinite source of information. We show how search applications may benefit from this kind of analysis by analyzing popular applications of query log mining and their influence on user experience. We conclude the paper by, briefly, presenting some of the most challenging current open problems in this field.

Contents

1
Introduction

"History teaches everything, even the future."
— Alphonse de Lamartine, speech at Macon 1847.

Think about it, for a moment: after checking e-mails, and checking your favorite on-line newspaper and comic strip, what is the first thing you do when connected to the web? You probably open a search engine and start looking for some information you might need either for work or for leisure: news about your favorite actor, news about presidential candidates, and so on.

Even though they are quite rooted in our lives, web search engines are quite new on the scene.

Query Log Mining is a branch of the more general *Web Analytics* [110] scientific discipline. Indeed, it can be considered a special type of web usage mining [213]. According to the Web Analytics Association, "*Web Analytics is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage* [11]".

In particular, query log mining is concerned with all those techniques aimed at discovering interesting patterns from query logs of web search engine with the purpose of enhancing either effectiveness or efficiency of an online service provided through the web.

Keeping into account that query log mining is not only concerned with the search service (from which queries usually come from) but also with more general services like, for instance, search-based advertisement, or web marketing in general [105].

1.1 Web Search Engines

Systems that can be considered similar to modern web search engines started to operate around 1994. The now-defunct *World Wide Web Worm* (*WWWW*) [146] created by Oliver McBryan at the University of Colorado, and the *AliWeb* search engine [124] created by Martijn Koster in 1994, are the two most famous examples. Since then many examples of such systems have been around the web: AltaVista, Excite, Lycos, Yahoo!, Google, ASK, MSN (just to name a few). Nowadays, searching is considered one of the most useful application on the web. As reported in 2005 by *Pew Research Center for The People & The Press* [161]:

> *"search engines have become an increasingly important part of the online experience of American internet users. The most recent findings from Pew Internet & American Life tracking surveys and consumer behavior trends from the comScore Media Metrix consumer panel show that about 60 million American adults are using search engines on a typical day"* [188].

Even if this quote dates back to 2005, it is very likely that those survey results are still valid (if not still more positives for search engines). On the other side of the coin, search engines' users are satisfied by their search experience [189].

In a paper overviewing the challenges in modern web search engines' design, Baeza-Yates *et al.* [14] state:

> *The main challenge is hence to design large-scale distributed systems that* ***satisfy the user expectations****, in which* ***queries use resources efficiently****, thereby reducing the cost per query.*

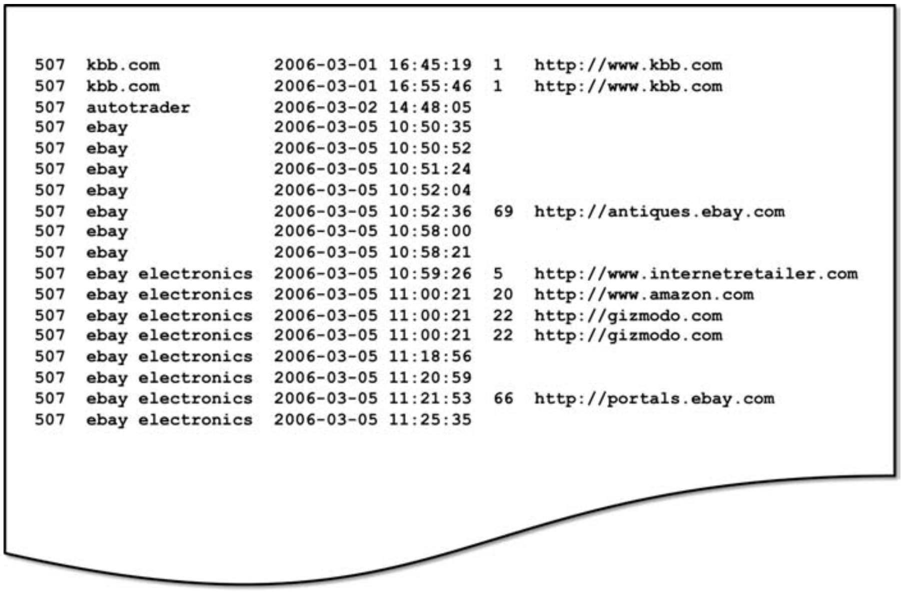
Therefore, the two key performance indicators in this kind of application, in order, are: (i) the quality of returned results (e.g. handle quality diversity and fight spam), and (ii) the speed with which results are returned.

Web search engines are part of a broader class of software systems, namely Information Retrieval (IR) Systems. Basically, IR systems were born in the early 1960s due to two major application needs. Firstly, allowing searching through digital libraries. Secondly, the need for computer users to search through the data they were collecting in their own digital repositories.

Intuitively, an IR system is a piece of software whose main purpose is to return a list of documents in response to a user query. Thus far, this description makes IR systems similar to what a DB system is. Indeed, the most important difference between DB and IR systems is that DB systems return objects that exactly match the user query, whereas IR systems have to cope with natural language that makes it simply impossible for an IR system to return perfect matches. Just to make a very simple example: what does meta refer to? A *meta* character? The *meta* key in computer keyboards? Every single query may mean different things to different users. Even worse, *polysemy* also happens. In Spanish the word *meta* means *goal*.

To this extent, a web search engine is in all respects an IR system [221] only on a *very large scale*. The uncertainty in users' intent is also present in web search engines. Differently from smaller scale IR systems, though, web IR systems can rely on the availability of a huge amount of usage information stored in *query logs*.

One of the most used ways of enhancing the users' search experience, in fact, is the exploitation of the knowledge contained within past queries. A query log, typically, contains information about users, issued queries, clicked results, etc. From this information knowledge can be extracted to improve the quality (both in terms of effectiveness and efficiency) of their system. Figure 1.1 shows a fragment of the AOL query log. The format of this query log represents a record using five features: user id, query, timestamp, rank of the clicked result, host string of the clicked URL.



```
507  kbb.com           2006-03-01 16:45:19  1   http://www.kbb.com
507  kbb.com           2006-03-01 16:55:46  1   http://www.kbb.com
507  autotrader        2006-03-02 14:48:05
507  ebay              2006-03-05 10:50:35
507  ebay              2006-03-05 10:50:52
507  ebay              2006-03-05 10:51:24
507  ebay              2006-03-05 10:52:04
507  ebay              2006-03-05 10:52:36  69  http://antiques.ebay.com
507  ebay              2006-03-05 10:58:00
507  ebay              2006-03-05 10:58:21
507  ebay electronics  2006-03-05 10:59:26  5   http://www.internetretailer.com
507  ebay electronics  2006-03-05 11:00:21  20  http://www.amazon.com
507  ebay electronics  2006-03-05 11:00:21  22  http://gizmodo.com
507  ebay electronics  2006-03-05 11:00:21  22  http://gizmodo.com
507  ebay electronics  2006-03-05 11:18:56
507  ebay electronics  2006-03-05 11:20:59
507  ebay electronics  2006-03-05 11:21:53  66  http://portals.ebay.com
507  ebay electronics  2006-03-05 11:25:35
```

Fig. 1.1 A fragment of the AOL query log [160].

How query logs interact with search engines has been studied in many papers. For a general overview, [12, 20] are good starting point references.

In this paper, we review some of the most recent techniques dealing with query logs and how they can be used to enhance web search engine operations. We are going to summarize the basic results concerning query logs: analyses, techniques used to extract knowledge, most remarkable results, most useful applications, and open issues and possibilities that remain to be studied.

The purpose is, thus, to present ideas and results in the most comprehensive way. We review fundamental, and state-of-the-art techniques. In each section, even if not directly specified, we review and analyze the algorithms used, not only their results. This paper is intended for an audience of people with basic knowledge of computer science. We also expect readers to have a basic knowledge of Information Retrieval. Everything not at a basic level is analyzed and detailed.

Before going on, it is important to make clear that all the analyses and results reported were not reproduced by the author. We only report results as stated in the papers referenced. In some cases we slightly adapted them to make concepts clearer.

1.2 Sketching the Architecture of a Web Search Engine

A search engine is one of the most complicated pieces of software a company may develop. Consisting of tens of interdependent modules, it represents one of the toughest challenge in today's computer engineering world.

Many papers and books sketch the architecture of web search engines. For example Barroso *et al.* [33] present the architecture of Google as it was in 2003. Other search engines are believed to have similar architectures. When a user enters a query, the user's browser builds a URL (for example http://www.google.com/search?q=foundations+trends+IR). The browser, then, looks up on a DNS directory for mapping the URL main site address (i.e., www.google.com) into a particular IP address corresponding to a particular data-center hosting a replica of the entire search system. The mapping strategy is done accordingly to different objectives such as: availability, geographical proximity, load and capacity. The browser, then, sends an HTTP request to the selected data-center, and thereafter, the query processing is entirely local to that center. After the query is answered by the local data-center, the result is returned in the form of an HTML page, to the originating client.

Figure 1.2 shows they way the main modules of a web search engine are connected.

Web search engines get their data from different sources: the web (primarily), Image and video repositories (e.g. Flickr, or YouTube), etc. In particular, in the case of web content, a crawler scours through hypertext pages searching for new documents, and detecting stale, or updated content. Crawlers store the data into a repository of content (also known as web document *cache*), and structure (the graph representing how web pages are interconnected). The latter being used, mainly, as a feature for computing static document rank scores (e.g. PageRank [157], or HITS [122]). In modern web retrieval systems, crawlers continuously run and download pages from the web updating

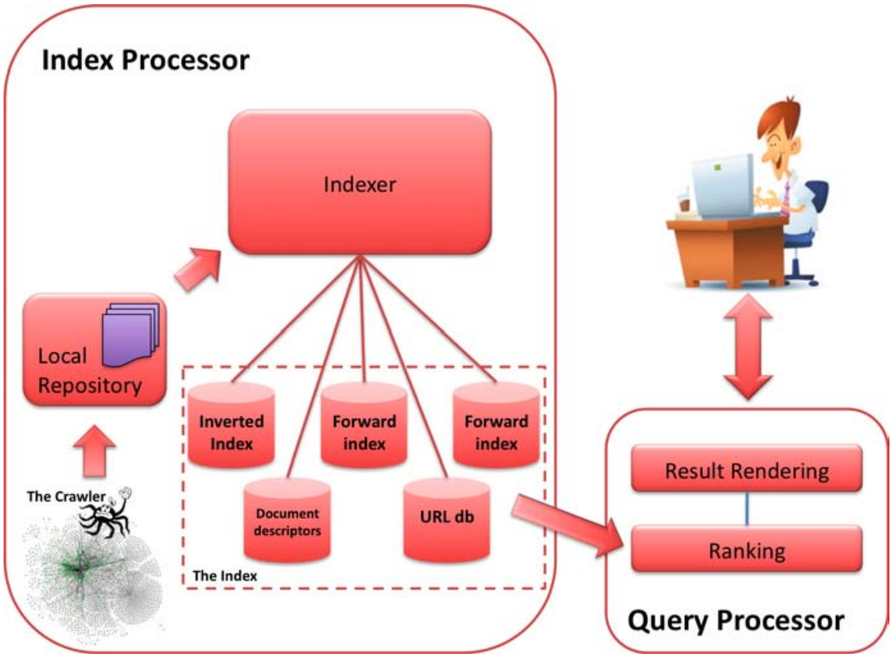


Fig. 1.2 The typical structure of a web search engine. Note that throughout the text IR core, and query server will be used interchangeably.

incrementally the content of the document cache. For more information on crawling, interested readers can refer to Castillo's Ph.D. thesis on web Crawling [57].

The textual (i.e., hypertextual) content is *indexed* to allow fast retrieval operations (i.e., *query requests*). The index (built by the *Indexer*) usually comprises of several different archives storing different facets of the index. The format of each archive is designed for enabling a fast retrieval of information needed to resolve queries. The format of the index is the subject of Section 5 where we review some of the most used techniques for optimizing index allocation policies.

Usually in real systems the design is tailored to favor aggregate request throughput not peak server response time [33].

In real-world search engines, the index is distributed among a set of *query servers* coordinated by a *broker*. The broker, accepts a query from the user and distributes it to the set of query servers. The index servers

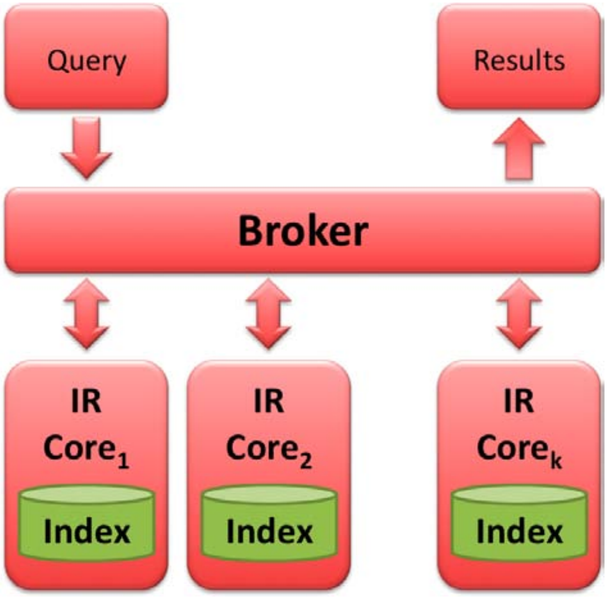


Fig. 1.3 The typical structure of a distributed web search engine.

retrieve relevant documents, compute scores, rank results and return them back to the broker which renders the result page and sends it to the user. Figure 1.3 shows the interactions taking place among query servers and the broker.

The broker is usually the place where queries are grabbed and stored in the query logs. A module dedicated to analyze past queries is also usually available within the architecture components.

1.2.1 The Index

An Inverted File index on a collection of web pages consists of several interlinked components. The principal ones are the *lexicon*, i.e., the list of all the *index terms* appearing in the collection, and the corresponding set of *inverted lists*, where each list is associated with a distinct term of the lexicon. Each inverted list contains, in turn, a set of *postings*. Each posting collects information about the *occurrences* of the corresponding term in the collection's documents. For the sake of simplicity, in the following discussion we consider that each posting only includes the identifier of the document (DocID) where the term appears, even if postings actually store other information used for document ranking purposes (e.g. in the implementation [203] each posting also includes the positions and the frequency of the term within the document, and context information like the appearance of the term within specific html tags).

Several sequential algorithms have been proposed in the past, which try to balance the use of memory hierarchy in order to deal with the large amount of input/output data involved in query processing. The inverted file index [221] is the data structure typically adopted for indexing the web. This occurs for three reasons. First, an inverted file index allows the efficient resolution of queries on huge collections of web data [246]. In fact, it works very well for common web queries, where the conjunction of a few terms is to be searched for. Second, an inverted file index can be easily compressed to reduce the space occupancy in order to better exploit the memory hierarchy [203]. Third, an inverted file can be easily built using a sort-based algorithm in time complexity that is the same order of a sorting algorithm [246].

Query answering using inverted file is a very straightforward task. We illustrate the basic AND operation and refer to other papers for a thorough analysis of the remaining operations. Given a query as a conjunction of two terms ($t_1 \wedge t_2$), the query resolution proceeds by firstly looking up t_1 and t_2 in the lexicon to retrieve the corresponding inverted lists l_1 and l_2. The result set is then built by intersecting the two lists, thus, returning those documents having the two terms in common. During the intersection step a scoring function is also computed to evaluate the likeliness of a document to be relevant for the query. The top r results are then selected (in typical web search engines r is usually set to 10 results) and successively returned to the users who originated the query. Query processing can be done in two different ways: *Document-At-A-Time* (DAAT), when document lists for terms are scanned contemporary, as opposed to the *Term-At-A-Time* (TAAT) strategy, where each term is considered separately [219].

Another important feature of inverted file indexes is that they can be easily partitioned. Let us consider a typical distributed web search engine: the index can be distributed across the different nodes

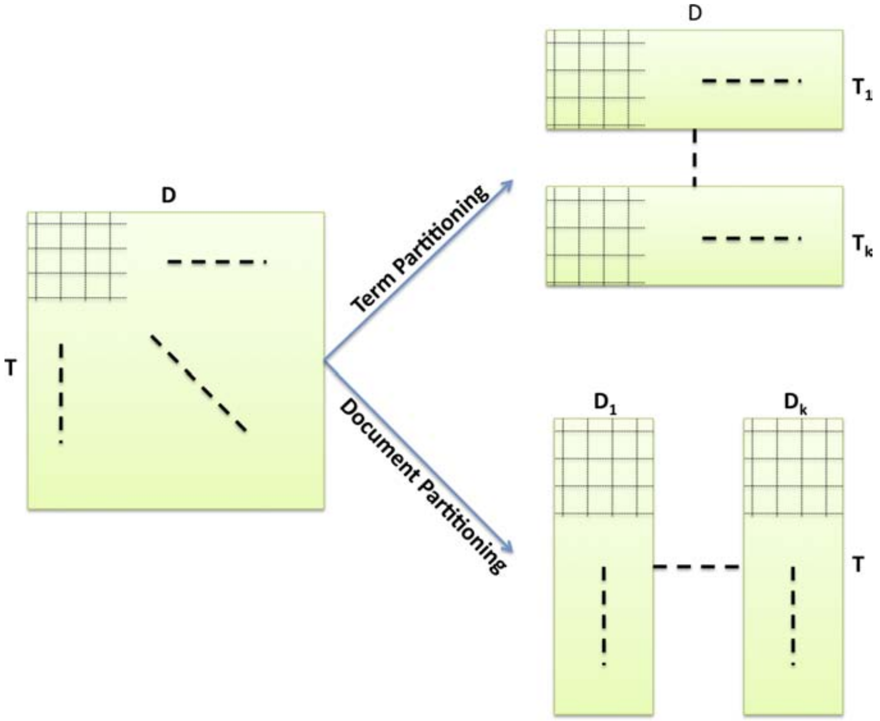


Fig. 1.4 The two different ways of partitioning an inverted index. Rows of the whole $T \times D$ matrix are the lexicon entries, columns represent the posting lists.

of the underlying architecture in order to enhance the overall system's throughput (i.e., the number of queries answered per each second). For this purpose, two different partitioning strategies can be devised.

The first approach requires to horizontally partition the whole inverted index with respect to the lexicon, so that each index server stores the inverted lists associated with only a subset of the index terms. This method is also known as *term partitioning* or *global inverted files*. The other approach, known as *document partitioning* or *local inverted files*, requires that each index server becomes responsible for a disjoint subset of the whole document collection (vertical partitioning of the inverted index). Figure 1.5 graphically depicts such partitioning schemes.

The construction of a document-partitioned inverted index is a two-staged process. In the first stage each index partition is built locally

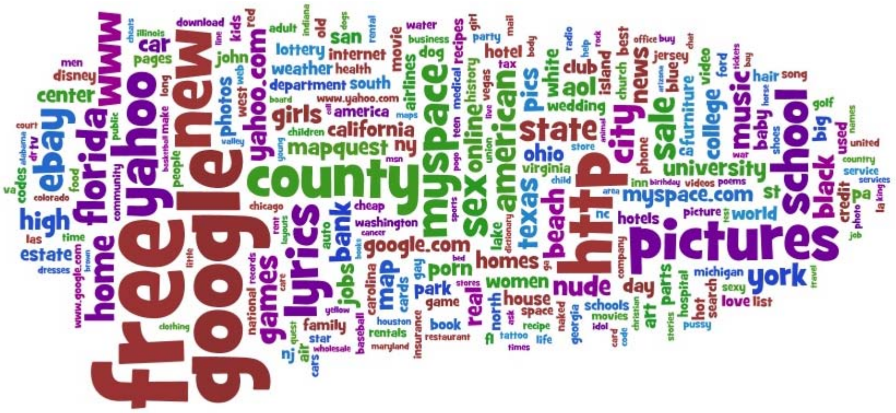


Fig. 1.5 A cloud of the 250 most frequent queried terms in the AOL query log [160]. Picture has been generated using http://www.wordle.net.

and independently from a partition of the whole collection. The second phase collects global statistics computed over the whole inverted index. One of the most valuable advantages of document partitioning is the possibility of easily performing updates. In fact, new documents may simply be inserted into a new partition to independently index separately from the others [169].

Since the advent of web search engines, a large number of papers have been published describing different architectures for search engines, and search engine components [10, 25, 47, 33, 96, 97, 147, 150, 153, 204]. Many other papers [13, 14, 100, 101] enumerate the major challenges search engine developers must address in order to improve their ability to help users in finding information they need. Interested readers shall find in the above referenced papers many interesting insights. Needless to say, you shall not find any particular details, in this survey, about the *real* structure of a search engine. Usually, this kind of information is highly confidential and it is very unlikely that search companies will ever disclose them.

1.3 Fun Facts about Queries

Due to their “commercial importance”, finding query logs has always been a difficult task. The very first publicly available query log dates back to 1997. Doug Cutting, representing Excite, a major search service to that date, made available for research a set of user queries as submitted to Excite. Since then, the other query logs made publicly available were the AltaVista log, the TodoBR query log, and the AOL log.

AOL eventually fired employees involved in the public release of their log. This confirms, even more strongly, the particular level of privacy characterizing such data. Obviously, this may sound worse than it is. Search Engine companies are still releasing their data, only that they adopt more conservative policies and release data under research licenses preventing broad distribution.

Figure 1.5 shows a cloud of the 250 most frequent queried terms in the AOL query log.

Queries posed by users are somewhat entertaining. To have an idea of what every day users search through search engines, consider these queries that were actually extracted from the (in)famous AOL Query Log.[1]

In today's hectic world, people often get very stressed. Stress produces distraction and user #427326 probably was a little more stressed than the average. At 2006-04-21 21:16:51, in fact, he was looking for the following sentence "where is my computer". Well, probably is closer than what you were suspecting. Actually, searching for this sentence on popular search engines result in around 200,000 results. Gosh! Many stressed people out there![2]

Again, people gets stressed easily today. I dare you to guess what was user #582088 looking for by entering the following keywords "can you hear me out there i can hear you i got you i can hear you over i really feel strange i wanna wish for something new this is the scariest thing ive ever done in my life who do we think we are angels and airwaves im gonna count down till 10 52 i can". Hint: try by yourself and enter the above sentence. What is the result? In your opinion, what was user doing while typing the query?

Search engines publish some of the most interesting submitted queries. *Interestingness*, here, is a relative concept. Depending on the

[1] http://www.techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/.

[2] Indeed, many results are on people asking where is "My Computer" icon on their desktop.

search engine company, interesting may mean different things. At Google, for instance, Zeitgeist[3] is a

> "*cumulative snapshot of interesting queries people are asking over time, within country domains, and some on Google.com that perhaps reveal a bit of the human condition.*"

Zeitgeist does not reveal the most searched queries, but only those having had a "sudden", and "unexpected" raise in popularity. For instance, late in 2007 Italian Zeitgeist ranked "*federico calzolari*"[4] as the most "inflated" query. Many (mainly Italian) newspapers, and blogs started to ask who is the person referred to in the query. The name was that of a Ph.D. student in Pisa that periodically queried Google for his name. This resulted in an unexpected raise in popularity for the query term thus ending up in the Zeitgeist. Many people, mainly journalists, started to discuss whether or not Federico Calzolari has hacked the Google ranking algorithm.

It is important to point out that the discussion above seems to imply that one could guess the intent of the users by looking at query session. This is far from being true. As it is shown later on, the identification of users' tasks is a very challenging activity. The main goal of this paragraph is to make readers aware of: (i) the variety of information in query logs, and (ii) the detail that, in principle, can be obtained about a single user.

An interesting recent paper dealing in a scientific way with discovering information about search engine index content by carefully probing it using queries out of a query log is Bar-Yossef and Gurevich [28].

1.4 Privacy Issues in Query Log Mining

The most recent scandal concerning privacy and query logs happened in 2006 at AOL. AOL compiled a statistical sampling of more than 20 million queries entered by more than 650,000 of their customers, and then made this DB available to the public for research purposes.

[3] http://www.google.com/press/zeitgeist.html.

[4] http://googleitalia.blogspot.com/2007/12/zeitgeist- di-novembre.html (in Italian).

While user names were replaced by numbers, these numbers provide a thread by which queries by a given user could be identified so that if, for example, a user entered some piece of information which permits their identity to be discerned, all the other queries they made during the sampling period could be identified as theirs. AOL received so much criticism for releasing this data that it eventually fired two employees. The real problem was that they released ALL off the data to EVERYONE. A Non-Disclosure-Agreement form for researchers to sign, would have saved a lot of pain to AOL people that were fired after the mishap.

Many commercial search engines overcome to this problem by simply not publishing their logs. Is this approach good? Yes for some reasons, no for others. Roughly speaking, it is good that people (in general) cannot access query log data. As already said above they might be used to infer users' preferences, tastes, and other personal information that might be used against their will. On the other hand, as pointed also out by Judit Bar-Ilan in [27]

> *"[...] interesting results can be obtained from query logs without jeopardizing the privacy of the users."*

While Bar-Ilan showed that it is possible to sanitize a query log in order to prevent private information to be disclosed, Jones *et al.* [117] showed that even heavily scrubbed query logs, still containing session information, have significant privacy risks.

This paper does not deal with this (extremely important) issue, but we would not have been comfortable without making the reader aware of this issues. More important, we think this would clarify why many studies reported here are made on (sometimes) old and outdated logs, or logs privately held by companies not sharing them.

The interested reader shall find an introduction and some thoughts about privacy and log publishing in recently published papers [1, 126, 164, 230]. Recently, Cooper published a very detailed survey on query log privacy-enhancing techniques [64], readers interested in this topic shall find a very thorough analysis of the most recent techniques dealing with privacy preserving analysis of query logs.

Recently ASK[5] has given the possibility to users to explicitly deny the storing of their usage data. On the other hand, Google, Yahoo, and Microsoft, continuously ask users for the permission to store their preferences, behaviors, and data in general. What is the most correct behavior? It depends on search engines' policies, thus we do not enter into details on how these are managed.

The remainder of this work presents the most recent results and advances that have used query logs as (the main) source of information. It is worth mentioning here that not always the experiments presented might be reproduced. This is something that in science should be avoided [87]. Unfortunately, as already said above, the main source of knowledge (the query logs) are mainly kept by search engine companies that for many reasons (not last, privacy issues) are very reluctant of give them away, even to scientists. Therefore, many times in this article, the experimental evaluation is based on results obtained by others and presented in the literature. We apologize in advance to both authors of the mentioned papers, and to readers.

Before entering into the details of our survey, it is important to remark that query log mining is a very hot topic nowadays. The material covered by this survey is to be considered as a valid starting point for those interested in knowing something more on the topic. Proceedings of the major conference series (e.g. SIGIR, WWW, SIGMOD, VLDB, SIGKDD, CIKM, etc. Just to name a few) and top journals (e.g. ACM TOIS, ACM TWEB, ACM TKDD, ACM TOIT, Information Processing & Management, JASIST, IEEE TKDE, etc.) are the best source for the state-of-the-art works on this field. Furthermore, we use the same notation used by the authors of the surveyed papers. This, in our opinion, makes each (sub)section of the survey more independent and leave to the reader the possibility of selecting the techniques he is interested on.

That said, let the journey into the marvelous world of queries begin . . .

[5] http://www.ask.com.

References

[1] E. Adar, "User 4xxxxx9: Anonymizing query logs," in *Query Log Analysis: Social And Technological Challenges. A Workshop at the 16th International World Wide Web Conference (WWW 2007)*, (E. Amitay, C. G. Murray, and J. Teevan, eds.), May 2007.

[2] E. Adar, D. S. Weld, B. N. Bershad, and S. S. Gribble, "Why we search: Visualizing and predicting user behavior," in *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pp. 161–170, New York, NY, USA: ACM, 2007.

[3] A. Agarwal and S. Chakrabarti, "Learning random walks to rank nodes in graphs," in *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, pp. 9–16, New York, NY, USA: ACM, 2007.

[4] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 19–26, New York, NY, USA: ACM, 2006.

[5] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning user interaction models for predicting web search result preferences," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–10, New York, NY, USA: ACM, 2006.

[6] E. Agichtein and Z. Zheng, "Identifying "best bet" web search results by mining past user behavior," in *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 902–908, New York, NY, USA: ACM, 2006.

[7] R. Agrawal, T. Imielinski, and A. N. Swami, “Mining association rules between sets of items in large databases,” in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26–28, 1993*, (P. Buneman and S. Jajodia, eds.), pp. 207–216, ACM Press, 1993.

[8] F. Ahmad and G. Kondrak, “Learning a spelling error model from search query logs,” in *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 955–962, Vancouver, Canada: Association for Computational Linguistic, October 2005.

[9] C. Anderson, *The Long Tail.* Random House Business, 2006.

[10] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, “Searching the web,” *ACM Transactions on Internet Technology*, vol. 1, no. 1, pp. 2–43, 2001.

[11] V. Authors, “About web analytics association,” Retrieved on August 2009. http://www.webanalyticsassociation.org/aboutus/.

[12] R. Baeza-Yates, *Web Mining: Applications and Techniques.* ch. Query Usage Mining in Search Engines, pp. 307–321, Idea Group, 2004.

[13] R. Baeza-Yates, “Algorithmic challenges in web search engines,” in *Proceedings of the 7th Latin American Symposium on Theoretical Informatics (LATIN’06)*, pp. 1–7, Valdivia, Chile, 2006.

[14] R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, and F. Silvestri, “Challenges in distributed information retrieval,” in *International Conference on Data Engineering (ICDE)*, Istanbul, Turkey: IEEE CS Press, April 2007.

[15] R. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri, “The impact of caching on search engines,” in *SIGIR ’07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 183–190, New York, NY, USA: ACM, 2007.

[16] R. Baeza-Yates, A. Gionis, F. P. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri, “Design trade-offs for search engine caching,” *ACM Transactions on the Web*, vol. 2, no. 4, pp. 1–28, 2008.

[17] R. Baeza-Yates, C. Hurtado, and M. Mendoza, *Query Recommendation Using Query Logs in Search Engines.* pp. 588–596. Vol. 3268/2004 *of Lecture Notes in Computer Science*, Berlin, Heidelberg: Springer, November 2004.

[18] R. Baeza-Yates, C. Hurtado, and M. Mendoza, “Ranking boosting based in query clustering,” in *Proceedings of 2004 Atlantic Web Intelligence Conference*, Cancun, Mexico, 2004.

[19] R. Baeza-Yates and A. Tiberi, “Extracting semantic relations from query logs,” in *KDD ’07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 76–85, New York, NY, USA: ACM, 2007.

[20] R. A. Baeza-Yates, “Applications of web query mining,” in *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21–23, 2005, Proceedings*, (D. E. Losada and J. M. Fernández-Luna, eds.), pp. 7–22, Springer, 2005.

[21] R. A. Baeza-Yates, "Graphs from search engine queries," in *SOFSEM 2007: Theory and Practice of Computer Science, 33rd Conference on Current Trends in Theory and Practice of Computer Science, Harrachov, Czech Republic, January 20–26, 2007, Proceedings*, (J. van Leeuwen, G. F. Italiano, W. van der Hoek, C. Meinel, H. Sack, and F. Plasil, eds.), pp. 1–8, Springer, 2007.

[22] R. A. Baeza-Yates, C. A. Hurtado, and M. Mendoza, "Improving search engines by query clustering," *JASIST*, vol. 58, no. 12, pp. 1793–1804, 2007.

[23] R. A. Baeza-Yates, C. A. Hurtado, M. Mendoza, and G. Dupret, "Modeling user search behavior," in *Third Latin American Web Congress (LA-Web 2005), 1 October - 2 November 2005, Buenos Aires, Argentina*, pp. 242–251, IEEE Computer Society, 2005.

[24] R. A. Baeza-Yates, F. Junqueira, V. Plachouras, and H. F. Witschel, "Admission policies for caches of search engine results," in *SPIRE*, pp. 74–85, 2007.

[25] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[26] R. A. Baeza-Yates and F. Saint-Jean, "A three level search engine index based in query log distribution," in *SPIRE*, pp. 56–65, 2003.

[27] J. Bar-Ilan, "Access to query logs — an academic researcher's point of view," in *Query Log Analysis: Social And Technological Challenges. A Workshop at the 16th International World Wide Web Conference (WWW 2007)*, (E. Amitay, C. G. Murray, and J. Teevan, eds.), May 2007.

[28] Z. Bar-Yossef and M. Gurevich, "Mining search engine query logs via suggestion sampling," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 54–65, 2008.

[29] R. Baraglia, F. Cacheda, V. Carneiro, F. Diego, V. Formoso, R. Perego, and F. Silvestri, "Search shortcuts: A new approach to the recommendation of queries," in *RecSys '09: Proceedings of the 2009 ACM Conference on Recommender Systems*, New York, NY, USA: ACM, 2009.

[30] R. Baraglia, F. Cacheda, V. Carneiro, V. Formoso, R. Perego, and F. Silvestri, "Search shortcuts: Driving users towards their goals," in *WWW '09: Proceedings of the 18th International Conference on World Wide Web*, pp. 1073–1074, New York, NY, USA: ACM, 2009.

[31] R. Baraglia, F. Cacheda, V. Carneiro, V. Formoso, R. Perego, and F. Silvestri, "Search shortcuts using click-through data," in *WSCD '09: Proceedings of the 2009 Workshop on Web Search Click Data*, pp. 48–55, New York, NY, USA: ACM, 2009.

[32] R. Baraglia and F. Silvestri, "Dynamic personalization of web sites without user intervention," *Communications of the ACM*, vol. 50, no. 2, pp. 63–67, 2007.

[33] L. A. Barroso, J. Dean, and U. Hölzle, "Web search for a planet: The google cluster architecture," *IEEE Micro*, vol. 23, no. 2, pp. 22–28, 2003.

[34] S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, "Temporal analysis of a very large topically categorized web query log," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 2, pp. 166–178, 2007.

[35] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, "Hourly analysis of a very large topically categorized web query log," in *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 321–328, New York, NY, USA: ACM, 2004.

[36] S. M. Beitzel, E. C. Jensen, O. Frieder, D. D. Lewis, A. Chowdhury, and A. Kolcz, "Improving automatic query classification via semi-supervised learning," in *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pp. 42–49, Washington, DC, USA: IEEE Computer Society, 2005.

[37] S. M. Beitzel, E. C. Jensen, D. D. Lewis, A. Chowdhury, and O. Frieder, "Automatic classification of web queries using very large unlabeled query logs," *ACM Transactions on Information Systems*, vol. 25, no. 2, p. 9, 2007.

[38] L. A. Belady, "A study of replacement algorithms for a virtual storage computer," *IBM Systems Journal*, vol. 5, no. 2, pp. 78–101, 1966.

[39] R. E. Bellman, *Dynamic Programming.* Princeton, NJ: Princeton University Press, 1957.

[40] "Beowulf Project at CESDIS," http://www.beowulf.org.

[41] M. Bilenko and R. W. White, "Mining the search trails of surfing crowds: Identifying relevant websites from user activity," in *WWW '08: Proceeding of the 17th International Conference on World Wide Web*, pp. 51–60, New York, NY, USA: ACM, 2008.

[42] B. Billerbeck, F. Scholer, H. E. Williams, and J. Zobel, "Query expansion using associated queries," in *Proceedings of the twelfth international conference on information and knowledge management*, pp. 2–9, ACM Press, 2003.

[43] P. Boldi and S. Vigna, "The webgraph framework i: Compression techniques," in *WWW '04: Proceedings of the 13th International Conference on World Wide Web*, pp. 595–602, New York, NY, USA: ACM Press, 2004.

[44] J. Boyan, D. Freitag, and T. Joachims, "A machine learning architecture for optimizing web search engines," in *Proceedings of the AAAI Workshop on Internet-Based Information Systems*, 1996.

[45] O. Boydell and B. Smyth, "Capturing community search expertise for personalized web search using snippet-indexes," in *CIKM '06: Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 277–286, New York, NY, USA: ACM, 2006.

[46] J. S. Breese, D. Heckerman, and C. M. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *UAI*, pp. 43–52, 1998.

[47] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *WWW7: Proceedings of the Seventh International Conference on World Wide Web 7*, pp. 107–117, Amsterdam, The Netherlands: Elsevier Science Publishers B. V., 1998.

[48] A. Z. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.

[49] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang, "Robust classification of rare queries using web knowledge," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pp. 231–238, New York, NY, USA: ACM, 2007.

[50] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic Clustering of the Web," in *Selected Papers from the Sixth International Conference on World Wide Web*, pp. 1157–1166, Essex, UK: Elsevier Science Publishers Ltd., 1997.

[51] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, New York, NY, USA: ACM, 2005.

[52] C. J. C. Burges, R. Ragno, and Q. V. Le, "Learning to rank with nonsmooth cost functions.," in *NIPS*, (B. Schölkopf, J. Platt, and T. Hoffman, eds.), pp. 193–200, MIT Press, 2006.

[53] R. Buyya, ed., *High Performance Cluster Computing*. Prentice Hall PTR, 1999.

[54] H. C. by Thomas, E. L. Charles, L. R. Ronald, and S. Clifford, *Introduction to Algorithms*. The MIT Press, 2001.

[55] J. Callan and M. Connell, "Query-based sampling of text databases," *ACM Transactions on Information Systems*, vol. 19, no. 2, pp. 97–130, 2001.

[56] J. P. Callan, Z. Lu, and W. B. Croft, "Searching distributed collections with inference networks," in *SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21–28, New York, NY, USA: ACM, 1995.

[57] C. Castillo, "Effective web crawling," PhD thesis, Department of Computer Science — University of Chile, Santiago, Chile, November 2004.

[58] J. Caverlee, L. Liu, and J. Bae, "Distributed query sampling: A quality-conscious approach," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 340–347, New York, NY, USA: ACM, 2006.

[59] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary clustering," in *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 554–560, New York, NY, USA: ACM, 2006.

[60] Q. Chen, M. Li, and M. Zhou, "Improving query spelling correction using web search results," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 181–189, Prague, Czech Republic: Association for Computational Linguistic, June 2007.

[61] F. Chierichetti, A. Panconesi, P. Raghavan, M. Sozio, A. Tiberi, and E. Upfal, "Finding near neighbors through cluster pruning," in *Proceedings of ACM SIGMOD/PODS 2007 Conference*, 2007.

[62] P. A. Chirita, C. S. Firan, and W. Nejdl, "Personalized query expansion for the web," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 7–14, New York, NY, USA: ACM, 2007.

[63] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe, "Collection statistics for fast duplicate document detection," *ACM Transactions on Information Systems*, vol. 20, no. 2, pp. 171–191, 2002.

[64] A. Cooper, "A survey of query log privacy-enhancing techniques from a policy perspective," *ACM Transactions on the Web*, vol. 2, no. 4, pp. 1–27, 2008.

[65] N. Craswell, P. Bailey, and D. Hawking, "Server selection on the world wide web," in *DL '00: Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 37–46, New York, NY, USA: ACM, 2000.

[66] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *WSDM '08: Proceedings of the International Conference on Web Search and Web Data Mining*, pp. 87–94, New York, NY, USA: ACM, 2008.

[67] S. Cucerzan and E. Brill, "Spelling correction as an iterative process that exploits the collective knowledge of web users," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pp. 293–300, July 2004.

[68] S. Cucerzan and R. W. White, "Query suggestion based on user landing pages," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 875–876, New York, NY, USA: ACM Press, 2007.

[69] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Probabilistic query expansion using query logs," in *WWW '02: Proceedings of the 11th International Conference on World Wide Web*, pp. 325–332, New York, NY, USA: ACM, 2002.

[70] E. Cutrell and Z. Guan, "What are you looking for? An eye-tracking study of information usage in web search," in *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 407–416, New York, NY, USA: ACM, 2007.

[71] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[72] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pp. 89–98, 2003.

[73] Z. Dou, R. Song, and J. Wen, "A large-scale evaluation and analysis of personalized search strategies," in *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, pp. 572–581, May 2007.

[74] T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data," *ACM Transactions on Information Systems*, vol. 24, no. 1, pp. 51–78, 2006.

[75] C. H. Fenichel, "Online searching: Measures that discriminate among users with different types of experience," *JASIS*, vol. 32, no. 1, pp. 23–32, 1981.

[76] P. Ferragina and A. Gulli, "A personalized search engine based on web-snippet hierarchical clustering," in *WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, pp. 801–810, New York, NY, USA: ACM, 2005.

[77] L. Fitzpatrick and M. Dent, "Automatic feedback using past queries: social searching?," in *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 306–313, New York, NY, USA: ACM, 1997.

[78] B. M. Fonseca, P. B. Golgher, E. S. de Moura, and N. Ziviani, "Using association rules to discover search engines related queries," in *LA-WEB '03: Proceedings of the First Conference on Latin American Web Congress*, p. 66, Washington, DC, USA: IEEE Computer Society, 2003.

[79] I. Foster and C. Kesselman, eds., *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan-Kaufmann, 1999.

[80] S. T. I. Foster and C. Kesselman, "The anatomy of the grid: Enabling scalable virtual organization," *Int'l Journal on Supercomputer Application*, vol. 3, no. 15.

[81] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, pp. 933–969, 2003.

[82] N. Fuhr, "Optimal polynomial retrieval functions based on the probability ranking principle," *ACM Transactions on Information Systems*, vol. 7, no. 3, pp. 183–204, 1989.

[83] N. Fuhr, "A decision-theoretic approach to database selection in networked ir," *ACM Transactions on Information Systems*, vol. 17, no. 3, pp. 229–249, 1999.

[84] N. Fuhr, S. Hartmann, G. Knorz, G. Lustig, M. Schwantner, and K. Tzeras, "AIR/X — a rule-based multistage indexing system for large subject fields," in *Proceedings of the RIAO'91, Barcelona, Spain, April 2–5, 1991*, pp. 606–623, 1991.

[85] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," in *VLDB '05: Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 181–192, VLDB Endowment, 2005.

[86] G. W. Furnas, S. C. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure," in *SIGIR*, pp. 465–480, 1988.

[87] G. Galilei, "Discorsi e dimostrazioni matematiche intorno a due nuove scienze," Leida : Appresso gli Elsevirii, 1638.

[88] L. A. Granka, T. Joachims, and G. Gay, "Eye-tracking analysis of user behavior in www search," in *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, pp. 478–479, New York, NY, USA: ACM, 2004.

[89] L. Gravano, H. Garcia-Molina, and A. Tomasic, "The efficacy of gloss for the text database discovery problem," Technical Report, Stanford University, Stanford, CA, USA, 1993.

[90] L. Gravano, H. García-Molina, and A. Tomasic, "The effectiveness of gioss for the text database discovery problem," in *SIGMOD '94: Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*, pp. 126–137, New York, NY, USA: ACM, 1994.

[91] L. Gravano, H. García-Molina, and A. Tomasic, "Gloss: text-source discovery over the internet," *ACM Transactions on Database Systems*, vol. 24, no. 2, pp. 229–264, 1999.

[92] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein, "Categorizing web queries according to geographical locality," in *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pp. 325–333, New York, NY, USA: ACM, 2003.

[93] Z. Guan and E. Cutrell, "An eye tracking study of the effect of target rank on web search," in *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 417–420, New York, NY, USA: ACM, 2007.

[94] T. H. Haveliwala, "Topic-sensitive pagerank," in *WWW '02: Proceedings of the 11th International Conference on World Wide Web*, pp. 517–526, New York, NY, USA: ACM, 2002.

[95] D. Hawking, "Overview of the trec-9 web track," in *TREC*, 2000.

[96] D. Hawking, "Web search engines: Part 1," *Computer*, vol. 39, no. 6, pp. 86–88, 2006.

[97] D. Hawking, "Web search engines: Part 2," *Computer*, vol. 39, no. 8, pp. 88–90, 2006.

[98] D. Hawking and P. Thistlewaite, "Methods for information server selection," *ACM Transactions on Information Systems*, vol. 17, no. 1, pp. 40–76, 1999.

[99] J. Hennessy and D. Patterson, *Computer Architecture — A Quantitative Approach*. Morgan Kaufmann, 2003.

[100] M. R. Henzinger, "Algorithmic challenges in web search engines," *Internet Mathematics*, vol. 1, no. 1, 2003.

[101] M. R. Henzinger, R. Motwani, and C. Silverstein, "Challenges in web search engines," *SIGIR Forum*, vol. 36, no. 2, pp. 11–22, 2002.

[102] T. C. Hoad and J. Zobel, "Methods for identifying versioned and plagiarized documents," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 3, pp. 203–215, 2003.

[103] I. Hsieh-Yee, "Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers," *JASIS*, vol. 44, no. 3, pp. 161–174, 1993.

[104] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[105] B. J. Jansen and M. Resnick, "An examination of searcher's perceptions of nonsponsored and sponsored links during ecommerce web searching," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 14, pp. 1949–1961, 2006.

[106] B. J. Jansen and A. Spink, "An analysis of web searching by european alltheweb.com users," *Information Processing and Management*, vol. 41, no. 2, pp. 361–381, 2005.

[107] B. J. Jansen and A. Spink, "How are we searching the world wide web? A comparison of nine search engine transaction logs," *Information Processing and Management*, vol. 42, no. 1, pp. 248–263, 2006.

[108] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real life information retrieval: A study of user queries on the web," *SIGIR Forum*, vol. 32, no. 1, pp. 5–17, 1998.

[109] B. J. Jansen, A. Spink, and S. Koshman, "Web searcher interaction with the dogpile.com metasearch engine," *JASIST*, vol. 58, no. 5, pp. 744–755, 2007.

[110] B. J. J. Jansen, "Understanding user-web interactions via web analytics," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 1, no. 1, pp. 1–102, 2009.

[111] T. Joachims, "Optimizing search engines using clickthrough data," in *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, New York, NY, USA: ACM Press, 2002.

[112] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 154–161, New York, NY, USA: ACM, 2005.

[113] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search," *ACM Transactions on Information Systems*, vol. 25, no. 2, p. 7, 2007.

[114] T. Joachims, H. Li, T.-Y. Liu, and C. Zhai, "Learning to rank for information retrieval (lr4ir 2007)," *SIGIR Forum*, vol. 41, no. 2, pp. 58–62, 2007.

[115] T. Joachims and F. Radlinski, "Search engines that learn from implicit feedback," *Computer*, vol. 40, no. 8, pp. 34–40, 2007.

[116] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments," *Information Processing and Management*, vol. 36, no. 6, pp. 779–808, 2000.

[117] R. Jones, R. Kumar, B. Pang, and A. Tomkins, ""I know what you did last summer": Query logs and user privacy," in *CIKM '07: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 909–914, New York, NY, USA: ACM, 2007.

[118] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating query substitutions," in *WWW '06: Proceedings of the 15th International Conference on World Wide Web*, pp. 387–396, New York, NY, USA: ACM Press, 2006.

[119] R. Karedla, J. S. Love, and B. G. Wherry, "Caching strategies to improve disk system performance," *Computer*, vol. 27, no. 3, pp. 38–46, 1994.

[120] M. Kendall, *Rank Correlation Methods.* Hafner, 1955.

[121] J. Kleinberg, "Bursty and hierarchical structure in streams," in *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 91–101, New York, NY, USA: ACM, 2002.

[122] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[123] S. Koshman, A. Spink, and B. J. Jansen, "Web searching on the vivisimo search engine," *JASIST*, vol. 57, no. 14, pp. 1875–1887, 2006.

[124] M. Koster, "Aliweb: Archie-like indexing in the web," *Computer Networks and ISDN Systems*, vol. 27, no. 2, pp. 175–182, 1994.

[125] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 49–86, 1951.

[126] R. Kumar, J. Novak, B. Pang, and A. Tomkins, "On anonymizing query logs via token-based hashing," in *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pp. 629–638, New York, NY, USA: ACM, 2007.

[127] T. Lau and E. Horvitz, "Patterns of search: analyzing and modeling web query refinement," in *UM '99: Proceedings of the Seventh International Conference on User Modeling*, pp. 119–128, Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1999.

[128] U. Lee, Z. Liu, and J. Cho, "Automatic identification of user goals in web search," in *WWW '05: Proceedings of the 14th International Conference on World Wide Web*, pp. 391–400, New York, NY, USA: ACM, 2005.

[129] R. Lempel and S. Moran, "Predictive caching and prefetching of query results in search engines," in *WWW '03: Proceedings of the 12th International Conference on World Wide Web*, pp. 19–28, New York, NY, USA: ACM, 2003.

[130] R. Lempel and S. Moran, "Competitive caching of query results in search engines," *Theoretical Computer Science*, vol. 324, no. 2–3, pp. 253–271, 2004.

[131] R. Lempel and S. Moran, "Optimizing result prefetching in web search engines with segmented indices," *ACM Transactions on Internet Technology*, vol. 4, no. 1, pp. 31–59, 2004.

[132] R. Lempel and F. Silvestri, "Web search result caching and prefetching," Encyclopedia of Database Systems, Springer Verlag, 2008.

[133] M. Li, Y. Zhang, M. Zhu, and M. Zhou, "Exploring distributional similarity based models for query spelling correction," in *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 1025–1032, Morristown, NJ, USA: Association for Computational Linguistics, 2006.

[134] Y. Li, Z. Zheng, and H. K. Dai, "Kdd cup-2005 report: facing a great challenge," *SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 91–99, 2005.

[135] F. Liu, C. Yu, and W. Meng, "Personalized web search by mapping user queries to categories," in *CIKM '02: Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 558–565, New York, NY, USA: ACM Press, 2002.

[136] Live Search Team at Microsoft, "Local, relevance, and japan!," http://blogs.msdn.com/livesearch/archive/2005/06/21/431288.aspx, 2005.

[137] X. Long and T. Suel, "Three-level caching for efficient query processing in large web search engines," in *WWW '05: Proceedings of the 14th International Conference on World Wide Web*, pp. 257–266, New York, NY, USA: ACM, 2005.

[138] R. M. Losee and L. C. Jr., "Information retrieval with distributed databases: Analytic models of performance," *IEEE Transactions on Parallel & Distributed Systems*, vol. 15, no. 1, pp. 18–27, 2004.

[139] C. Lucchese, S. Orlando, R. Perego, and F. Silvestri, "Mining query logs to optimize index partitioning in parallel web search engines," in *InfoScale '07: Proceedings of the 2nd International Conference on Scalable Information Systems*, New York, NY, USA: ACM, 2007.

[140] T.-Y. Lui, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, 2008.

[141] Y. Lv, L. Sun, J. Zhang, J.-Y. Nie, W. Chen, and W. Zhang, "An iterative implicit feedback approach to personalized search," in *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pp. 585–592, Morristown, NJ, USA: Association for Computational Linguistics, 2006.

[142] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[143] M. Marchiori, "The quest for correct information on the web: Hyper search engines," *Computer Networks*, vol. 29, no. 8–13, pp. 1225–1236, 1997.

[144] E. P. Markatos, "On caching search engine query results," *Computer Communications*, vol. 24, pp. 137–143, 1 February 2000.

[145] M. Mat-Hassan and M. Levene, "Associating search and navigation behavior through log analysis: Research articles," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 9, pp. 913–934, 2005.

[146] O. A. McBryan, "Genvl and wwww: Tools for taming the web," in *Proceedings of the First International World Wide Web Conference*, (O. Nierstarsz, ed.), p. 15, CERN, Geneva, 1994.

[147] S. Melink, S. Raghavan, B. Yang, and H. Garcia-Molina, "Building a distributed full-text index for the web," *ACM Transactions on Information Systems*, vol. 19, no. 3, pp. 217–241, 2001.

[148] T. Mitchell, *Machine Learning*. McGraw-Hill International Editions, 1997.

[149] A. Moffat, W. Webber, and J. Zobel, "Load balancing for term-distributed parallel retrieval," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 348–355, New York, NY, USA: ACM, 2006.

[150] A. Moffat, W. Webber, J. Zobel, and R. Baeza-Yates, "A pipelined architecture for distributed text query evaluation," *Information Retrieval*, vol. 10, no. 3, pp. 205–231, 2007.

[151] A. Moffat and J. Zobel, "Information retrieval systems for large document collections," in *TREC*, 1994.

[152] E. J. O'Neil, P. E. O'Neil, and G. Weikum, "An optimality proof of the lru-k page replacement algorithm," *Journal of the ACM*, vol. 46, no. 1, pp. 92–112, 1999.

[153] S. Orlando, R. Perego, and F. Silvestri, "Design of a parallel and distributed WEB search engine," in *Proceedings of Parallel Computing (ParCo) 2001 conference*, Imperial College Press, September 2001.

[154] H. C. Ozmutlu, A. Spink, and S. Ozmutlu, "Analysis of large data logs: An application of poisson sampling on excite web queries," *Information Processing and Management*, vol. 38, no. 4, pp. 473–490, 2002.

[155] S. Ozmutlu, H. C. Ozmutlu, and A. Spink, "Multitasking web searching and implications for design," *JASIST*, vol. 40, no. 1, pp. 416–421, 2003.

[156] S. Ozmutlu, A. Spink, and H. C. Ozmutlu, "A day in the life of web searching: An exploratory study," *Information Processing and Management*, vol. 40, no. 2, pp. 319–345, 2004.

[157] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Technical Report, Stanford Digital Library Technologies Project, 1998.

[158] S. Pandey and C. Olston, "User-centric web crawling," in *WWW '05: Proceedings of the 14th International Conference on World Wide Web*, pp. 401–411, New York, NY, USA: ACM, 2005.

[159] S. Pandey and C. Olston, "Crawl ordering by search impact," in *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pp. 3–14, New York, NY, USA: ACM, 2008.

[160] G. Pass, A. Chowdhury, and C. Torgeson, "A picture of search," in *InfoScale '06: Proceedings of the First International Conference on Scalable Information Systems*, p. 1, New York, NY, USA: ACM, 2006.

[161] "Pew research center for the people & the press," WWW page, 2007. http://people-press.org/.

[162] J. Piskorski and M. Sydow, "String distance metrics for reference matching and search query correction," in *Business Information Systems, 10th International Conference, BIS 2007, Poznań, Poland, April 2007*, (W. Abramowicz, ed.), pp. 356–368, Springer-Verlag, 2007.

[163] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, "Personalized search," *Communications of the ACM*, vol. 45, no. 9, pp. 50–55, 2002.

[164] B. Poblete, M. Spiliopoulou, and R. Baeza-Yates, "Website privacy preservation for query log publishing," in *First International Workshop on Privacy, Security, and Trust in KDD (PINKDD'07)*, August 2007.

[165] S. Podlipnig and L. Böszörmenyi, "A survey of web cache replacement strategies," *ACM Computing Surveys*, vol. 35, no. 4, pp. 374–398, 2003.

[166] A. L. Powell and J. C. French, "Comparing the performance of collection selection algorithms," *ACM Transactions on Information Systems*, vol. 21, no. 4, pp. 412–456, 2003.

[167] A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles, "The impact of database selection on distributed searching," in *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–239, New York, NY, USA: ACM, 2000.

[168] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*. Cambridge University Press, Second ed., 1992.

[169] D. Puppin, "A search engine architecture based on collection selection," PhD thesis, Dipartimento di Informatica, Università di Pisa, Pisa, Italy, December 2007.

[170] D. Puppin and F. Silvestri, "The query-vector document model," in *CIKM '06: Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 880–881, New York, NY, USA: ACM, 2006.

[171] D. Puppin, F. Silvestri, and D. Laforenza, "Query-driven document partitioning and collection selection," in *InfoScale '06: Proceedings of the First International Conference on Scalable Information Systems*, p. 34, New York, NY, USA: ACM, 2006.

[172] D. Puppin, F. Silvestri, R. Perego, and R. Baeza-Yates, "Tuning the capacity of search engines: Load-driven routing and incremental caching to reduce and balance the load," *ACM Transactions on Information Systems*.

[173] D. Puppin, F. Silvestri, R. Perego, and R. Baeza-Yates, "Load-balancing and caching for collection selection architectures," in *InfoScale '07: Proceedings of the 2nd International Conference on Scalable Information Systems*, New York, NY, USA: ACM, 2007.

[174] F. Qiu and J. Cho, "Automatic Identification of User Interest for Personalized Search," in *WWW '06: Proceedings of the 15th International Conference on World Wide Web*, pp. 727–736, New York, NY, USA: ACM, 2006.

[175] F. Radlinski and T. Joachims, "Query chains: learning to rank from implicit feedback," in *KDD '05: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 239–248, New York, NY, USA: ACM Press, 2005.

[176] F. Radlinski and T. Joachims, "Active exploration for learning rankings from clickthrough data," in *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 570–579, New York, NY, USA: ACM, 2007.

[177] K. H. Randall, R. Stata, J. L. Wiener, and R. G. Wickremesinghe, "The link database: Fast access to graphs of the web," in *DCC '02: Proceedings of the Data Compression Conference (DCC '02)*, p. 122, Washington, DC, USA: IEEE Computer Society, 2002.

[178] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *SIGIR '94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, New York, NY, USA: Springer-Verlag New York, Inc., 1994.

[179] S. E. Robertson and S. Walker, "Okapi/keenbow at trec-8," in *TREC*, 1999.

[180] J. T. Robinson and M. V. Devarakonda, "Data cache management using frequency-based replacement," *SIGMETRICS Performance Evaluation Review*, vol. 18, no. 1, pp. 134–142, 1990.

[181] J. Rocchio, *Relevance Feedback in Information Retrieval.* Prentice-Hall, 1971.

[182] G. Salton and C. Buckley, "Parallel text search methods," *Communications of the ACM*, vol. 31, no. 2, pp. 202–215, 1988.

[183] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," *JASIS*, vol. 41, no. 4, pp. 288–297, 1990.

[184] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval.* New York, NY, USA: McGraw-Hill, Inc., 1986.

[185] M. Sanderson and S. T. Dumais, "Examining repetition in user search behavior," in *ECIR*, pp. 597–604, 2007.

[186] P. C. Saraiva, E. S. de Moura, N. Ziviani, W. Meira, R. Fonseca, and B. Ribeiro-Neto, "Rank-preserving two-level caching for scalable search

engines," in *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 51–58, New York, NY, USA: ACM, 2001.

[187] F. Scholer, H. E. Williams, and A. Turpin, "Query association surrogates for web search: Research articles," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 7, pp. 637–650, 2004.

[188] "Search engine use shoots up in the past year and edges towards email as the primary internet application," WWW page, 2005. http://www.pewinternet.org/pdfs/PIP_SearchData_1105.pdf.

[189] "Search engine users," WWW page, 2005. http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf.

[190] "Search engine users," White paper, 2005. http://www.enquiroresearch.com/personalization/.

[191] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[192] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang, "Q^2c@ust: Our winning solution to query classification in kddcup 2005," *SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 100–110, 2005.

[193] X. Shen, B. Tan, and C. Zhai, "Ucair: A personalized search toolbar," in *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 681–681, New York, NY, USA: ACM, 2005.

[194] M. Shokouhi, J. Zobel, and Y. Bernstein, "Distributed text retrieval from overlapping collections," in *ADC '07: Proceedings of the Eighteenth Conference on Australasian Database*, pp. 141–150, Darlinghurst, Australia: Australian Computer Society, Inc., 2007.

[195] M. Shokouhi, J. Zobel, S. Tahaghoghi, and F. Scholer, "Using query logs to establish vocabularies in distributed information retrieval," *Information Processing and Management*, vol. 43, no. 1, pp. 169–180, 2007.

[196] L. Si and J. Callan, "Using sampled data and regression to merge search engine results," in *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 19–26, New York, NY, USA: ACM, 2002.

[197] L. Si and J. Callan, "Relevant document distribution estimation method for resource selection," in *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 298–305, New York, NY, USA: ACM, 2003.

[198] S. Siegfried, M. J. Bates, and D. N. Wilde, "A profile of end-user searching behavior by humanities scholars: The getty online searching project report no. 2," *JASIS*, vol. 44, no. 5, pp. 273–291, 1993.

[199] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of a very large altavista query log," Technical Report, Systems Research Center — 130 Lytton Avenue — Palo Alto, California 94301, 1998.

[200] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," *SIGIR Forum*, vol. 33, no. 1, pp. 6–12, 1999.

[201] F. Silvestri, "High performance issues in web search engines: Algorithms and techniques," PhD thesis, Dipartimento di Informatica, Università di Pisa, Pisa, Italy, May 2004.

[202] F. Silvestri, "Sorting out the document identifier assignment problem," in *Proceedings of the 29th European Conference on Information Retrieval*, April 2007.

[203] F. Silvestri, S. Orlando, and R. Perego, "Assigning identifiers to documents to enhance the clustering property of fulltext indexes," in *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 305–312, New York, NY, USA: ACM, 2004.

[204] F. Silvestri, S. Orlando, and R. Perego, "Wings: A parallel indexer for web contents," in *International Conference on Computational Science*, pp. 263–270, 2004.

[205] D. D. Sleator and R. E. Tarjan, "Amortized efficiency of list update and paging rules," *Communications of the ACM*, vol. 28, no. 2, pp. 202–208, 1985.

[206] A. J. Smith, "Cache memories," *ACM Computing Surveys*, vol. 14, no. 3, pp. 473–530, 1982.

[207] M. Speretta and S. Gauch, "Personalized search based on user search histories," in *Web Intelligence*, pp. 622–628, 2005.

[208] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "From e-sex to e-commerce: Web search changes," *Computer*, vol. 35, no. 3, pp. 107–109, 2002.

[209] A. Spink, S. Koshman, M. Park, C. Field, and B. J. Jansen, "Multitasking web search on vivisimo.com," in *ITCC '05: Proceedings of the International Conference on Information Technology: Coding and Computing, (ITCC'05) Volume II*, pp. 486–490, Washington, DC, USA: IEEE Computer Society, 2005.

[210] A. Spink, H. C. Ozmutlu, and D. P. Lorence, "Web searching for sexual information: An exploratory study," *Information Processing and Management*, vol. 40, no. 1, pp. 113–123, 2004.

[211] A. Spink and T. Saracevic, "Interaction in information retrieval: Selection and effectiveness of search terms," *JASIS*, vol. 48, no. 8, pp. 741–761, 1997.

[212] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic, "Searching the web: the public and their queries," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 226–234, February 2001.

[213] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *SIGKDD Explorations*, vol. 1, no. 2, pp. 12–23, 2000.

[214] J. Teevan, E. Adar, R. Jones, and M. Potts, "History repeats itself: Repeat queries in yahoo's logs," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 703–704, New York, NY, USA: ACM, 2006.

[215] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts, "Information re-retrieval: Repeat queries in yahoo's logs," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 151–158, New York, NY, USA: ACM, 2007.

[216] J. Teevan, S. T. Dumais, and E. Horvitz, "Beyond the commons: Investigating the value of personalizing web search," in *Proceedings of Workshop on New Technologies for Personalized Information Access (PIA '05)*, Edinburgh, Scotland, UK, 2005.

[217] J. Teevan, S. T. Dumais, and E. Horvitz, "Personalizing search via automated analysis of interests and activities," in *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 449–456, New York, NY, USA: ACM Press, 2005.

[218] "The associated press: Internet ad revenue exceeds $21b in 2007," 2008. http://ap.google.com/article/ALeqM5hccYd6ZuXTns2RWXUgh6br4n1UoQ D8V1GGC00.

[219] H. Turtle and J. Flood, "Query evaluation: Strategies and optimizations," *Information Processing and Management*, vol. 31, no. 6, pp. 831–850, 1995.

[220] M. van Erp and L. Schomaker, "Variants of the borda count method for combining ranked classifier hypotheses," in *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition*, pp. 443–452, International Unipen Foundation, 2000.

[221] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 2nd ed., 1979.

[222] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos, "Identifying similarities, periodicities and bursts for online search queries," in *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, pp. 131–142, New York, NY, USA: ACM, 2004.

[223] M. Vlachos, P. S. Yu, V. Castelli, and C. Meek, "Structural periodic measures for time-series data," *Data Mining and Knowledge Discovery*, vol. 12, no. 1, pp. 1–28, 2006.

[224] D. Vogel, S. Bickel, P. Haider, R. Schimpfky, P. Siemen, S. Bridges, and T. Scheffer, "Classifying search engine queries using the web as background knowledge," *SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 117–122, 2005.

[225] X. Wang and C. Zhai, "Learn from web search logs to organize search results," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 87–94, New York, NY, USA: ACM, 2007.

[226] R. Weiss, B. Vélez, and M. A. Sheldon, "Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering," in *HYPERTEXT '96: Proceedings of the the Seventh ACM Conference on Hypertext*, pp. 180–193, New York, NY, USA: ACM, 1996.

[227] R. W. White, M. Bilenko, and S. Cucerzan, "Studying the use of popular destinations to enhance web search interaction," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 159–166, New York, NY, USA: ACM, 2007.

[228] R. W. White, M. Bilenko, and S. Cucerzan, "Leveraging popular destinations to enhance web search interaction," *ACM Transactions on the Web*, vol. 2, no. 3, pp. 1–30, 2008.

[229] R. W. White and D. Morris, "Investigating the querying and browsing behavior of advanced search engine users," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 255–262, New York, NY, USA: ACM, 2007.

[230] L. Xiong and E. Agichtein, "Towards privacy-preserving query log publishing," in *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007)*, (E. Amitay, C. G. Murray, and J. Teevan, eds.), May 2007.

[231] J. Xu and J. Callan, "Effective retrieval with distributed collections," in *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 112–120, New York, NY, USA: ACM, 1998.

[232] J. Xu and W. B. Croft, "Cluster-based language models for distributed retrieval," in *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 254–261, New York, NY, USA: ACM, 1999.

[233] J. Xu and W. B. Croft, "Improving the effectiveness of information retrieval with local context analysis," *ACM Transactions on Information Systems*, vol. 18, no. 1, pp. 79–112, 2000.

[234] J. L. Xu and A. Spink, "Web research: The excite study," in *WebNet 2000*, pp. 581–585, 2000.

[235] Yahoo! Grid, "Open source distributed computing: Yahoo's hadoop support," http://developer.yahoo.net/blog/archives/2007/07/yahoo-hadoop.html, 2007.

[236] Y. Yang and C. G. Chute, "An example-based mapping method for text categorization and retrieval," *ACM Transactions on Information Systems*, vol. 12, no. 3, pp. 252–277, 1994.

[237] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 271–278, New York, NY, USA: ACM, 2007.

[238] B. Yuwono and D. L. Lee, "Server ranking for distributed text retrieval systems on the internet," in *Proceedings of the Fifth International Conference on Database Systems for Advanced Applications (DASFAA)*, pp. 41–50, World Scientific Press, 1997.

[239] O. R. Zaïane and A. Strilets, "Finding similar queries to satisfy searches based on query traces," in *OOIS Workshops*, pp. 207–216, 2002.

[240] J. Zhang and T. Suel, "Optimized inverted list assignment in distributed search engine architectures," in *IPDPS*, pp. 1–10, 2007.

[241] Y. Zhang and A. Moffat, "Some observations on user search behavior," in *Proceedings of the 11th Australasian Document Computing Symposium*, Brisbane, Australia, 2006.

[242] Z. Zhang and O. Nasraoui, "Mining search engine query logs for query recommendation," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pp. 1039–1040, New York, NY, USA: ACM, 2006.

[243] Q. Zhao, S. C. H. Hoi, T.-Y. Liu, S. S. Bhowmick, M. R. Lyu, and W.-Y. Ma, "Time-dependent semantic similarity measure of queries using historical clickthrough data," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pp. 543–552, New York, NY, USA: ACM, 2006.

[244] Z. Zheng, K. Chen, G. Sun, and H. Zha, "A regression framework for learning ranking functions using relative relevance judgments," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 287–294, New York, NY, USA: ACM, 2007.

[245] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley, 1949.

[246] J. Zobel and A. Moffat, "Inverted files for text search engines," *ACM Computing Surveys*, vol. 38, no. 2, p. 6, 2006.