Webology, Volume 1, Number 2, December, 2004

Shifts in Search Engine Development: A Review of Past, Present and Future Trends in Research on Search Engines

Saeid Asadi
School of Information Technology & Electrical Engineering, The University of Queensland, Australia

Hamid R. Jamali
School of Library, Archive & Information Studies, University College London, United Kingdom

Received November 15, 2004; Accepted December 9, 2004

Abstract

The World Wide Web has developed fast and many people use search engines to capture information from the Web. This article reviews past, present and future of search engines. Papers published in four major Web and information management conferences were surveyed to track research interests in the last five years. Web search and information retrieval topics such as ranking, filtering and query formulation are still hot topics among researchers. The most important shifts and issues of the future of search engines are mentioned too. Search engine companies are trying to capture the Deep Web and extract structured data to offer high quality results. Using Web page structure, shared search engines, expert recommendations and different mobile search facilities seem to be features of the next generation of search engines.

Keywords

World Wide Web, Web searching, Information retrieval, Search engines, Personalization, Localized search, Federated search, Deep Web, Web page structure analysis, Structured data

---

1. Introduction

The World Wide Web with its short history has become a major area of interest for different people. Millions of people all around the world use the Web for their daily life needs and many companies invest on developing a better system for information retrieval on the Web. The rate of the growth of the Web is exponential. The number of domains from 16,300 in July 1992 increased to 30,000,000 in July 2001 (Gromov, 2002). A big issue is how to manage and control the unstructured and fast-developing body of the Web. While the Web adopted traditional information storage and retrieval methods as its basics, it was quickly obvious that those traditional methods do not work efficiently for the Web. Not surprisingly, a survey for Realnames (Sullivan, 2000) reports that 44% of users are frustrated by navigation and search engine use. Current search engines build indices

mostly based on keyword occurrence and frequency for query negotiation using these indices (Watters & Amoudi, 2003).

A review on the Web's short history and its popularity shows that not only many developments and innovations have been made in the past, but also there is an extensive desire among people and the business world to develop Web-based resources and services more and more. Search tools and services have developed synchronously and it seems that the Web is ruled and directed highly by the search engine industry. So far, we have seen emerging useful services for Web search tools. General search engines such as Google, MSN and AltaVista have covered a huge amount of information and resources for us. Specialized search engines can retrieve special subjects or specific forms of resources for us. Crawling, indexing, ranking and query formulation have improved and many special-purposed search engines have been developed. However, the more the Web grows the more problems in and needs for search engines appear. In this paper, we review previous challenges and changes of search engines and then we examine the current subjects of related research in the fields of Web search and information retrieval. Then, we will introduce new issues and the outlook of search engines.

2. Web Searching Overview

2.1 *A Brief History of Web Searching*: search engines as we know them today began to appear in 1994 when the number of HTTP resources increased (Schwartz, 1998). However, Internet search engines were in use before the emergence and growth of the Web. The first pre-Web search engine was Archie, which allowed keyword searches of a database of names of files available via FTP (Poulter, 1997). The first robot and search engine of the Web was Wandex, which was developed by Matthew Gray in 1993 (Wall, 2004). Since the appearance and exponential growth of the Web, hundreds of search engines with different features have appeared.

Primary search engines were designed based on traditional information retrieval methods. AltaVista, Lycos and Excite made huge centralized indices of Web pages. To answer a query, they simply retrieved results from their indexed databases and showed the cached pages based on keyword occurrence and proximity. While traditional indexing models have been successful in databases, it was revealed that these methods are not sufficient for a tremendously unstructured information resource such as the Web. The completeness of the index is not the only factor in the quality of search results. "Junk results" often wash out any results that a user is interested in. In order to increase the quality of search, Google made an innovative ranking system for the entire Web. PageRank used the citation graph of the Web and Google introduced link analysis in the search engine systems (Brin & Page, 1998). Other efforts have been made to customize and specialize search tools.

2.2 *Specialization*: specialization has been considered in the Web search industry and has occurred widely in different search tools and techniques. Web designers and search engine developers had to focus on special skills (Holzschlag, 2001). Beside designers' specialization, another important aspect of specialization concerns users. The Web is used by different people with different backgrounds and needs. The huge size of the Web and thousands of resources in each subject usually causes users' frustration. Special-purpose search engines such as local, music, image or even metasearch engines are some efforts for specializing the Web searching. As well, we have seen specialization in geographical search engines and personalized search services.

2.2.1 *Specialized Search*: the Web was initially designed to share textual documents but textual and non-textual documents developed synchronously. While there has been some success in developing search engines for text, search engines for other media on the Web (images, audio, and video) are still rare and not very powerful. People frequently need to

locate such materials to use them as a source of information or for illustrations (Kherfi et al, 2004). Multimedia appears to require greater interactivity between the user and search engine, relative to general Web searching. The increase in the query and session lengths and the increase in the number of results pages being viewed indicate this greater interactivity (Jansen et al, 2003). Many special search tools have been designed to find image, sound, music, movie and other multimedia resources. Google, AltaVista, Yahoo and other general search engines have developed search facilities for images. MSN can search movies. Different search engines have been designed specifically to search a certain kind of multimedia documents. Corbis, PicSearch, MusicSearch, FindSounds, AudioFind and WaveSearch are special multimedia search engines.

2.2.2 *Personalization*: different search engine users have different information needs and interests. A query with a search term such as "soccer" will have the same results for all people who are searching in the same time but with different purposes. Personalized or customized search tries to find out interests of a special user. Search history and user profile are two common ways for personalizing search results. A personalized search engine may track and record a user's search history in order to learn the user's long-term interests. In subsequent searches for the same query, the search engine refers to history of previous searches to find the pages that have been reviewed by users. The frequency of opening a page or the length of time a user sees a page can be taken into account to find out the importance of every page for that user. User profiles also can be used to represent users' interests and to infer their intentions for new queries. Liu et al (2002) made a personalized search tool with personal categorized profiles in which a user profile consists of a set of categories and for each category, a set of terms (keywords) with weights. Each category represents a user interest in that category. The weight of a term in a category reflects the significance of the term in representing the user's interest in that category.

2.2.3 *Location-Based Search*: it has been declared that one-third of Web search queries are related to a special location on the earth. While general search engines can be used for geospatial queries, their results are not very successful and reliable. Many queries have geospatial dimensions, when physical interaction is ultimately anticipated, such as driving, touring, or shipping. Online shopping, for example, is built on the premise that distance and location are irrelevant (with the possible exception of shipping charges) while tourism and onsite inspection of goods have a geospatial dimension and do depend on distance and location (Watters & Amoudi, 2003). Many algorithms have been applied to current search engines to solve the problem of location. Unlike online shopping Websites such as eBay (www.ebay.com), search engines have not been able to solve the issues of location-based searching properly. Google's local service is usable just for the United States and Canada (Perez, 2004). Yahoo and MSN have local Web pages for many countries but they offer only coarse-grained results. The most important issue with geospatial search is identifying and extracting correct addresses and location names from Web pages. Current algorithms are poor for addressing this problem.

3. The Survey

3.1. *Methodology*: all the papers presented in some of the top conferences in the fields of the Web and information management were reviewed in order to track the shifts in research interests in the Web information retrieval domain. WWW, SIGMOD, SIGIR and CIKM are four international conferences that have been examined in this study. We analysed subjects of papers published in these seminars in a 5-year period from 2000 to 2004. ACM Portal has covered full-text papers and posters presented in these conferences. The subject classification of ACM was used for classifying papers' topics. We considered as many topics as were mentioned for a paper in ACM. Therefore, many papers were classified under more than one subject.

3.2. *Findings*: CIKM, SIGMOD, WWW and SIGIR have mostly concentrated on information systems and services. Figure one shows that 70 percent of papers have focused on information systems, Web services, and information storage and retrieval techniques. In addition, as it is illustrated in Figure 2, we can see a big shift and increase in the research on information systems in 2001. The figure declined for next two years; however, in 2004 there was another slight growth in the number of papers on this topic.

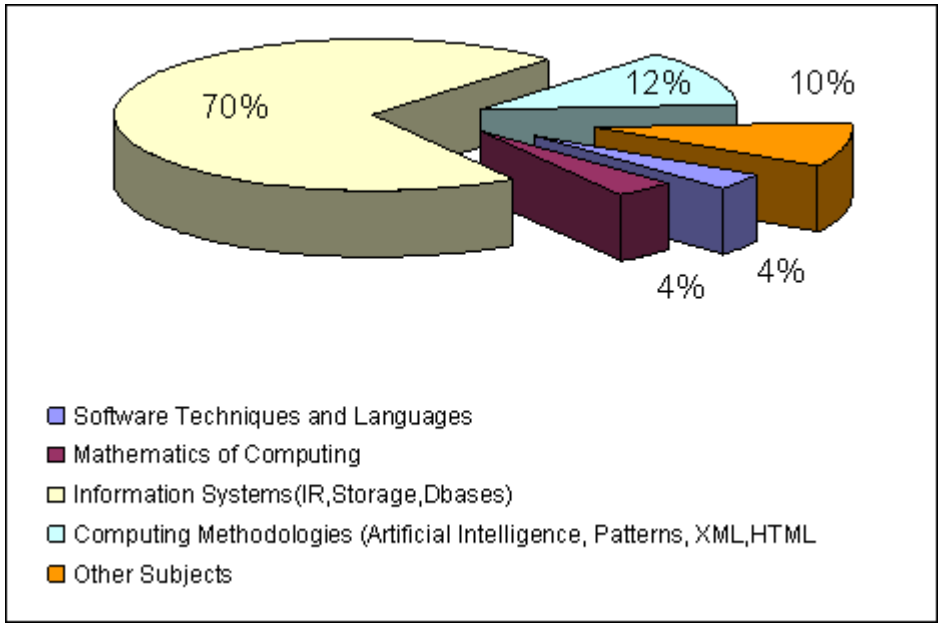

Fig. 1 - Major Topics Published in CIKM, SIGIR, SIGMOD and WWW from 2000 to 2004

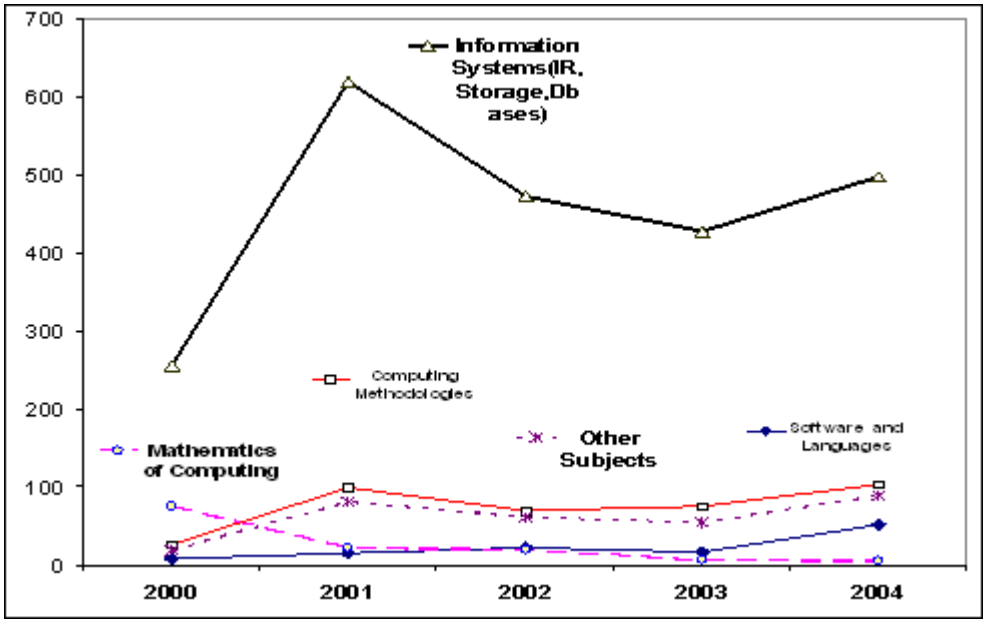

Fig. 2- Subjects of Research Papers Published in Different Years

At this stage of the study, the authors ignored all other areas other than the area of Information Systems. We divided remaining papers in five categories: models and principles, database management, applications of information systems, interfaces and presentation techniques, and finally information storage and retrieval (figure 3).

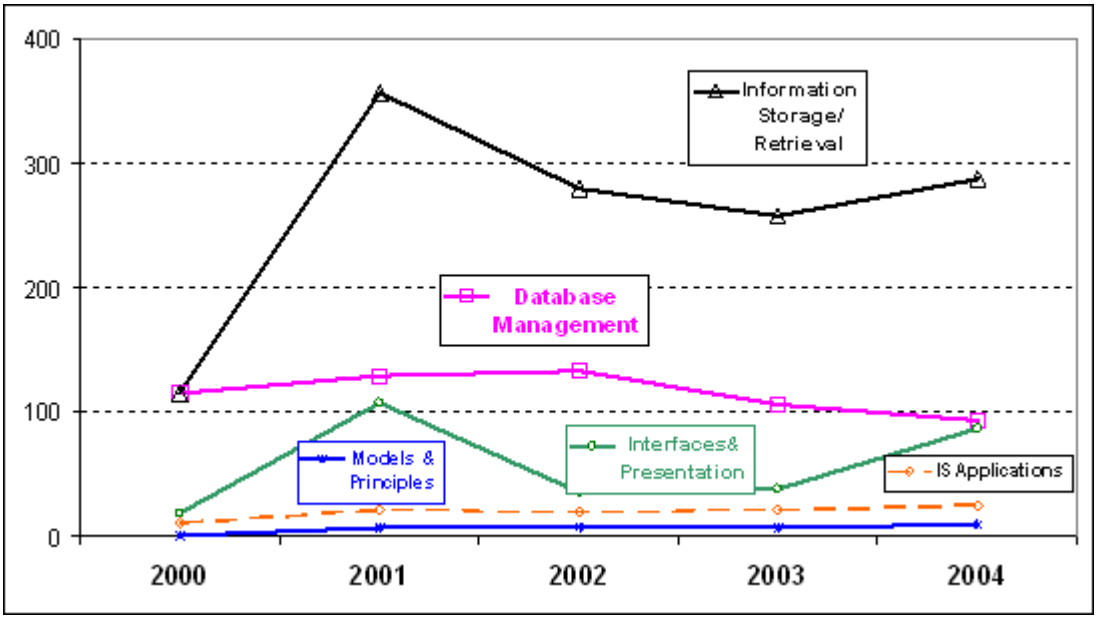

Fig. 3 - Different Issues of Information Systems Published in 2000-2004

The most studied area of research is information storage and retrieval. Database management has been another field of interest for researchers. We can see a considerable increase in the number of papers on interfaces & presentation of information.

Because researchers paid special attention to information storage and retrieval, we broke this area down to more specific subject areas. Indexing and abstracting, content analysis, information search and retrieval, systems and software evaluation, online systems including the World Wide Web and finally digital libraries are the most distinguishable fields among research on information storage and retrieval. As it is illustrated in Figure 4, many researchers have focused on information retrieval and Web searching. Designing and implementing different distributed systems and online databases and resources are still favourable.

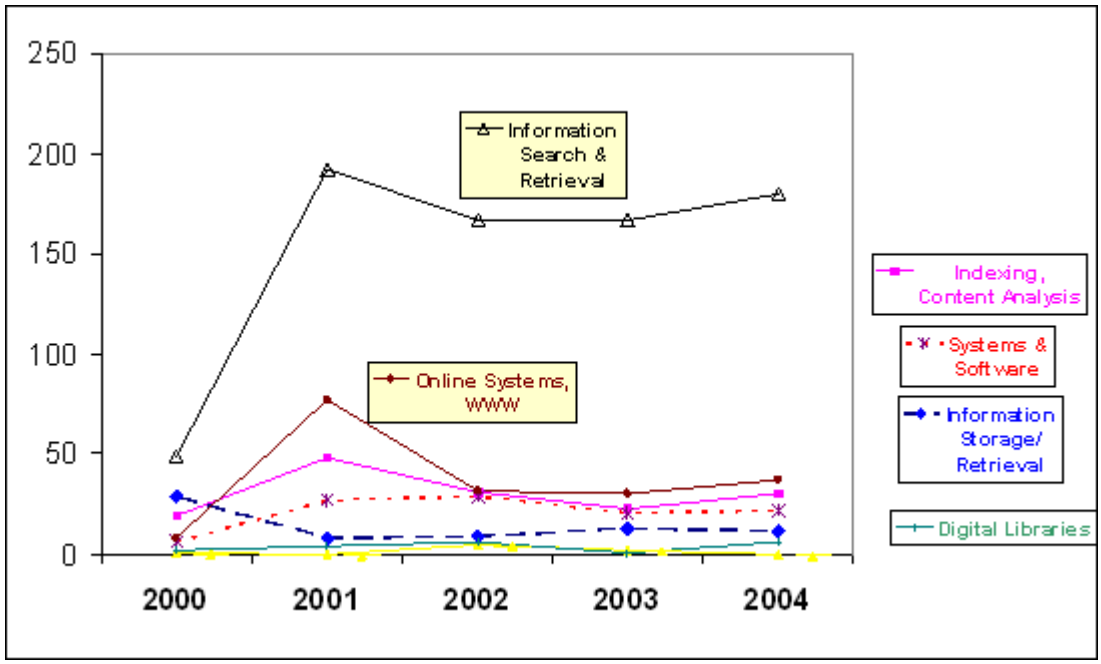

Fig. 4 - Different Aspects of Information Storage and Retrieval

Finally, we focused on papers with subjects such as information retrieval or different technical issues of Web search tools. It was difficult to divide those papers in different categories. Therefore, we decided to classify them based on an IT-based approach. Query formulation, search process and algorithms, relevance feedback and ranking of results, retrieval models, information filtering, clustering and Web page selection were the most important specific topics in Web search and information retrieval.

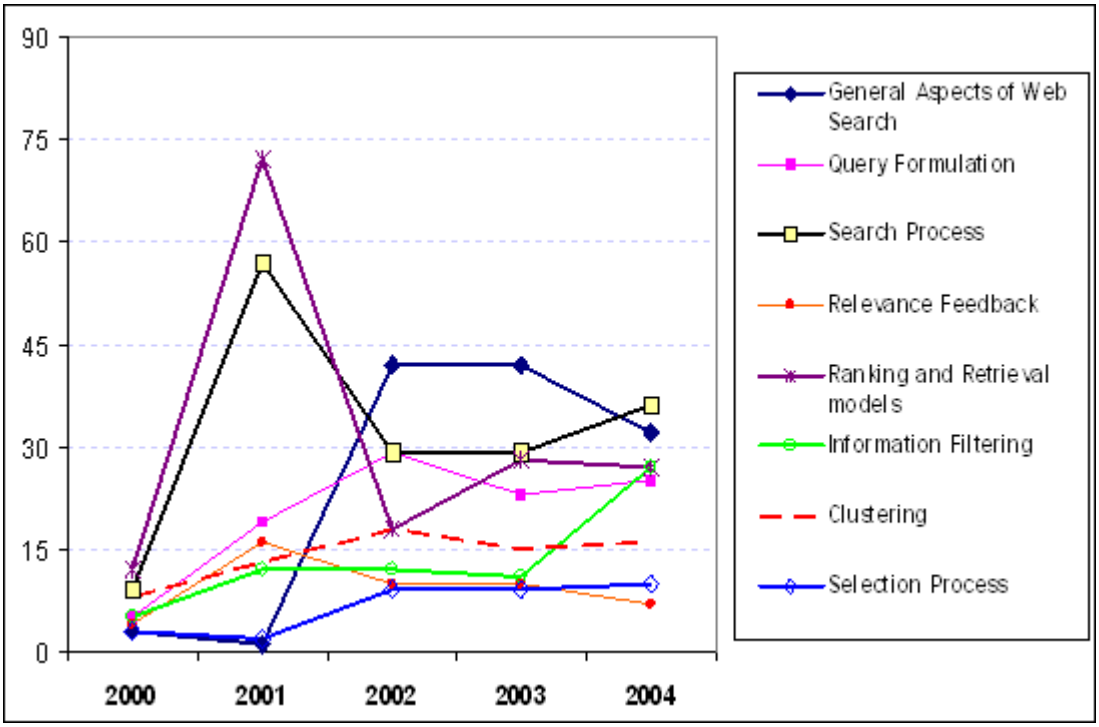

Fig. 5 - Different Aspects of Information Search and Retrieval

Figure 5 shows that retrieval models and search results ranking methods were most examined and on demand topics in the Web search and information retrieval domain. The peak of research on search processes, retrieval models and rank methods was in 2001 and after that, there was a drop in the amount of research in these areas. This can be interpreted either as researchers have succeeded in solving many technical problems of search systems or that the technical and hardware issues are too difficult to be resolved with current knowledge and technology. Also there has been a big focus on search process, crawling and indexing of Web pages. Query formulation and general aspects of Web search such as personalization, and Website design and classification have been hot topics as well.

4. New Features for Web Searching

The incredible development of Web resources and services has become a motivation for many studies and for companies to invest on developing new search engines or adding new features and abilities to their search engines. By looking at the papers published in the mentioned conferences and other journals and seminars, we can track several specifications and shifts in the future. Ma (2004) from Asian Microsoft Research Centre reported features of the next generation of search engines in WISE04. Deep Web with structured information is a potential resource that search companies are trying to capture. Meanwhile, researchers have focused on Web page structure to increase the quality of search. Microsoft has started a big competition on Web searching through working on Web page blocks, the Deep Web and mobile search. MSN new ranking model will be based on object-level ranking rather than document-level.

4.1 *Page Structure Analysis*: the first search engines concentrated on Web page contents. AltaVista and other old search engines were made based on indexing the content of Web pages. They built huge centralized indices and this is still a part of every popular search engine. However, it was clear that the contents of a Web page could not be sufficient for capturing the huge amount of information. In 1996-1997 Google was designed based on a novel idea that the link structure of the Web is an important resource to improve the results of search engines. Backlinks were used based on the Hyperlink-Induced Topic Search (HITS) algorithm to crawl billions of Web pages. Google not only used this approach to capture the biggest amount of Web pages but also established PageRank - the ranking system that improved the search results (Brin & Page, 1998). After content-based indexing and link analysis the new area of study is page and layout structures. HTML and XML are important in this approach. It is thought that Web page layout is a good resource for improving search results. For example, the value of information presented in < heading > tags can be more than information in < paragraph > tags. We can imagine also that a link in the middle of Web page is more important than a link in footnote. Web Graph algorithms such as HITS might be implemented to a sub-section of Web pages to improve search result ranking models. The automatic thesaurus construction method is a page structure method, which extracts term relationships from the link structure of Websites. It is able to identify new terms and reflect the latest relationship between terms as the Web evolves. Experimental results have shown that the constructed thesaurus, when applied to query expansion, outperforms traditional association thesaurus (Chen et al, 2003).

4.2 *Deep Search*: current search engines can only crawl and capture a small part of the Web, which is called the "visible" or "indexable" Web. A huge amount of scientific and other valuable information is behind closed doors. It is believed that the size of invisible or deep Web is several times bigger than the size of the surface Web. Different databases, library catalogues, digital books and journals, patents, research reports and governmental archives are examples of resources that usually cannot be crawled and indexed by current search engines. Web content providers are moving toward Semantic Web by applying technologies such as XML and RDF (Resource Description Framework) in order to create more structured Web resources. New search engines are trying to find suitable methods for penetrating the database barriers. BrightPlanet's "differencing" algorithm is designed to transfer queries across multiple deep Web resources at once, aggregating the results and letting users compare changes to those results over time. Google, MSN and many other popular search engines are competing to find solution for the invisible Web. Recently, Yahoo has developed a paid service for searching the deep Web that is called the Content Aggregation Program (CAP). The method is secret but the company does acknowledge that its Content Aggregation Program will give paying customers a more direct pipeline into its search database (Wright, 2004).

4.3 *Structured Data*: the World Wide Web is considered a huge collection of unstructured data presented in billions of Web pages. As we already mentioned, the amazing size and valuable resources of the deep Web have affected the industry of search engines and the next generation of search engines are supposed to be able to investigate deep Web information. As a part of both surface and deep Web, structured data resources are very important and valuable. In many cases, data is stored in tables and separated files. The concept of *structured searching* is different from the way search engines currently operate. Most of search engines just save a copy of Web pages in their repository and then make several indexes from the content of these pages. Most documents available on the Web are unstructured resources. So, search engines can just judge them based on the keyword occurrence. As Rein (1997) says a search engine supporting XML-based queries can be programmed to search structured resources. Such an engine would rank words based on their location in a document, and their relation to each other, rather than just the number of times they appear. Traditional information retrieval and database management techniques

have been used to extract data from different tables and resources and combine them to respond users' queries. Current search engines cannot resolve this problem efficiently, but in the future an intelligent search engine will be able to distinguish different structured resources and combine their data to find a high quality response for a complicated query.

4.4 *Recommending Group Ranking*: while many search engines are able to crawl and index billions of Web pages, sorting the results of each query is still an issue. Page ranking algorithms have been utilized to present a better ranked result. The idea is simple: more relevant pages must take a higher rank. Basic ranking algorithms are based on the occurrence rate of index terms in each page. Simply, if the search term is *mathematics* then a page that has the word *mathematics* 20 times must be ranked before a page which has *mathematics* 10 times. As we already mentioned, this alone is not a sufficient way; recently link information and page structure information have been used to improve rank quality. These methods are automatic and are done by machines. However, it is believed that the best judgement about the importance and quality of Web pages is acquired when they are reviewed and recommended by human experts. Discussion thread recommendation or peer reviews are expected to be used by search engines to improve their results. In the future, search results will be ranked not only based on the automatic ranking algorithms but also by using the ideas of scholars and scientific recommending groups.

4.5 *Federated Search*: also known as parallel search, metasearch or broadcast search, it aggregates multiple channels of information into a single searchable point. Federated search engines are different from metasearch engines. Metasearch engines services for users are free while federated search engines are sold to libraries and other interested information service providers. Federated searche mostly covers subscription based databases that are usually a part of Invisible Web and ignored by Web-oriented metasearch engines. Usually there is no overlap between databases covered by federated search engines. Federated searching has several advantages for users. It reduces the time that is needed for searching several databases and also users do not need to know how to search through different interfaces (Fryer, 2004). One of the important reasons of the growing interest in federated searching is the complexity of the online materials environment such as the increasing number of electronic journals and online full-text databases. Webster (2004) maintains that although federated searching tools offer some real immediate advantages today, they cannot overcome the underlying problem of growing complexity and lack of uniformity. We need an open interoperable and uniform e-content environment to provide fully the interconnected assessable environment that librarians are seeking from metasearching. One of the disadvantages of federated search engines is that they cannot be used for sophisticated search commands and queries, and are limited to basic Boolean search.

4.6 *Mobile Search*: the number of people who have a cell phone seems to be more than the number of people who have a PC. Also many other mobile technologies such as GPS devices are used widely. Search engine companies have focused on the big market of mobile phones and wireless telecommunication devices. In the future everyone will have access to the Web information and services through his/her wireless phone without necessarily having a computer. Recently, Yahoo developed its mobile Web search system and mobile phone users can have access to Yahoo Local, Image and Web search, as well as quick links to stocks, sports scores and weather for fee. The platform also includes a modified Yahoo Instant Messaging client and Yahoo Mobile Games (Singer, 2004).

5. Conclusion

The World Wide Web with its short history has experienced significant changes. While the first search engines were established based on the traditional database and information retrieval methods, many other algorithms and methods have since been added to them to improve their results. The gigantic size of the Web and vast variety of the users' needs and interests as well as the big potential of the Web as a commercial market have brought about many changes and a great demand for better search engines. In this article, we reviewed the history of Web search tools and techniques and mentioned some big shifts in this field. Google utilized Web graph or link structure of the Web to make one of the most comprehensive and reliable search engines. Local services and the personalization of search tools are two major ideas that have been studied for several years.

By looking at papers published in popular conferences on Web and information management, we see not only a considerable increase in the quantity of Web search research papers since 2001, but also we can see that Web search and information retrieval topics such as ranking, filtering and query formulation are still hot topics. This reveals that search engines have many unsolved and research-interesting areas.

We mentioned several important issues for the future of search engines. The next generations of search tools are expected to be able to extract structured data to offer high quality responses to users' questions. The structure of Web pages seems to be a good resource with which search engines can improve their results. As well, there will be a shift towards providing specialised search facilities for the scholarly part of the Web that encompasses a considerable part of the deep Web. Having the Beta version of Google Scholar (http://scholar.google.com) released in November 2004, other major players in search engine industry are expected to invest on rivals for this new service. Search engines are trying to consider recommendations of special-interest groups into their search techniques. Limitation in funds has enforced libraries and other major information user organizations to share their online resources. Federated search is a sample of future cooperative search and information retrieval facilities. Finally, we addressed the efforts of search engine companies in breaking their borders through making search possible for mobile phones and other wireless information and communication devices.

The World Wide Web will be more usable in the future. The Web's security and privacy are two important issues for the coming years. Web search industry is opening new horizons for the global village. Meanwhile many issues have remained unsolved or incomplete still. Information extraction, ambiguity in addresses and names, personalization and multimedia searching among others are major issues in the next few years.

References

- Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International WWW Conference*, Brisbane, Australia, 107-117.
- Chen, Z., Liu, S., Wenyin, L., Pu, G. & Ma, W. (2003). Building a web thesaurus from web link ltructure. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. Toronto, 48– 55.
- Fryer, D. (2004). Federated search engines. *Online*, 28(2), 16-19.
- Gromov, G. R. (2002). History of Internet and WWW: the roads and crossroads of Internet history. Retrieved December 5, 2004, from http://www.netvalley.com/intvalstat.html
- Holzschlag, M. E. (2001). How specialization limited the Web. Retrieved December 4, 2004, from http://www.webtechniques.com/archives/2001/09/desi/
- Jansen, B. J., Spink, A. & Pedersen, J. (2003). An analysis of multimedia searching on AltaVista. *Proceedings of the 5th ACM SIGMM international workshop on*

Multimedia information retrieval, 186-192.
- Kherfi, M. L., Ziou, D. & Bernardi, A. (2004). Image retrieval from the World Wide Web: issues, techniques and systems. *ACM Computer Surveys*, 36(14), 35-67.
- Liu, F., Yu, C. & Meng, W. (2002). Personalized web search by mapping user queries to categories. *Proceedings of the eleventh international conference on Information and knowledge management CIKM'02*, McLean, Virginia, USA, 558-565.
- Ma, W., Zhang, H. and Hon, H. (2004). Towards Next Generation Web Information Retrieval. *Web Information Systems – WISE04: Proceedings of the fifth international Conference on Web Information System Engineering* , Brisbane, Australia, 17.
- Perez, C. (2004). Google offers new local search service. Retrieved December 2, 2004, from http://www.infoworld.com/article/04/03/17/HNgooglelocal_1.html
- Poulter, A. (1997). The design of World Wide Web search engines: a critical review, *Program*, 31(2), 131-145.
- Rein, L. (1997). XML Ushers in Structured Web Searches. Retrieved November 20, 2004, from http://www.wired.com/news/technology/0,1282,7751,00.html
- Schwartz, C. (1998). Web search engines. *Journal of the American Society for Information Science*, 49(11), 973-982.
- Singer, M. (2004, October 27). Yahoo sends search aloft. Retrieved November 28, 2004, from http://www.internetnews.com/bus-news/article.php/3427831
- Sullivan, D. (2000, June 2). Survey reveals search habits. *The Search Engine Report*. Retrieved December 1, 2004, from http://www.searchenginewatch.com/sereport/00/06-realnames.html
- Wall, A. (2004). History of search engines & web history. Retrieved December 3, 2004, from http://www.search-marketing.info/search-engine-history/
- Watters, C. & Amoudi, G. (2003). GeoSearcher: location-based ranking of search engine results. *Journal of the American Society for Information Science and Technology*, 54(2), 140-151.
- Webster, P. (2004). Metasearching in an academic environment. *Online*, 28(2), 20-23.
- Wright, A. (2004, March 9). In search of the deep Web. Retrieved December 5, 2004, from http://www.salon.com/tech/future/2004/03/09/deep_web/index_np.html

---

Bibliographic information of this paper for citing:

Asadi, S., & Jamali, H.R. (2004). "Shifts in search engine development: A review of past, present and future trends in research on search engines". *Webology*, **1**(2), Article 6. Available at: http://www.webology.org/2004/v1n2/a6.html

---

This article has been cited by other articles.
