Site-Specific versus General Purpose Web Search Engines: A Comparative Evaluation

G. Atsaros, D. Spinellis, P. Louridas
Department of Management Science and Technology
Athens University of Economics and Business
gatsar@dmst.aueb.gr, dds@aueb.gr, louridas@aueb.gr

Abstract

We can distinguish two types of web search engines: general use ones that index and search all the web, and site-specific ones that are provided by individual websites for local searching. A comparison of the effectiveness of the two types allows search engine users to choose the right engine and organizations to decide whether they should develop their own search software or purchase the search function as a service. We evaluate the performance of two general purpose search engines and 10 site-specific ones. The criteria we used are precision and relative recall. We entered 20 queries in each website's search engine and evaluated the first 10 links. According to the results, Google is in most cases the most efficient search engine. However, in some cases general purpose search engines do not index the website's content as well as a site-specific engine.

1. Introduction

The continuous and fast development of the internet has made it an integral part of everyday life for people and organizations. According to statistic data from Internet World Stats [1], on November 27th 2007, internet users reached the number of approximately 1,262 million or 19.1% of the world population. But, as the internet is sharply increasing, the amount of data available via the web is increasing as well. That is why internet users use search engines in order to locate the data they want, without wasting much time and avoiding the risk to get lost into the immense amount of data available through the net. The search engines are information retrieval systems that match the queries of the users with relevant documents and links. About 85% of internet users utilize web search engines for their informational needs [2], while search engines usage is the second most popular web service, after e-mail [3]. The search engines developed for internet users are not only engines for general use that search all the web, but also site-specific search engines that are provided by individual websites for data searching in their databases.

Search engines evaluations became more and more popular as the web was expanding rapidly and its users started to get dependent on search engines for their information retrieval needs. Search engine evaluations serve both search engine constructors, as they provide a means to differentiate and rank their product, and users, as they provide a means to assess the quality available search engines.

Most of research papers about search engine evaluation compare general purpose search engines or meta-search engines. This paper attempts to provide an answer to the research question whether general purpose search engines are better than site-specific search engines. The answer to this question is going to enable organizations to appreciate whether they should develop their own search software or they should purchase the search function as a separate service. Moreover, the answer to this question can also help search engine users to find more easily the information they are interested in, choosing every time the right search engine. This paper carries out a comparative evaluation of the performance of two well known general purpose search engines, "Google" and "Yahoo!", and the site-specific search engines of 10 websites , "Amazon", "Scopus", "IMBD", "IEEE Xplore", "IngentaConnect", "Barnes & Noble", "PubMed", "ACM Portal", "SpringerLink", "JSTOR". The criteria in this evaluation are the well known and established criteria of precision and relative recall.

This paper starts with a literature review on the research that has been made in the field of search engines evaluation. After that, the methodology and the criteria of the evaluation process are presented, followed by the evaluation results. Finally, the results of the evaluation are shown, followed by a discussion

978-0-7695-3323-0/08 $25.00 © 2008 IEEE
DOI 10.1109/PCI.2008.47

about the conclusions drawn from the performance of the search engines.

2. Literature review

As the web expands, the research about the performance of web search engines acquires a lot of importance and attention. One problem associated with search engines evaluations is that search engines keep changing constantly, evolving their mechanisms and their communication platform with the users. In addition, quite often new search engines emerge or already existing search engines stop operating. This, together with the dynamical nature of the web, does not let the evaluations of search engines remain valid for a long time [4]. Initially, the criteria used for search engines evaluations came from the field of information retrieval systems, the idea being that web search engines, as information retrieval systems [5], could be evaluated with the same criteria and methods. Most information retrieval systems' evaluations follow a systemic approach, putting emphasis on the search algorithms of search engines using quantitative criteria. One of the most popular evaluations of information retrieval systems, which founded the basis for later research, was the Cranfield experiments [6]. These experiments established the well known criteria of precision and recall [4]. Although the use of these measures is controversial [4], the majority of published evaluations use these criteria. The Cranfield evaluation model [6] is widely acceptable because it includes quantitative evaluation measures and it relies on meticulous experimental conditions [15]. Thus, it places powerful scientific foundations for the information retrieval systems evaluation science [15]. Later, some other similar experiments followed, like SMART [7] and STAIRS [8], but the most important follow-up of the Cranfield experiments in the information retrieval field are the TREC (Text REtrieval Conference) conferences [9], which first took place in 1992, and whose purpose is to support the research in the information retrieval field, providing the essential infrastructure for big scale evaluations of information retrieval methods. The majority of the well-known search engines incorporate technology that was first developed in these conferences, and while the research papers about search engines evaluations have grown in numbers, the criteria for these evaluations have still come from the Cranfield experiments. Recently, however, the number of criteria used for search engines evaluations increased, including factors as the stability of the results retrieved, the web coverage, the capabilities of search engines, the interfaces of search engines with the user, the duplicate links retrieved, the percentage of inactive links retrieved, the bias of search engines, the index mechanism of search engines, the level of difficulty of queries inserted, the ranking of results retrieved, the quality of results retrieved and many more. Finally, it is particularly important that much recent research deviates from the traditional focus on the technical and systemic characteristics of search engines and evaluate them from a user's perspective. These types of evaluations examine facts such as the user's satisfaction, their perceptions and their preferences. However, in this study the criteria of precision and relative recall are selected, mainly because they are precise, easy to understand and convenient for accurate comparisons [15].

3. Methodology

3.1 Selection of Search Engines

General purpose search engines use crawler mechanisms in order to scan as many web documents as possible. Usually, they focus on the general population of internet search engine users and they cover all types of information needs [10]. According to statistics data from Nielsen//NetRatings [11], the two most popular general purpose search engines are Google and Yahoo!, which were selected for this work.

Concerning the site-specific search engines, the sites were selected according to their popularity using the measurements from Alexa Research [12]. The chosen websites varied from e-commerce websites to academic databases.

3.2 Selection of Queries

The queries used in search engines evaluation were collected from the database of each website in order to insure the fact that the content of each query existed in the website's database. The type of queries used in the evaluation was informational as their intent was to acquire information which was supposed to be present on certain web pages [16]. Thus, 20 queries were inserted in each website's search engine and then in "Google" and "Yahoo!". The queries used for the evaluation were selected randomly, in a stochastic way in order to efface any kind of possible bias. For the academic databases, the research papers were picked first from a random journal, using the following procedure. Initially, a number from 1 to 26 was selected, where each number represented a different letter of the alphabet in order to choose the first letter of the journal. Once a specific letter was selected, a random journal was picked from the journals available

for the selected letter. If there were not any journals referring to the selected letter, the process was repeated. Then, having picked a journal, a random year was chosen (for the years that the specific journal has been published), a random issue (between 1-N, N the number of issues of a specific journal for a specific year) and a random article (between 1-M, M the number of articles included in the specific issue already selected). For example, for first letter "C", journal "Computer", year 2007, issue 5, article 4.

For IMDB, the selections were made for a random movie genre (between 1-N, where N the number of genres of movies) and then a random movie was picked (between 1-50, the top 50 of each genre were taken into consideration). For Amazon and Barnes & Noble, the books were picked for a random category (between 1-N, N the number of book categories), then a random sub-category (between 1-M, M the number of sub-categories of a specific category) and then a random book X (between 1-N, the top 20 books of each sub-category were taken into consideration).

3.3 Evaluation Criteria and Assumptions

The criteria used for the evaluation of selected search engines are precision and relative recall. These criteria have been used in a large number of research papers and have proved to be fair and reliable. Although there are a lot of criteria tried and proposed by well known research papers for search engines evaluation, precision and recall remain the most convenient and efficient ones. Regarding the recall as a criterion, as it is very difficult to calculate the "absolute recall" considering the huge size of the web, in this evaluation the relative recall was used, which was introduced by Clarke and Willet [13]. From the retrieving documents of the search engines, the top 10 were taken into consideration, based on the fact that search engine users have the tendency to follow the first results retrieved by search engines [14]. Thus, the precision and relative recall scores were calculated from the quotients below:

- Relative recall = Total number of results retrieved by a search engine / Sum of results retrieved by all search engines

- Precision = Number of relevant results retrieved by a search engine / 10

For the searches carried out with the general purpose search engines, the jargon "site:" (e.g. site: jstor.org) was used in order to guide the search process.

4. Results and discussion

The precision and relative recall scores of each search engine result were calculated as the average score of the precision and relative recall scores achieved by each search engine for the 20 searches carried out. When duplicate links between the results of search engines were found, they are counted into the sum as one. The relevancy of the links retrieved was calculated in a binary fashion, marking them as relevant or non relevant. A link was defined as relevant if the provided information by its title, its summary and its content contained all the words of the query and was significantly related to it. Additionally, the links were considered as relevant if they contained most of the words of the query and at the same time they could be considered useful for the search engine user. The evaluator who judged the relevancy of each link was the first author.

The evaluation results are presented below in Figures 1 & 2. Figure 1 depicts the precision scores, while Figure 2 depicts the corresponding relative recall scores.

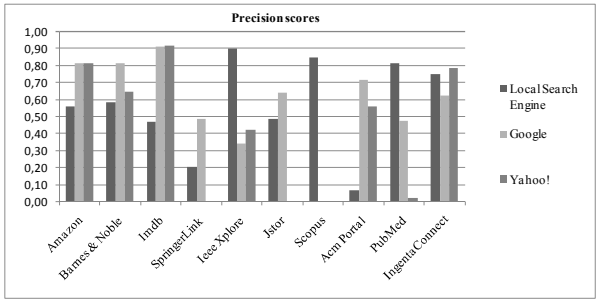

Fig.1 Precision scores of site-specific search engines, Google and Yahoo!

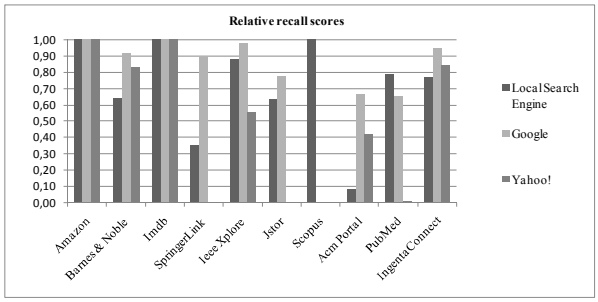

Fig.2 Relative recall scores of site-specific search engines, Google and Yahoo!

The low precision of Amazon can be explained by the fact that all the information about each book is well organized and filed under the link of each book and the search engine retrieved in most cases only book titles

and not interspersed data about them such as "customer reviews", "special offers" or "related forums" like the other two search engines did. So in the top 10 results of Amazon there were a lot of books-links that had no relation with the searched query. The general purpose search engines retrieved links which were all related to the searched queries. But the relative recall (score 1 for all) proved that all of them retrieved the same amount of information.

For the Scopus search engine, the results show it as being very efficient and competent. The Google and Yahoo! search engines could not index the content of its database, so it was not possible to be comparatively evaluated.

It was quite remarkable that in most of the searches which were carried out, IEEE Xplore search engine retrieved the exact query and nothing else. The relative recall score of IEEE Xplore was equally high, which indicates a very good search engine. It was also noticed that Google had the best score of relative recall because it located some articles that the IEEE Xplore search engine was unable to track. Yahoo! achieved the lowest score of recall.

This low precision of IMBD can be explained by the fact that all the information about each movie is well organized and filed under the title of the movie and the search engine retrieved only titles of movies and not interspersed data about them. The situation is similar to that observed with Amazon. This is supported by the fact that the relative recall was the same for all the search engines (score 1 for all).

Barnes & Noble has the lowest precision and the lowest relative recall compared to the general purpose search engines. Google appears the most efficient search engine. But, these low results of Barnes & Noble search engine can be mainly explained by the fact that this search engine does not give the user the opportunity to search all the possible categories, but only in specific categories such as "books" or "used books". As a result, it was inevitable to choose a specific search domain.

The search engine of IngentaConnect cannot overcome general purpose search engines' performance.

The JSTOR website proved to be better indexed by Google. Yahoo! does not index at all the content of this website.

The search engine of PubMed proved very efficient, effective and better than the general purpose search engines that were used. Most of PubMed content could not be indexed by Yahoo!.

On the contrary, the search engine of ACM achieved a quite poor performance. It is quite remarkable that the search engine of ACM failed to retrieve 11 articles of its own database. The relative recall score of ACM search engine was equally low, whereas Google achieved the highest score.

Finally, the content of SpringerLink seems to be indexed better by Google. Yahoo! does not index at all the content of this website.

5. Conclusions and future work

As it can be shown from the results of the evaluation, general purpose search engines appear to be more efficient and effective than site-specific search engines, with some exceptions where either the websites own very competent search engines (e.g. IEEE Xplore, PubMed), or the general purpose search engines cannot index some web pages, which, according to them, are part of the "deep web". Google has the best performance overall, justifying in this way its good reputation. Although, Yahoo! may be somewhat worse than Google, it is still better than site-specific search engines when it indexes their content. Google and Yahoo! stand to benefit over site-specific search engines by using the "Sitemaps protocol", which supplements their crawl-based mechanisms by providing them with information about URLs to retrieve data on websites that are available for crawling [17]. They benefit by having a global view of the popularity of each page through links pointing to it from the whole world, and also by being able to track the users' interests across many sites. This protocol help web crawlers search more efficiently a website and enable web searches in the "deep web". It can then be arguably concluded that it is in the interest of online organizations to purchase the search operation as a service from providers who are site-specific in the development of such services, than developing the appropriate search software on their own. This way, the organizations have the opportunity to take advantage of the efficient and innovative search mechanisms and generally the know-how of experienced and successful providers like Google and Yahoo!, along with the data that such companies gather about users' preferences. Furthermore, as far as search engines' users are concerned, they can use general purpose search engines not only as an alternative choice when site-specific search engines do not satisfy them, but even better as primary tools.

It should be noted, though, that the results of this research are indicative only for the period in which this evaluation was carried out, because the progress in the field of search engines is very rapid and if this evaluation takes place sometime in the near future, it may show different results. Also, since the decision of whether a retrieved link is relevant or not with the searched query is "in the eye of the beholder" and the

fact that only one evaluator judged the relevancy of the queries, in the future the methods and criteria used in similar studies can be more objective if more than one evaluators participate and a different scale of relevance is used. Finally, since no similar evaluation of search engines has been made so far, the future prospects in this type of evaluations with the same or other criteria are highly favorable. For instance, it would be interesting a comparative evaluation between site-specific search engines of academic databases such as IEEE Xplore and ACM with Google Scholar.

6. Acknowledgment

This work is partially funded by the European Community's Sixth Framework Programme under the contract IST-2005-033331 "Software Quality Observatory for Open Source Software" (SQO-OSS).

7. References

[1] Internet Statistics, Online, Available: http://www.internetworldstats.com/emarketing.htm.

[2] Kobayashi M., and Takeda K, "Information retrieval on the Web*", ACM Computing Surveys*, 32(2), 2000, pp. 144–173.

[3] L. Rainie, "Search engine use shoots up in the past year and edges towards email as the primary internet application", *Pew Internet & American Life Project*, 2005, Washington, DC.

[4] Oppenheim C., Morris A., McKnight C. and Lowley S., "The evaluation of WWW search engines", *Journal of Documentation*, vol. 56, no. 2, 2000, pp. 190-211.

[5] Salton G., and McGill M., "Introduction to modern information retrieval". New York: McGraw-Hill, 1983.

[6] Cleverdon C. W., Mills J. and Keen E. M., "Factors determining the performance of indexing systems", *Aslib Cranfield Research Project*, Cranfield UK, College of Aeronautics. (Volume 1: Design; Volume 2: Results), 1966.

[7] Salton G., "The SMART retrieval systems: experiments in automatic document processing", Englewood Cliffs, NJ: Prentice-Hall, 1971.

[8] Blair D. C., and Maron M. E., "Full-text information retrieval: further analysis and clarification", *Information Processing & Management*, 26(2), 1990, pp. 437–447.

[9] Text Retrieval Conference, Online, Available: http://trec.nist.gov/overview.html (Accessed on 5 January 2008).

[10] B. J. Jansen and P. Molina, "The effectiveness of Web search engines for retrieving relevant ecommerce links", *Information Processing & Management*, vol. 42, 2006, pp. 1075-1098.

[11] Nielsen Online Announces December U.S. Search Share Rankings, Online, Available: http://www.nielsen-netratings.com/pr/pr_080118.pdf.

[12] Alexa, the Web Information Company. Online. Available: http://www.alexa.com/browse?CategoryID=1.

[13] Clarke S.J. and Willett P., "Estimating the recall performance of Web search engines", *Aslib Proceedings*, 49(7), 1997, pp. 184–189.

[14] Jakob Nielsen, "The Power of Defaults", 2005, Online, Available: http://www.useit.com/alertbox/defaults.html.

[15] Hildreth C. R., "Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study", *Information Research*, 6(2), 2001, Available: http://InformationR.net/ir/6-2/paper101.html.

[16] Andrei Broder, "A taxonomy of web search", *ACM SIGIR Forum*, 36(2), 2002, Available: http://www.acm.org/sigir/forum/F2002/broder.pdf.

[17] Sitemaps.org, Online, Available: http://www.sitemaps.org/.