A Big Leap Forward - The Next Step of Educational Information Retrieval

Hendrik Speck, Frédéric Philipp Thiele, Sven Wagenhöfer
University of Applied Science Kaiserslautern, Germany
E-Mail: contact@hendrikspeck.com, contact@fp-thiele.de, sven@wagenhoefer.com

Abstract: This paper discusses the current implications of global internet search engines and their hidden algorithms on education and scientific research, including the changing student behavior in research. It presents the development of the search engine market and its political implications along with specific problems that current centralized search engines have and how new information becomes rather unavailable through current search engine popularity algorithms. It proposes a new decentralized open source engine which will circumvent these problems and grant institutions like universities a way to gain access to a specialized search system with unique characteristics for their sole purpose.

Historical Context: Search Engine Development

In 1989 the World Wide Web was born, based on concepts developed Bush Vannemar, using technologies developed in the 1960s by ARPA and the ARPANET, and finally refurbished by Tim Berners-Lee at the CERN Institute. In 1993 the new medium gained momentum with the publication of the mosaic browser by Marc Andreesen – it received it's role as mass media and more and more information was added. The explosion of content meant an increasing amount of knowledge stored in the networks.

At this time, no possibility of searching into this vast information was available – the user had to get to it directly via URLs and Links. To bring order to chaos, 1994 Lycos started as first WWW-search-engine with a graphical user interface from a Carnegie-Mellon-University-Project with 54,000 pages in its index. The idea became a trend with the founding of other search-engines like Infoseek, Altavista, HotBot and Excite. Another concept led the former market leader yahoo to success – a human-indexed, high quality web catalogue. In 1999 Yahoo had a market share of 50 percent, the rest was divided within Altavista, Excite and some minor search engines.

In September 1999 Google started officially, but at first without great importance. With the domination of Internet Explorer, Microsofts MSN could hold position two behind Yahoo in 2000. But this success did not last long – slowly but certain Google fought its way up to the top. As the Yahoo management recognized the importance of full-text-search-engines and abandoned the web catalogue, they integrated a google-based search into their portal. With the increased popularity, Google could displace MSN from the second rank – even market leader Yahoo had to face a decrease in popularity and stranded at about 40 percent market share.

Till the end of 2002 Google was able to gain a market share of 50 percent and so become the most popular search engine, leaving Yahoo with 20, MSN at about 15 and AOL search at around 5 percent. Midst 2002 AOL also decided to buy its search-results from Google. So, the market leader is currently at about 50 percent direct share, and including the results other companies buy from them, at about 75 percent of market share. As only real competitor remained MSN with about 15 percent of market share. The remaining 10 percent are divided into several smaller, most special interest, search-engines.

This monopoly greatly changed user behavior on the net – direct links become less important than links from the market leader Google – one reason, why people pay for advertisement or sponsored links there.

The question remains, what quality these links have – especially for educational purposes or scientific research.

Current Educational Research Problems

Most students start their research on the internet. Current statistics demonstrate, that even the media literate among them have changed their usage profiles. Students are less inclined to read newspapers, periodicals, books and encyclopedias, instead they consume television, multi media, video and computer games - they prepare their assignments on the internet. Many term papers actually start with the very same search engine, a market leader and monopoly that has "googled" its way even into the (paper based) Encyclopedia Britannica – redefining the word “search” once and forever.

This trend is actually supported by e-learning, where the web becomes the contact between the students and their study-issue. So we see that search engines are crucial for education, knowledge and research. Some term papers also “end” at the search engine result – introducing a new problem of digital “plagiarism” and cyber vigilantes.

As the search engine market only consists of proprietary, centralized systems, none of the underlying algorithms have been exactly published or are known by the greater mass of web-searchers. Nevertheless, some of the general principles are known to the web masters and therefore being used to push their sites to the top. As a consequence, most of the pages listed in modern search engines are optimized in a way that they do not generally reflect the real order in importance but rather in knowledge of the webmaster or site creator. Commercial and trade secret interests are so standing between the general knowledge and finding of the wanted sites.

An exception is a groundbreaking paper, in which students at Stanford University published their ranking and evaluation system called page rank. This system is generally based on the popularity of a site, which is defined by how much other web-pages link to it. Although the general principle and algorithm is known, this kind of web-popularity-based search also defines some problems.

So search engines algorithms define and monopolize the way how we access and perceive information. It must be questioned, if algorithms based on popularity are capable of providing the most valuable, creative, or innovative information. Since new and innovative material is not popular and therefore has not many supportive pages that link on it, it sinks to the ground in search engines and is not easily found. The students therefore concentrate on old, known material – innovation is stopped or greatly delayed, since evolution never emerges out of the centers of gravity.

Popularity based systems will always represent the mainstream only, search engine results will be distorted by contemporary trends, fashions, favors. It is no surprising that such a algorithm has been devised in a political system that chooses to elect an actor as its leader, the representation of the reality, while still struggling under its own weight.

As a result it would be desirable to have access to an global (and private) information retrieval system whose algorithms can be modified and customized according to the needs and requirements of the user. An example for such algorithms could be the "opt of the hill"-algorithm. There, some major players are identified with common methods and the pages which they link to get higher ratings.

Proposal: Decentralized, Open Source Search Engine

As seen above, personalized search and evaluation algorithms are not feasible on centralized information retrieval systems. The technological problems can not be accomplished with centralized hierarchical and commercial search engines. Too much computer load had to be established to grant every institution or different searcher the personalized information they need.

Our proposal is a decentralized, open source search engine, which is fit to support the above mentioned ideas and should be considered the next step in educational information retrieval.

As the system is open source, all algorithms are known to the public – and can be easily adjusted to new environment, new technology (like flash) and special interests. It can be customized and specialized, including search functionality for academic institutions and universities. If desired, every institution could build it's own repository including information from the main distributed system that is specialized on its own needs and desires.

A global distributed framework or search engine can successfully circumvent censorship efforts. The freedom of expression can be preserved and information can be exchanged even when regional authorities try to eliminate the freedom of speech. The framework can per definition not be censored – encryption, network administration, self organization, and open standards render manipulation efforts ineffective.

The integration of local data into the results of a search query can redefine the concept of information retrieval and publication. Although web based services have already lowered the entry level of global publishing services the very same barriers still continue to exist. The publication of information and the freedom of expression on the internet are tied to the benevolence of site owners, web space provider and server or hosting services. Commercial, religious and political interests often use these points of contact as a leverage to suppress critical information.

The decentralized system nevertheless has to include some kind of self-organizing hierarchy, based on computing power, internet-availability and bandwidth. Four general different functions can be established within the network – the crawlers, database-clients, group leaders and super nodes. The crawlers and database-clients are clustered to generate themed data bases – organized by the domain names of included web content and other attributes. Subgroups of at least three machines form a network consisting of single clusters which cover the entire web – providing enough redundancy to avoid data loss. While one machine is serving the search engine requests, a second computer updates its index or database, and the third machine interchanges data with neighboring or related clusters.

The subgroups are controlled by one (or more) group-leader-machines. They process all requests, query the database clients for search terms and keywords and cache the results for search terms for a given period of time (For details on data queries see 3.4). Group leaders are supervised by one (or more) super-nodes. Super nodes forward the user queries to group leaders and coordinate all network activity – assigning a new group function when necessary or replacing a machine which has been turned off.

Conclusion

We have analyzed the present search engine market and its global implications. We have provided a solution for a scalable, distributed information retrieval framework. The distributed open source search engine will offer a guarantee for better, more complete and more transparent results. Our framework will not only improve the actuality and reliability of search engine services, but also address several concerns relating to privacy, censorship and educational research.

We are currently deploying the first prototypes of our framework and continue to develop new features and modules. Our concept depends on the contributions of thousands of users, therefore we provide special incentive structures motivating users to participate and improve the global information architecture. As other distributed projects have shown, the net is ready to evolve from proprietary search systems and take the next step into the future of our information and education society.

Acknowledgements

Our thanks to all the authors and developers of published concepts and open source project who allowed us to develop this system. We would like to thank our mentors and partners Maiko Katayama, Stefanie Weber and Annika Held for their guidance, patience, and continuing support.