New Methods and Tools for the World Wide Web Search

Vlatko Ceric
Faculty of Economics, University of Zagreb
Kennedyjev trg 6, 10000 Zagreb, Croatia
E-mail: vceric@efzg.hr

Abstract: *Explosive growth of the World Wide Web as well as its heterogeneity asks for powerful and intuitive methods and tools for Web retrieval that provide provide the user with a moderate number of relevant answers. This paper presents some of the innovative Web search methods and tools, especially in the area of subject trees and search engines. In the area of subject trees we will pay special attention to Hyperbolic Tree that enables visual intuitive approach in exploring large quantities of data by presenting focused part of the tree in the framework of the unfocused data set, and to ThemeScape that produces a topological map of textual documents based on the similarity of their content. In the area of subject trees we will discuss the Northern Light search engine that automatically groups resulting Web documents in meaningful categories, Clever project that uses hyperlink Web structure in identifying "authorities" and "hubs" Web pages and thus enables finding of high quality answers to user request, as well as to Simpli.com and Oingo search engines that try to resolve the problem of lexical ambiguity of search terms.*

Key words: World Wide Web, search, ranking of results, search engines, subject trees

1. Introduction

World Wide Web had an astonishing growth in the last few years. Number of Web servers increased from 1.7 million in December 1997, to 3.7 million in December 1998, and then to 9.6 million in December 1999[1]. The same source estimates that the number of Internet hosts increased from 29.7 million in January 1998, to 43.2 million in January 1999, and then to 72.4 million in January 2000. The number of people online is assessed to be about 275 million as of March 2000[2].

It was estimated (Lawrence and Giles, 1999) that in February 1999 the publicly indexable Web contained approximately 800 million pages, the quantity of information equivalent to 6 terabytes of pure text, as well as some 180 million images equivalent to about 3 terabytes (Library of Congress, the biggest library today, contains about 20 terabytes of text). No single search engine was indexing more than about 16% of Web sites. Combined, major search engines were covering only about 42% of the Web.

Without software tools for searching and cataloguing of the huge and heterogeneous Web information space it would be impossible to find relevant information residing on the Web (Baeza-Yates and Ribeiro-Neto, 1999; Ceric, 2000). The need for retrieval of information on the Web is so large that the major search engines receive about 15-20 thousand queries per minute. Users need powerful, easy to use search tools capable to provide a moderate number of appropriately ranked relevant answers.

In the last few years numerous advancements were made in various areas of the World Wide Web search. This paper presents analysis of key aspects of recently developed Web search methods and tools, as well as most important trends in Web search methods.

---

[1] Hobbes' Internet Timeline, http://www.isoc.org/zakon/Internet/History/HIT.html

[2] Nua Ltd., http://www.nua.ie/surveys/how_many_online/

2. Subject trees

Both subject trees (directories) and search engines have the problem to follow the extremely fast growth of the Web. Since Web links are brought into subject trees by human indexing teams, the size of the team determines the coverage of Web contents by the subject tree. Biggest subject trees have the following approximate sizes (at the end of 1999): Yahoo! has the staff of about 150 editors and has about 1.2 million Web links, LookSmart has about 200 editors and consist of about 1 million Web links, while Snap has about 60 editors and contains about 400,00 Web links. Special case is Open Directory with some 16,000 editors and content of about 1 million Web links. One of the most valuable mid-sized subject trees Britannica contains about 130,000 Web links.

Subject trees have become so big that the traditional user interfaces via hierarchic trees became inadequate for dealing with information since they don't enable the view of the whole information structure, and navigation is done by slow fold-and-unfold process. Such situation led to development of user interfaces with novel representations of the subject tree structure, like Hyperbolic Tree and ThemeScape.

Hyperbolic Tree[3], developed by the Xerox Palo Alto Research Center, is user interface designed for managing and exploring large hierarchies of data such as Web subject trees, product catalogs or document collections. It enables an overview of the whole data set as well as drilling down and accessing the primary information (e.g. specific Web links). Hyperbolic Tree is an intuitive and highly interactive tool that allows focusing on the part of data represented in context of the whole unfocused data set. User can easily change the focus by animated moving of the elements of the hyperbolic tree by means of click-and-drag with the mouse.

ThemeScape[4] organizes a topographical map of the textual documents based on the similarity of their content. This software reads documents from the data set and examines their contents using sophisticated statistical and natural language filtering algorithms. Related topics found in the documents are associated, forming different concepts. These concepts are assigned to the spatial structure that is transformed into topological maps in such a way that the greater the similarity between two documents the closer together they appear on the map. Peaks on the map denote the concentration of several documents about the same topic. Resulting topographical map clearly shows which topics are covered in the data set, how much emphasis is given to the specific topic, and how different topics relate to one another. Search across the documents results in highlighting of relevant documents right on the map, instead of showing the list of discovered documents.

3. Search engines

Fast growth of the Web led to the growth of search engines data bases. Several major search engines have the following size of databases (as of April, 2000): INKTOMI database has 500 million Web pages indexed, FAST has 340 million pages, AltaVista has 250 million pages and Northern Light has 240 million pages.

One of the most important things for efficient use of the Web search is suitable ranking of the resulting Web links. Two new approaches developed in this field were done by Northern Light and Google search engines. *Northern Light* introduced grouping of resulting Web links into

---

[3] Inxight: Hyperbolic Tree, http://www.inxight.com/products/developer/hyperbolic_tree.html

[4] Cartia: ThemeScape, http://www.cartia.com/products/index.html

meaningful categories, so called Custom Search Folders, that contain only information relevant to that folder. These folders group results by subject, type (e.g. press releases or product reviews), source (e.g. commercial sites or magazines) and language. Users can concentrate on links from the appropriate folder and thus narrow the results considerably and save time. Folders are not preset, but rather dynamically created for each search. *Google*[5] search engine uses the PageRank algorithm based on the importance (popularity) of Web pages. Importance of Web pages is discovered through analysis of the Web link structure, and doesn't depend on the specific search request. PageRank is a characteristic of the Web page itself – it is higher if more Web pages link to this page, as well as if these Web pages have high PageRank. Therefore, important Web pages help to make other Web pages important. One consequence of this approach is that search result may include links to Web pages that were not found by Google spider, but were linked by some Web page accessed by the spider.

Simpli.com and Oingo are search engines dealing with the problem of lexical ambiguity of the traditional search with one or more search terms put out of context. Since words can have many different meanings (polysemy) traditional search engines respond by selecting all Web pages that fits each possible meaning of the search terms. This leads to a large list of resulting links that includes many links not related to the desired meaning of the search. Because there are many words with the same or similar meaning (synonymy) it can happen that keyword-based search cannot find relevant pages because it uses one synonym while authors of Web pages may use some other.

Simpli.com[6] is using principles of linguistic and cognitive science as well as the interaction with users to place search terms in context. After the user enters the search term Simpli.com activates its knowledge base and automatically generates the list of possible contexts of the search term. User than interacts with the search engine and chooses the appropriate meaning (concept). Search engine now consults its database to choose related words based on the search term and the chosen concept, and automatically expands the query with these words.

Oingo[7] takes the similar approach and enables using more than one keyword for search. Oingo offers the user a list with possible meanings for all of the terms used in a query. The user than chooses the exact meaning for each query term, and these meanings are used for the search.

4. Trends

Several new approaches to search engines mechanisms were developed around the idea of exploiting rich Web hyperlinking structure for clustering and ranking of Web documents. *Clever* project (Members of the Clever team, 1999), run by researchers from IBM, Cornell University and the University of California at Berkeley, analysis Web hyperlinks and automatically locates two types of pages: "authorities" and "hubs". "Authorities" are the best sources of information on a particular broad search topic, while "hubs" are collection of links to these authorities. A respective authority is a page that is referred to by many good hubs, while a useful hub is a page that points to many valuable authorities. For any query Clever first performs an ordinary text-based search using an engine like AltaVista, and takes a list of 200 resulting Web pages. This set of links is then expanded with Web pages linked to and from those 200 pages – this step is

[5] Google, http://www.google.com

[6] Simpli.com, http://www.simpli.com

[7] Oingo, http://www.oingo.com

repeated to obtain a collection of about 3,000 pages. Clever system then analysis the interconnections between these documents, giving higher authority scores to those pages that are frequently cited, and higher hub scores to pages that link to those authorities. This procedure is repeated several times with iterative adjusting of authorities and hub scores: authority that has many high-scoring hubs pointing to it earns a higher authority score, while a hub that points to many high-scored authorities gets a higher hub score. The same page can be both the authority and the hub. A side-effect of this iterative processing is that the algorithm separates Web sites into clusters of similar sites. Clever system is also used for automatic compilation of lists of Web resources similar to subject trees – these lists appear to be competitive with handcrafted ones.

Traditional search in the batch mode leads to the situations where users may miss recently-added Web pages because of the slow update of search engines data bases (crawlers often revisit the same Web site after one or two months). This led IBM to develop the *Fetuccino*[8] software, an interesting combination of traditional search in batch mode and dynamic search in real time. In the first stage some traditional batch search engine is exploited. In the second stage results obtained in the first stage are refined via dynamic searching and crawling. This is done by feeding the search results from the first phase to the Mapuccino's dynamic mapping system which augments those results by dynamically crawling in directions where relevant information is found.

5. Conclusions

In the last few years significant improvements were made in a number of areas of the World Wide Web search. This paper first presented the current state of the World Wide Web, the speed of its development, as well as its heterogeneity. After that, the paper discussed some of the most advanced recent approaches to the Web search, as well as various novel operational Web search elements or systems in the area of subject trees, search engines and database search.

There is certainly room for further advancements of both technology and services. One type of service that would be extremely helpful for professionals are *specialty search services* that would exclusively cover the domain of interest of specific groups of professionals. Coverage should be deep, and information should be fresh and almost without any dead links. *Academic search engine* is another higly important service that should cover academic Web sites. It should carry out a deep and very frequent crawl of these Web sites. Covering a fresh state of research and research publishing could e.g. lead to decrease duplication of research work.

References

1. Baeza-Yates, R. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley, Harlow, England.

2. Ceric, V. (2000), Advancements and trends in the World Wide Web search, *Proc. 20th Internat. Conf. on Information Technology Interfaces ITI'2000*, Pula, Croatia, 211-220.

3. Lawrence, S. and Giles, L. (1999), Accessibility of information on the web, *Nature* **400**, 107-109.

[8] IBM: Fetuccino, http://www.ibm.com/java/fetuccino/index.html

4. Members of the Clever team (1999), Hypersearching the Web, *Scientific American*, June 1999. (Online at http://www.sciam.com/1999/0699issue/0699raghavan.html)