A Survey on the "Performance Evaluation of Various Meta Search Engines"

K.Srinivas[1], P.V.S.Srinivas[2] and A.Govardhan[3]

[1] Associate Professor, Department of IT, Geethanjali College of Engineering and Technology, Cheeryal (V), Keesara (M), R.R.Dist, A.P- 501 301, India.

[2] Professor, Department of CSE, Geethanjali College of Engineering and Technology, Cheeryal (V), Keesara (M), R.R.Dist, A.P- 501 301, India.

[3] Professor of CSE and Principal, JNTU College of Engineering and Technology, Jagityal, Karimnagar (D),A.P, India.

Abstract

Though a Search Engine (SE) helps in the process of retrieving the information required to the user, a Meta Search Engine (MSEs) on the other hand uses new methodologies or fusion schemes for the information retrieval from the Web, and helps the user to collect more, relevant documents from the Web. This paper proposes a survey on various Meta Search Engines and the various parameters on which the efficiency of a MSE lies.

Keywords: *Web, Search Engine (SE), Meta Search Engine (MSE) and Information retrieval.*

1. Introduction

Search engines play a pivotal role in the process of retrieving information from the Web. When the user gives a Query, as a response, a Search engine returns a list of relevant results ranked in order. As a human, it is the tendency of the user to use top-down approach of the list displayed by the Search Engine and examines one result at a time, until the required information is found. However, while search engines are definitely good for certain search tasks like finding the home page of an organization. They may be less effective for satisfying broad or ambiguous queries. The results on different subtopics or meanings of a query will be mixed together in the list, thus implying that the user may have to sift through a large number of irrelevant items to locate those of interest. On the other hand, there is no way to exactly figure out what is relevant to the user given that the queries are usually very short and their interpretation is inherently ambiguous in the absence of a context.

An Effective and alternate approach to the information retrieval on the Web in recent years is by using the Meta Search Engine (MSE), instead of simply a Search Engine[5].This paper proposes a comprehensive survey on various Meta Search Engines and their performance in the process of information retrieval from the Web.

Section 2 explains about various types of Search Engines which are under use, Section 3 discusses the challenges posed by Search Engines, Sections 4, 5 and 6 briefly explains about different types of Meta search Engines, their properties and architecture. Finally Section 7 gives the conclusions.

2. Different types of Search Engines

Ask is a Search Engine, which is also known as Ask Jeeves. It is basically designed to answer the user's queries in the mode of Q&A and is proved to be a focused search engine. Ask was developed in 1996 by Garrett Gruener and David Warthen in Berkeley, California. Originally, the software was developed and implemented by Gray Chevsky [13]. Easier AskJeevs.com was built on core engine by Warthen, Chevsky, and Justin Grant.Three venture capital firms, Highland Capital Partners, Institutional Venture Partners, and the RODA Group were early investors of Ask.com, and it is currently owned by Inter Active Corp under the NASDAQ symbol IACI. In late 2010, facing insurmountable competition from Google, the company outsourced its web search technology to an unspecified third party and returned to its roots as a question and answer site [10].

Bing is a Search Engine, which was formerly known as Live Search, Windows Live Search, and MSN

Search. It is a web search engine (advertised as a "decision engine") that was owned by Microsoft [7]. Bing was unveiled by Microsoft CEO Steve Ballmer on May 28, 2009 at the *All* Things Digital conference in San Diego. It went fully online on June 3, 2009, with a preview version released on June 1, 2009. Notable changes include the listing of search suggestions as queries are entered and a list of related searches (called "Explorer pane") based on semantic technology from Powerset that Microsoft purchased in 2008. On July 29, 2009, Microsoft and Yahoo! announced a deal in which Bing would power Yahoo! Search. All Yahoo! Search global customers and partners are expected to be transitioned by early 2012.

Google Search or Google Web Search is a web search engine owned by Google Inc. and is the most-used search engine on the Web. Google receives several hundred million queries each day through its various services. The main purpose of Google Search is to hunt for text in webpages, as opposed to other data, such as with Google Image Search. Google search was originally developed by Larry Page and Sergey Brin in 1997. Google Search provides at least 22 special features beyond the original word-search capability. These include synonyms, weather forecasts, time zones, stock quotes, maps, earthquake data, movie show times, airports, home listings, and sports scores[9]. There are special features for numbers, including ranges, prices, temperatures, money/unit conversions, calculations, package tracking, patents, area codes, and language translation of displayed pages. The order of search results (ghits for *Google hits*) on Google's search-results pages is based, in part, on a priority rank called a "PageRank". Google Search provides many options for customized search, using Boolean operators such as: exclusion ("-xx"), inclusion ("+xx"), alternatives ("xx OR yy"), and wildcard ("x * x").

Yahoo! Search is a web search engine, owned by Yahoo! Inc. till December 2009, the 2nd largest search engine on the web by query volume, at 6.42%, after its competitor Google at 85.35% and before Baidu at 3.67%, according to Net Applications. Originally, Yahoo! Search started as a web directory of other websites, organized in a hierarchy, as opposed to a searchable index of pages. In the late 1990s, Yahoo! evolved into a full-fledged portal with a search interface and, by 2007, a limited version of selection-based search. Yahoo! Search**,** originally referred to as *Yahoo!* provided Search interface, would send queries to a searchable index of pages supplemented with its directory of sites. The results were presented to the user under the Yahoo! brand. Originally, none of the actual web crawling and storage/retrieval of data was done by Yahoo! itself. In 2001 the searchable index was powered by Inktomi and later was powered by Google until 2004, when Yahoo! Search became independent. On July 29, 2009, Microsoft and Yahoo! announced a deal in which Bing would power Yahoo! Search. All Yahoo! Search global customers and partners are expected to be transitioned by early 2012[13].

3. Challenges Posed by Search Engines (SEs)

Using a Search Engine (SE), an index is searched rather than the entire Web. An index is created and maintained by automated web searching by programs commonly known as spiders. Plain search engines prove to be very effective for certain types of search tasks, such as retrieving of a particular URL and transactional queries (where the user is interested in some Web-mediated activity).
However, Search Engines can't address informational queries, where the user has information that needs to be satisfied.

A Meta Search Engine overcomes the above by virtue of sending the user's query to a set of search engines, collects the data from them displays only the relevant records by using clustering algorithm.

4. Metasearch engines

Meta Search engine combines the strength of multiple search engines, but it is worth pausing to consider in more detail how exactly we expect Meta search to improve the performance of search engines, and in each case how we test how well it works.

A metasearch engine is a search tool that sends user requests to several other search engines and/or databases and aggregates the results into a single list or displays them according to their source. Metasearch engines enable users to enter search criteria once and access several search engines simultaneously. Metasearch engines operate on the premise that the Web is too large for any one search engine to index it all and that more comprehensive search results can be obtained by combining the results from several search engines. This also may save the user from having to use multiple search engines separately.

The term "metasearch" is not only being used but also to describe the paradigm of searching multiple data sources in real time. The National Information Standards Organization (NISO) uses the terms Federated Search and Metasearch interchangeably to describe this web search paradigm.

Benefits of Meta Search Engines

As we are aware, a metasearch engine represents the combination of multiple search engines where in it exhibits a better performance than any search engine The advantages of metasearch engines are that the results can be sorted by different attributes such as host, keyword, date, etc; which can be more informative than the output of a single search engine [16]. Therefore browsing the results should be simpler. On the other hand, the result is not necessarily all the web pages matching the query, as the number of results per search engine retrieved by the metasearch engines are limited. None the less, pages returned by more than one search engine should be more apt.

We observe the following benefits from a metasearch engine.

a. Large data base: As a metasearch engine represents fusion to which more search engines with overlapping data bases are added, user can retrieve more amount of information. Depending on the fusion scheme a document appearing in only one data base may not be as likely to be retrieved by the metasearch engine as a document appearing in all of the data bases [21]. Using a metasearch to obtain a large data base is very important on the web where it is shown that major search engines cover only relatively a small portion of the entire index able web. It is also observed that the amount of the web that is been covered by search engine data bases is Using a metasearch to obtain a large data base is very important on the web where it is shown that major search engines cover only relatively a small portion of the entire index able web. It is also observed that the amount of the web that is been covered by search engine data bases is actually shrinking [8]. A search engine's performance is normally measured with precision and recall.

b. Improved recall: Recall is defined as the ratio of retrieved relevant documents to the total relevant documents. Intrinsically a Meta Search engine uses data fusion scheme, it provides a better and improved recall. Indeed it is observed that different systems retrieve different documents [17]. In one data set each of 61 search engines retrieved 1000 documents for each of 50 queues. An average intersection between pairs of systems on each query is only 238 documents, that different systems are returning many a documents. But it is found out that to achieve higher recall via fusion, it is necessary that the input systems retrieve not just different documents, but they provide different relevant documents [7].

c. Improved Precision: Precision is clearly understood as the ratio of retrieved relevant documents to retrieved documents [3]. It was proved that the odds of a document being relevant, increases monotonically with the number of search engines that retrieve it[18]. There is also another argument that an "Unequal Overlap Property" holds in ranked list fusion. Different retrieval algorithms retrieve many of the same relevant documents, but different irrelevant documents, and if it is true any fusion technique that more heavily weighs common documents should improve precision, but it may likely harms recall as rear relevant documents are de-emphasized.

d. More consistent in Performance: Reliable behavior is considered to be another important and desirable quality of a search engine. It was proved that the same search engine often response to the same query very differently over time, which may be due to the evolution of the data base [15]. Even with a fixed data base it is observed that each search engine will have its strengths and weakness, performing well on some queries and poorly on others.

e. Modular Architected: While designing a search engine, one is faced with many different sources of information about each document. Word frequencies, Phrase frequencies, textual structure within a document, hyper link structure between documents etc. Meta search engine is the answer that provides all of these information that can be incorporated sensibly in such a way that it takes the advantage of each. The architecture of a search engine is modular and a highly specialized sub engine module can be developed and fine tuned for each information source. Each sub engine can alone be used as a search engine, but it may exhibit relatively pure performance. On the other hand queried in parallel and combined by a Meta search core results into a high performance search engine.

f. Focused Ranking Algorithms: Effective Meta search engines may yield unexpected benefits and this may lead towards designing of focused algorithms for ranking documents that can take advantage of novel, highly specific information sources with in document. These focused ranking algorithms are not expected to function well in isolation, but they can improve the search engine's performance when combined with other ranking algorithms [4].

5. Different types of Meta Search Engines

WebCrawler is a metasearch engine that blends the top search results from Google, Yahoo!, Bing Search (formerly MSN Search and Live Search), Ask.com, About.com, MIVA, LookSmart and other popular search engines. WebCrawler also provides users the option to search for images, audio, video, news, yellow pages and white pages [2]. WebCrawler is a registered trademark of InfoSpace, Inc.

WebCrawler was the first Web search engine to provide full text search [1]. It went live on April 20, 1994 and was created by Brian Pinkerton at the University of Washington. It was bought by America Online on June 1, 1995 and sold to Excite on April 1, 1997. WebCrawler was acquired by InfoSpace in 2001 after Excite, (which was then called Excite@Home), went bankrupt. InfoSpace also owns and operates the metasearch engines Dogpile, MetaCrawler and Excite.

WebCrawler was originally a separate search engine with its own database, and displayed advertising results in separate areas of the page[12]. More recently it has been repositioned as a metasearch engine, providing a composite of separately identified sponsored and non-sponsored search results from most of the popular search engines.

Metacrawler is a metasearch engine that blends the top web search results from Google, Yahoo!, Bing (formerly Live Search), Ask.com, About.com, MIVA, LookSmart and other popular search engines. MetaCrawler also provides users the option to search for images, video, news, yellow pages and white pages. It used to provide the option to search for audio. MetaCrawler is a registered trademark of InfoSpace, Inc. MetaCrawler was originally developed in 1994 at the University of Washington by then graduate student Erik Selberg and Professor Oren Etzioni as Selberg's Ph.D. Qualifying Exam project[6]. Originally, it was created in order to provide a reliable abstraction layer to early Web search engines such as WebCrawler, Lycos, and InfoSeek in order to study semantic structure on the Web. However, it became clear that it was a useful service in its own right, and had a number of research challenges.

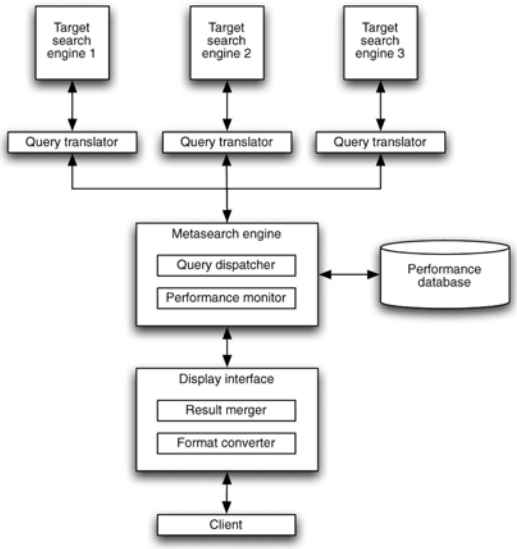
Dogpile is a metasearch engine that fetches results from Google, Yahoo!, Bing, Ask.com,[9] About.com and several other popular search engines, including those from audio and video content providers. It is a registered trademark of InfoSpace, Inc.

Brainboost is a metasearch engine designed to provide specific answers to questions asked in natural language. Currently it only supports English. The Brainboost engine uses machine learning and natural language processing AI techniques to answer the questions [19]. Traditional engines return the links to the pages that appear most relevant. Additionally the results page may include a summary of the page. The user then needs to download the pages and read them to see if the answer to this question exists. Brainboost however, generates a number of different queries that it submits to traditional search engines, downloads several hundred pages returned by the search engines, reads the pages and isolates the answers in the text of these pages, and ranks the different answers based on its *AnswerRank* algorithm [20]. The engine is highly suited to relatively common questions that might have been already dealt somewhere on the Web in one form or another. In December 2005, Brainboost was acquired by Answers Corporation.

Clusty is a metasearch engine which is developed by Vivisimo offered clusters of results. Vivísimo is a company built on Web search technology developed by Carnegie Mellon University researchers. Interestingly Lycos, a search engine that was popular a decade ago, developed at the same university. Clusty adds new features and a new interface to the previous Vivisimo clustering web metasearch [11]. Different tabs also offer metasearches for news, jobs (in partnership with Indeed.com), U.S. government info and blogs. Customized tabs allow users to select sources for their own metasearch to create personalized tabs. Clusty had free toolbars for Internet Explorer and Mozilla Firefox, as well as a Mycroft Project search plugin for Mozilla and Firefox. On May 14 2010, Clusty was acquired by Yippy, Inc., an Internet startup based in Fort Myers, Florida.

Chunklt (Yolink) is a powerful search technology developed by Tiger Logic that mines links and documents to retrieve keyword-rich blocks of information. Based on years of research & development, yolink employs semantic and propriety technology, built in house, to uncover information hidden inside links and documents that is important to you. With an expertise in XML and large-scale, multidimensional databases, yolink technology effortlessly reveals websites hidden information and presents it in a logical fashion [14]. In a phrase, we're proficient at producing highly structured data in a highly unstructured environment.

6. Architecture of a Meta Search Engine



Comparison among various MSEs

	Parameters		
Search Engine	No. of Search Engines Used	Response Time	Ranking Algorithm
Brain boost	16	Less	Clustering Algorithm
Dogpile	14	More	Display the results provided by different search engines
Metacrawler	5	Less	Eliminates duplicates
Chunkit	9	Average	Use Star rating system
Clusty	8	Average	Clustering Algorithm

7. Conclusions

This paper presented a comprehensive survey and understanding of various Meta Search Engines. It is understood that Meta Search Engine exhibits superior performance than any Search Engine and its performance also depends on various factors like recall and precision.

References

[1] Allen, R. B., Obry, P., and Littman, M. 1993. An interface for navigating clustered document sets returned by queries. In *Proceedings of the ACM Conference on Organizational Computing Systems.* ACM Press,166–171.

[2] Anagnostopoulos A , Broder, A and Punera. K. 2006Effective and efficient classification on a searchengine model. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. ACM Press, 208–217.

[3] Carpineto, C. And Romano, G. 2004a. *Concept Data Analysis: Theory and Applications*. Wiley.

[4] Carpineto, C. and Romano, G. 2004b. Exploiting the potential of concept lattices for information retrieval with Credo. *J. Univ. Comput. Sci. 10*, 8, 985–1013.

[5] I. Kang and G. Kim, BQuery type classification

for web document retrieval,[in Proc. 26th Annu. Int. ACM SIGIR Conf. Research Development Information Retrieval, 2003, pp. 64–71.
[6] J. Pitkow, H. Schutze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, BPersonalized search,[Commun. ACM, vol. 45, no. 9, pp. 50–55, 2002.
[7] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird, BLearning collection fusion strategies,[in Proc. 18th Annu. Int. ACM SIGIR Conf. Research Development Information Retrieval, New York, 1995, pp. 172–179.
[8] Cole, R., Eklund, P., and Stumme, G. 2003. Document retrieval for email search and discovery using formal concept analysis. *Appl. Artif. Intell. 17*, 3, 257–280.
[9] Hartigan, J. A. 1975. *Clustering Algorithms*. Wiley
[10] Ling, Y., Meng, X., and Liu, W. 2008. An attributes correlation based approach for estimating size of web databases. Journal of Software, 19, 2(Mar/Apr. 2007), 224-236.
[11] Wu, W., Doan, A., Yu, C., and Meng, W. 2009. Modeling and Extracting Deep-Web Query Interfaces. In Advances in Information and Intelligent Systems. Springer Berlin, Heidelberg, 65-90.
[12] Wu, W., Yu, C., Doan, A., and Meng, W. 2004. An interactive clustering-based approach to integrating source query interfaces on the deep web. In Proc. of the ACM International Conference on Management of Data (Paris, France, June 13-18, 2004) SIGMOD '04.ACM, New York, NY, 95 - 106.
[13] Wikipedia.org
[14] Manoj.M and Elizabeth Jacob Information retrieval on Internet using meta-search engines: A review Journal of Scientific & Industrial Research Vol. 67, October 2008, pp.739-746.
[15] Claudio Carpineto, Stanisiaw Osinski and Giovanni Romano on A Survey on Web Clustering Engines. ACM Computing Surveys, Vol. 41, No. 3, Article 17, Publication date: July 2009.
[16] Ritu Khare, Yuan An and II-Yeol Song on Understanding Deep Web Search Interfaces: A Survey. SIGMOD Record, March 2010 (Vol. 39, No 1).
[17] Harmunish Taneja and Karan Madan on Fusion Based Metasearch : An improved approach towards efficient Web searching.Proceedings of National Conference on Challenges & Opportunities in Information Technology (COIT-2007) RIMT-IET, Mandi Gobindgarh, March 23,2007
[18] Satinder Ball and Rajender Nath To Evaluate the Performance of Metasearch Engines: A Comparative study. Journal of Technology and Engineering Sciences. Vol 1, No. 1 January-June 2009
[19] Lyndon Kennedy and Shih Fu Chang on Query Adaptive Fusion for Multimodal Search. Proceedings of the IEEE vol. 96, No. 4, April 2008.
[20] Ricardo Baeza and Berthier Ribeiro Modern Information Retrieval Pearson Education.
[21] Lyndon Kennedy and Shih Fu Chang on Query Adaptive Fusion for Multimodal Search. Proceedings of the IEEE vol. 96, No. 4, April 2008.