Survey of Web Search Engines: Classifications, Characteristics and Effectiveness

Juryon Paik, Junghyun Nam, Junghwan Kim, Ung Mo Kim, Dongho Won

Abstract

The explosive growth of data available in the internet exposes a serious problem, information-overflow, where each user gets rarely necessary information and which brings blind spot of information search. The blind spot means the areas which cannot be accessed by search engines. Hence, there is no way users can get the information in blind spots. They are getting wider, which cause loss of valuable information for users' queries. The problem of blind spots stems from the way of navigating the web for current leading search engines, Google or Yahoo; 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. Besides, they still retrieve useless results for depending on a few keywords, where users wander again for really necessary information. 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. Recently, fresh search engines are developed and issued with Silicon Valley as the center. Their objectives are the intelligent and specialized search results as well as easy user interfaces. In this manuscript, we introduce some representatives of the newly published search engines along with surveying and classifying systematically current existing web search engines.

Keywords: Search Engines, Types of Search Engines, Search Engines Characteristics

1. Introduction

The growing trend of web users and developers is to discover or provide the specialized and professional information limited on some specified interest or topics. This is not irrelevant to the massive data in the web. According to the site WorldWideWebSize.com, which is based on the most referred researches, the indexed web is estimated to contain at least 14.33 billion pages (visited on February 19, 2013). 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.

In 2005 Yahoo announced that its search engine index contained more than 19.2 billion documents [1]. The Google blog reported in July 2008 that the company's systems that processed links on the web to find new content hit a milestone: 1 trillion unique URLs on the web at once [2]. It is a gigantic number of pages even Google cannot index every one of them.

Such the web data causes two opposite problems: excessive amount of unreliable data & insufficient amount of credit information. This phenomenon indicates that the quantity of accessible data has grown about terabytes while each individual user just accesses a tiny portion of the web along with seizing a small quantity of necessary information. Even worse the data overflow threats the merits of using the web because of the following reasons: 1) it is not easy task to obtain the exact information in
time users want to search. 2) Many users are bothered by spam pages. 3) Valuable and helpful data is under the sea and is getting bigger without any access.

In fact, Google and Yahoo considered the best search engines today cannot cover the entire web. They only access a small portion of the web. Exactly speaking, these best engines only give you access to less than 20% of the entire web. This is because that crawlers of the major search engines cannot reach to the remaining more than 80% of the web, and the data in the portion is usually specialized and dedicated to a specific activity, technology or profession. The characteristics of that portion exactly corresponds the current trend of using or developing the web. However, most data in 80% of the web is hidden in order to avoid crawling. Users have no chance to get valuable and necessary information for their queries. Even worse, the invisible data is getting bigger and the areas the crawlers cannot reach are getting wider.

Then, can’t we navigate the hidden web? The answer is YES. There are some specialized search engines can be the open-sesame for the hidden web. However, these engines are not panaceas, because they focus only on certain resources not the whole. There are no search engines reach all the information available in the entire web. The best way to catch and obtain valuable data from the web is to consult with a higher number of search engines as possible. To do, it is necessary to know what kinds of search engines are there and which characteristics qualify for each type.

There are more than 400,000 web search engines, and they can be grouped by their main characteristics; way of indexing, searching process, searching area, and so on. In this paper, we classify them according by a number of search engines to process a given query, and summarize the classes with some notable examples. Especially focusing on vertical search engines (which will be mainly described in a moment) we state related works in accordance with our research analysis.

2. Preliminaries

2.1. Areas of the Web

Actually, the web is 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 [3, 4], 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; information in databases is generally inaccessible to the software spiders and crawlers that create search engine indexes. Only a human reader can see the information by directly visiting its web sites and making direct database requests [5, 6].

According to the study conducted by Michael K. Bergman [7], public information on the deep web has been thought to be 400 to 550 times larger than that on the surface web. Considered the best search engine today, even Google only indexes a fraction of this monstrous content, which is an unbelievable story but it is true.

Of course, some users just want to stick with what they can discover through Google or Yahoo. However, it is not always easy to get what they are looking for, especially if they want something a bit complicated or obscure. You might wander the web. This is because the conventional search engines index most web pages in the surface web and this web just accounts for less than 20% of the entire web. Without being helped by the deep web, conventional search engines will not necessarily help searchers, because 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 it are generally better quality and more relevant than those of 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 [3].

Since the web sites in the deep web are specialized and dedicated to a specific interest of activities or technologies, they usually consist of documents rather than web pages. This is the reason the deep web is thought to be a vast collection of specialized databases. Finding and searching inside those databases are only way of accessing the deep web, and major search engines developers recognize its importance.
3. Search Engine Categories

Google is not the end all and be all of search engines. There are actually some better and more efficient search engines out there, and there are different types of search engines worthy of information seeker’s attention. Different search engines return different search results due to the variation in indexing and search process. The most influencing factor in the searched content is a portion of the web; Surface or Deep. In this paper, we summarize and categorize existing search engines with two aspects: The first criterion is the two main portions of the web, which is a publicly known categorization and the categories have introduced partially in some articles and white papers. The second criterion is importantly set in this paper, which categorizes search engines with respect to a number of the engines used to answer users’ queries.

3.1. Search Engine Categories in accordance with Web Portions

Search engines can be broadly classified into two types as follows depending on the web portion which is used in order to retrieve results for given queries [5, 8].

(1) General search engines: They 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 Bing, which have been hot and newsworthy for many years. 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 [7] Dogpile.com conducted two times overlap researches in April 2005 and 2007. The latest study from Queensland University of Technology and the Pennsylvania State University was that the top 4 search engines, Google, Yahoo!, Windows Live™ (formerly MSN search) and Ask™ (formerly Ask Jeeves) 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. The overlap across the first page of search results was found to be only 0.6% on average for a given query but 88.3% of the results were unique. To get more useful and reliable information, users must therefore take into account the indexes of several general search engines not just one [3].

(2) Vertical search engines: They 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.com). 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; spam free, superior quality, access to more of the surface web, and access to the deep web.

3.2. Search Engine Categories in accordance with Number of Running Search Engines

Currently, it is general to classify web search engines with respect to their crawling areas. However, the explosive growth of web data hinders accurate and useful information retrieval with just a single search engine from the web. 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% [9]. The more search engines are used, the higher quality information is produced.

Another obstacle is spams. The search results often contain spam pages on top of the lists, which are usually unwanted and deceptive pages. 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 susceptible 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.

The subsection newly classifies a variety of search engines with regard to their number of running engines to retrieve results for a given query, as well as gives proper examples for each classification.

3.2.1. Single Search Engine

It is the simplest way to search the web with just a single search engine. It can be either using a vertical search engine for a specific area to get superior results or using a general search engine for common topics to get usual results. The old version of Google was a proper example of a single general, but it is not any more. The reason will be given in later.

However, the type of a single vertical is continually retained. One of proper examples is SearchMedica (www.searchmedica.com), which is a notable vertical engine in medical niche. Figure 1 shows a retrieval page for a query ‘fibroid’. It gives sorted lists in increasing order, where the top is the most hit and recent page, with some categories upper the lists and query expansion in left menu. The categories allow users to refine results by content category such as Research/Reviews, Practical Articles and News, Patient Education Materials, and Practice Guidelines.

SearchMedica's index contains over 1,000 websites organized into numerous therapeutic categories, which appear as links across the top of the screen. Searches might also be refined by clicking one of the therapeutic category links to narrow results to that area.
3.2.2. Multi-Search Engines

The current 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. In order to gain as complete coverage as possible of the web, at least two search engines should be employed. The more search engines are used, the higher quality information is produced.

Search engine developers are continually trying to improve algorithms of existing search engines in various ways. Several methods have been issued; adopt the benefits of vertical search engines into the generals, play as an intermediate to find proper search engines for queries, use as many as possible search engines simultaneously together, and so on.

Multi-general search engines are using different general search engines simultaneously in order to retrieve a result for a given query. According to the white paper [9] published by Dogpile.com, a research study conducted in April 2007 with researchers from Queensland University of Technology and the Pennsylvania State University measured the overlap and ranking differences of the leading web search engines. The top 4 search engines, Google, Yahoo!, Windows Live™ and Ask™ were evaluated and 19,332 user-entered queries were measured. The results from the study highlight the fact there are wide differences between the four most popular search engines. According to the research, the overlap across the first page of search results was found to be only 0.6% on average for a given query. It means that first page search results from them are largely unique and this trend will most likely continue as each engine keeps modifying their crawling and ranking technologies. For the better quality search results from the surface web, it is important to use multiple general engines.

‘Clusty’ search engine developed by Vivismo is one of well-known search engines for this group. Figure 2 shows a part of result pages for a query ‘Barack Obama’ retrieved from Clusty. It can be known how many search engines are run simultaneously and how many pages are retrieved from the individual engines by clicking the sources tab on the left side.

Each general search engine navigates and finds various web pages with varying contents related to the query, thus, they provide users with variety of information. However on the other hand, it causes such a problem that users should search again by themselves for really necessary information because there are huge uninterested data in returned pages. One more thing is spam. Users cannot distinguish which one is spam page without reading the page.

Federated search engines provide a single-user interface to multiple search engines, especially multiple vertical search engines. For the surface web contents, the solutions for general search engines are effective. However, they do not fit to the deep web contents, because the information can only be accessed via source-specific search engines. Federated Search, also known as distributed information retrieval, searches information that cannot be accessed by general search engines [10].

The person using a federated search system may know that his query is sent to multiple sources and searched simultaneously but does not have to choose 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 [11, 12]: resource discovery, wrapper induction, resource representation, resource selection, and resource merging.

Federated search interfaces normally list topical databases - the deep web contents - by subject or category or in alphabetical order. For users it seems to be more complicated than the classical interface given by Google. Screens on Figure 3 are the retrieval results after initial searches from MetaLib and WebFeat which are the two most popular federated search engines [13]. Both engines list their results by vertical search engines or databases first with a number of hits for each one (we cannot try Webfeat due to its commercial policy; therefore, we just refer its white paper).
Google familiar users are not accustomed to the provided interfaces and may not understand at a glance how to use them, because the lists are not web pages themselves. In case of MetaLib for a query ‘xml mining’ in Science/Engineering area, it first shows the list of vertical search engines or databases closely related to the query. In the meantime, each vertical engine or database searches proper results inside its resources and returns them to MetaLib. Afterwards, MetaLib lists the detailed results from the highest rank score search engines, where users have some difficulties for choosing certain vertical search engines if they have no prior information of such engines.

Even though users select proper vertical engines or specialized databases, it’s not the end. The picked search engine can retrieve inefficient results or have rarely related results, because vertical search engines depend on their searching algorithms.

Another problem for Google friendly persons is searching speed [13]. 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. It is a fundamental limitation of federal search engines.

Meta-search 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. It operates 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 a user avoid to using multiple search engines separately. The way is similar to that of the federated. However, meta-search engines enabling searches in deep web databases and vertical engines are fewer. In addition, most of them only search in their own selection of pre-configured search engines.

Figure 2. Search Result from Clusty for a Query ‘Barack Obama’
Figure 3. Displaying Results from the Prominent Federated Search Engines – MetaLib and WebFeat
Meta-search 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.

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. Figure 4 depicts the retrieved result from Dogpile. The same query ‘xml mining’ was given. Compared to Figure 3, the returned interface is similar to that of Google, but not that of federated engines. And actually its resources are Google, Yahoo, and Yandex which are general search engines usually for the surface web.

Figure 4. Returned Result from Dogpile for a Query ‘xml mining’

Aggregated search engines also address the task of searching and assembling information from different sources on the web and placing it in a single interface [14, 15]. However, it has now been implemented by major search engines, especially Google. For a query ‘xml mining’ Google returns different types of information presented on Figure 5. With text results, images and news results are listed together.

As stated in [14], 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. As pointed in [15], 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.

Each type of search engines has its own strengths that should be emphasized. General search engines are commonly used when a user has an obscure topic or wants to retrieve a large number of web sites on the topic. On the contrary, vertical search engines are when a user has a focused topic on a specific area or is having difficulty locating what he/she wants on general. Federated search engines are used when a user wants to know hidden and professional information in the deep web for a topic with multiple vertical search engines. Meta search engines are when a user is tend to see combined different lists for a same query from different search engines, mostly general search engines. And, aggregated search engines are used when a user wants to get information from different sources in a single interface. It is recommended that web searchers pick appropriate search engines for their purposes, because what they all share in common is the aim of delivering relevant, helpful, and timely search results for their users.

Currently there are more than 400,000 search engines available online and this number increases fast stated in [11]. It is already so gigantic that it turns into a problem. How can the web users know which search engines are the best to their search? What are the best choices to every single search? What about the really time spending and low effective searches through vertical engines? There still remain a lot of open problems for the web searching.

4. Conclusion

The volume of information in the web is so enormous and the speed of its increase is too rapid. However, the technical capacity of search engines’ indexes has not grown to access all web pages. With a single search engine it is absolutely slim to reach the specialized and professional information from the web, especially the deep web. Looking for best way to satisfy user’s need to be able to search the deep web, high numbers of web search engines have been developed, issued, or disappeared. Among them, some engines such as Google and Yahoo have become popular and gained significance.
This paper has organized existing web search engines according to their domains and number of simultaneous running search engines. Each group of search engines was explained systematically its pros and cons particularly from the deep web point of view. Even more, the certain vertical search engines were introduced and stated for their benefits, because they have been highly regarded again in recent years. Vertical search engines had actually lost prominence in the last decade because of the increased dominance of companies like Google, but they are rebounding in the current web environment.

There have been so many search engines in the web, and still developing. Whatever their purposes are diverse, the truly required searching way would be 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.

5. References