A Survey of Web Search Engines*

Juryon Paik', Junghyun Nam[e], Ung Mo Kim', and Dongho Won**,'

' Department of Computer Engineering, Sungkyunkwan University,
300 Cheoncheon-dong, Jangan-gu, Suwon-si, Gyeonggi-do 440-746, Korea
{wise96, umkim}@ece.skku.ac.kr, dhwon@security.re.kr [2]
Department of Computer Engineering, Konkuk University,
322 Danwol-dong, Chungju-si, Chungcheongbuk-do 380-701,
Korea jhnam@kku.ac.kr

Abstract. Search engines such as Google or Yahoo allow us to navigate the web; they crawl web pages periodically and automatically, store them into indexed databases, and retrieve search results via queries. However, the rapid growth of the web data brings a limit of indexing pages, which mass-produces data areas that cannot be accessed by the search engines. The truly required searching way is to provide valuable and accurate search results to users in a customized way and to deliver the information from the viewpoint of a user, not from the viewpoint of a search engine provider. One of the best thing for it is to consult with a higher number of search engines as possible, since no search engines reach all the information available on the entire web. In this paper, we survey major types of web search engines and highlight the main requirements for the design of the engine and research agendas that need to be addressed.

Keywords: Web search engine, surface web, deep web, federated search engine, meta search engine, aggregated search engine.

1 Introduction

The web allows people to share the information available on the websites through generically known web search engines. Many information seekers spend a lot of their time to mine materials from the vast and unorganized web. According to the site WorldWideWebsize.com, which is based on the most referred researches, the indexed Web is estimated to contain at least 7.82 billion pages. (http://www.worldwidewebsize.com/) The actual size of the Web, however, is supposed to be much larger, somewhere between 15 and 40 billion and probably closer to the latter than the former due to the presence of dynamically generated pages. In 2005 Yahoo announced that its search engine index contained more than 19.2 billion documents [9]. Recently in July 2008 the Google blog reported that the company's systems that processed links

* This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0005861).

** The corresponding author (e-mail: dhwon@security.re.kr).

on the web to find new content hit a milestone: 1 trillion unique URLs on the web at once [7]. It is a gigantic number of pages even Google cannot index every one of them. In fact, Google and Yahoo considered the best search engines today do not cover all the Web. They only access a small portion of the entire Web.

The Web can be actually divided into two regions, the *Surface Web* and the *Deep Web.* The former is a portion of the information that can be reached by conventional search engines such as Google, Yahoo, etc. According to [4,6], the Surface Web already composed of more than 70 billion pages and duplicates its volume every year. The latter mainly refers a portion of the vast repository of information that cannot be accessed directly by the conventional search engines. Its characteristics are similar to those of databases. Only a human reader can see the information by directly visiting its web sites and making direct database requests.

The problem is that Google and Yahoo only give you access to less than 20% of the entire web, even which is in part of the Surface Web. That remaining 80% is in the Deep Web and the technology of the conventional search engines does not provide an open-sesame to the massive Deep Web. According to the study conducted by Michael K. Bergman [2], public information on the Deep Web has been thought to be 400 to 550 times larger than the Surface Web. Now, it can be much more.

The Deep Web has a lot more information out there than we could ever imagine and most of them are very tempting resources. The resources on the Deep Web are generally of better quality and more relevant than those on the Surface Web. Besides, the quality of the Deep Web is thought to be 3 times better than that of the Surface Web. This is basically because the content of the Deep Web sites is created, written or validated by professionals, specialists and authorities in their particular area of expertise. Finding and searching inside those databases is only way of accessing the Deep Web, which is still in the mature stage and needs to be further researched.

2 Types of Search Engines

Different search engines return different results due to the variation in indexing and search process; some of search engines use crawlers and robots to copy the entire interne onto their own hard disk. Other search engines use real people to look at web pages and then decide whether to include them in search results. The differential is mainly caused by searched content whether is in the surface web or the deep web. Each type of search engines has its own strengths that should be emphasized. What they all share in common are the duty of fetching the vast content of information and store it in an efficiently searchable, manipulatable way and the aim of delivering relevant, helpful, and timely search results for their users. In this section, we will make a brief survey over the literatures for search engines, trying to summarize their types and characteristics.

2.1 General Search Engine

General search engines are search engines that cover all areas of interest and attempt to index large portion of the surface web using crawlers. When a user sends his query to the search engine, the engine looks for relevant keywords and retrieves the best matching web pages from its indexed databases. This type of search engines is typical and popular in current search engines. Some of the most famous general search engines are Google, Yahoo and Ask Jeeves, which have been hot and newsworthy for many years. General search engines are commonly used when a user has a well-defined topic, user's topic is obscure, or a user wants to retrieve a large number of web sites on the topic.

However, their engines' indexes do not seem to be built as fast as the growth of the surface web and have a hard time to index significant portion of the web. According to the white paper [5] Dogpile.com conducted two times overlap researches in April 2005 and 2007. The latest study was that the top 4 search engines, Google, Yahoo!, Windows Live^{TM} and Ask^{TM} were evaluated and 19,332 user-entered queries were measured. The results from this study highlight the fact there are wide differences between the four most popular search engines [4]. To get more useful and reliable information, users must therefore take into account the indexes of several general search engines not just one.

2.2 Vertical **Search Engine**

Vertical search engines focus on a specific segment of interest. The vertical content area may be based on topicality, media type, or genre of content. It may help to think that vertical search is as a search for a particular interest. Targeting one specific niche, a vertical search engine directly uses a focused crawler that attempts to index only web pages that are relevant to a pre-defined topic. Some examples are: HealthLine, for Health information only (www.healthline.com); Citeseer, for academic and scientific papers (citeseer.ist.psu.edu); Codase, for source codes only (www.codase. corn). Vertical search engines are used when a user's topic is focused on a specific area or a user is having difficulty locating what he/she wants on general. Therefore, the topic of vertical search is closely related to that of the Deep Web [6].

The retrieved results are differentiated from those of general search engines. They profit to be used for searches with particular aims, where much more accurate and higher quality of results is produced for users. It causes relatively low user-traffic, but the concern of visiting user is pretty high. These engines have lost prominence in the last decade because of the increased dominance of companies like Google, Yahoo and Microsoft. However, they are highly regarded again in recent years due to their own benefits.

2.3 Multi-Search Engines

The recent trend of web search is to get accurate and specialized information limited to a specific area or interest rather than general and massive information covering all areas. This is closely relevant to the fact that the web has too much data but does not have the data that help users make good decisions. If a single general or single vertical is used to find information, users pass up on a considerable amount of results from the current web. In order to gain as complete coverage as possible of the web, at least two search engines should be employed, even which will produce coverage of around 10%. The more search engines are used, the higher quality information is produced.

Another problem is spams. The search results often contain so called spam pages on top of the lists, which are usually unwanted and deceptive pages like adwares or spywares. This mainly occurs when users search the web via general search engines. Because vertical engines only index the information that fit their specific area of interest, their indexing algorithm is more precise, and the content involves professional revising, their results have superior quality and less susceptive to spam. Therefore, search engine developers are continually trying to improve the algorithms of general search engines, where one of methods is to adopt the benefits of vertical engines into the generals.

Federated Search Engine. Federate search systems provide a single-user interface to multiple search engines. The person using the federated search system probably knows that the query is sent to multiple sources and searched simultaneously but does not have to select which databases to search or worry about the process of how queries are submitted or results obtained. The mechanism of federated search is more complex than that of general search, which is commonly viewed as consisting of five phases described in **[1,11]:** resource discovery, wrapper induction, resource representation, resource selection, and resource merging.

Federated search interfaces normally list specialized databases - the deep web contents - by subject or category or in alphabetical order. Compared with viewing Google results, viewing of federated search results is more complicated. One challenging is to determine which vertical engines users should pick after initial searching if they have no knowledge of the resources. Another challenging is the speed of federated search engines [3]. Google normally takes a very short period time to complete a search. There is probably no way that the federated search can compete with Google in speed, because the speed of a federated search is dependent not only on the speed of its own server, but also on the speed of servers of specialized vertical search engines with various response times.

Meta Search Engine. A metasearch engine is a search tool that sends user queries to several different search engines and/or databases, and aggregates the results into a single list or displays them according to their sources, usually with the duplicates removed. Meta engines operate on the premise that the Web is too large for just a single search engine to index it all and that more comprehensive search results can be obtained by combining the results from heterogeneous search engines. This helps the user avoid to use multiple search engines separately. Currently Dogpile, owned by Infospace, is probably the most popular meta search engine on the market, but like all other meta search engines, it has limited market share.

Metasearch engines were popular 10-15 years ago. Now, however, the influence on the web space seems to be rather weak due to the current major search engines, Google ,Yahoo!, and so on. One of the larger problems with meta search in general is that most meta search engines tend to mix pay per click ads in their organic search results, and for some commercial queries 70% or more of the search results may be paid results.

Aggregated Search Engine. As described in [8,10], an Aggregated search system also addresses the task of searching and assembling information from different sources on the web and placing it in a single interface. However, it has now been implemented by major search engines, esp. Google. The information sources are powered by dedicated vertical search engines, all mostly within the remit of the general search engines, and not several and independent search engines. The big difference compared to the "standard" general search engines, is that the individual information sources in aggregated search retrieve from very different collection of documents, such as images, news, videos.

The heterogeneous information items cannot be ranked using the same algorithms because they have different features. The main challenge is how to identify and integrate relevant heterogeneous results for each given query into a single result page. However, the result page is a limited space that the verticals share. In many cases, a vertical engine cannot have more than a few results in an aggregated result page, even if the vertical is closely relevant to a given query. Consequently, users can suffer from significant distortion on the aggregated page caused by the cut-off rank.

Acknowledgments. This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0005861).

References

1. Avrahami, T. T., Yau, L., Callan, J.: The FedLemur Project: Federated search in the real world. J. of the American Society for Information Science and Technology, vol. 57, pp. 347-358, 2006.
2. Bergman, M. K.: The deep web: Surfacing hidden value. The J. of Electronic Publishing, vol. 7, no. 1, 2001.
3. Chen, L.: Metalib, WebFeat, and Google. Online Information Review, vol. 30, pp. 413-427, 2006.
4. Digimind: Discover and exploit the invisible web for competitive intelligence. White paper, January 2006.
5. Dogpile: Different engines, different results. White paper, April 2007.
6. Goshme: A new way of finding information in the internet. White paper, May 2006.
7. Jesse, A., Nissan, H.: We knew the web was big. Google Blog, July 2008.
8. Lalmas, M.: Aggregated search. In: Melucci, M., Baeza-Yates, R. (eds.) Advanced Topics in Information Retrieval, The Information Retrieval Series, vol. 33, pp. 109-264, 2011.
9. Markoff, J.: Debating the size of the web. Available on www.nytimes.com, August 2005.
10. Seo, J., Croft, W.B., Kim, K.H., Lee, J.H.: Smoothing click counts for aggregated vertical search. In: Clough, P. et al. (eds.) ERIC 2011. LNCS vol. 6611, pp. 387-398, 2011.
11. Si, L.: Federated search of text search engines in uncooperative environments. ACM SIGIR Forum, vol. 41, no. 1, p. 120, 2007.