International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue (6) NCRTCST-2013, ISSN 2249–071X

National Conference on Recent Trends in Computer Science and Technology (NCRTCST)-2013

An Analytical Study of Meta Search Engines Performance Based on Precision and Relative Recall

Nagaraju Mamillapally
Lecturer, Department of Computer Science
Adarsh Degree & P.G College
Mahabubnagar, Andhra Pradesh, India.
Email: nagaraj.mavilla29@gmail.com

Trivikram Mulukutla
Lecturer, Department of Computer Science
Adarsh Degree & P.G College
Mahabubnagar, Andhra Pradesh, India
Email: mulukutlatrivikram@yahoo.in

***Abstract* – Finding more useful information quickly in the internet poses a challenge to both the ordinary users and the information professionals. Even the search engines performance is improved with powerful search capabilities of various types, lack of inability to predict the quality of the search results make it difficult for users to use search engines effectively. This paper compared the retrieval effectiveness of the Meta Search Engines Dogpile, Clusty, Zapmeta, Google, Yahoo and Search. Precision and Relative Recall are considered for evaluating the effectiveness of the search engines. Queries using concepts in the field of technology and research were tested and were divided into simple one-word queries and simple multi-word queries. Results of the study showed that the precision of Dogpile was high for simple one-word queries (9.77) and precision of Search was high for simple multi-word queries (9.56). Relative Recall of Google was high for simple one-word queries (5.75) and also for simple multi-word queries (6.37)**

***Keywords* – Internet, Meta Search Engines, Dogpile, Clusty, Zapmeta, Google, Yahoo, Search, Precision, Relative Recall.**

I. INTRODUCTION

Every day many users work on internet to search their necessary information. This is because web search is a key technology and also a primary way to access and read the content available all over the world. Web can be used as a quick and direct reference to get any type of information. Finding useful information quickly on the internet poses a challenge to all web surfers. Sometimes they cannot get appropriate results quickly because of millions of documents which are relevant and irrelevant are coming. The only solution for this is selection of a best Meta search engine. Meta Search Engine provides access to many resources over the web. They extract the content from multiple search engines and filter them as per the user requirements. The internet surfers may not be aware of many search engines functionality and how best they suited for their need.

The major benefits of Meta search engines are their capabilities to combine the related documents from multiple search engines. According to a survey, web has more than 550 millions of pages and user still accessing only 1% of them.

This paper proposes a comprehensive survey on various Meta Search Engines and their performance in the process of information retrieval and their precision and recall values from the Web. Among plenty of Meta search engines, we selected only most frequently used Meta Search engines mainly Dogpile, Clusty, Zapmeta, Google, Yahoo and Search for the study.

II. META SEARCH ENGINES

Meta search engines receives a simple word or group of words as a query from the user and when clicking on search button, it retrieves some millions of documents related to user query and displays as web page links in the order based on their ranks.

The ability of a Meta search engine depends on the display of most relevant documents as per user query. This can be examined by analyzing their precision and relative recall values.

a. Dogpile

Dogpile is a search engine that fetches results from Google, Yahoo! and Yandex, and includes results from several other popular search engines, including those from audio and video content providers. It is a registered trademark of Blucora, Inc. Dogpile began operation in November 1996. The site was created and developed by Aaron Flin and later sold to Go2net (in turn acquired by Infospace).The Dogpile search engine earned the J.D. Power and Associates award for best Residential Online Search Engine Service in both 2006 and 2007. In July 2010, Dogpile was ranked the 770th most popular website in the U.S., and 2548th most popular in the world by Alexa. Quantcast estimated 2.0 million unique U.S. visitors a month, and Compete estimated 1,953,280. Dogpile formerly fetched results from Ask and Bing.

b. Clusty

Clusty was developed by Vivísimo. Vivísimo is a company built on Web search technology developed by Carnegie Mellon University researchers, much like Lycos was a decade earlier. Clusty added new features and a new interface to the previous Vivisimo clustering web metasearch. Different tabs also offer Meta searches 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 plug-ins for Mozilla and Firefox.

By April 2011, Alexa.com was reporting that "while approximately 30% of visitors to the site come from the US, where it is ranked 11,500, it is also popular in Ireland, where it is ranked 2,002." Alexa also reports that the audience for Yippy.com was overrepresented by male cohorts over 45 years old and that it was underrepresented by youth.

c. Zapmeta

ZapMeta is a Meta search engine; which means that the user can search listings from multiple search engines all in one place. ZapMeta gets its' search results from AOL Search, Google, Yahoo,Ask.com, Wisenut, and Yahoo Web. ZapMeta also retrieves directory results from the Open Directory, and shopping listings powered by PriceGrabber. ZapMeta's search is easy. I typed in flipflops and here's what I got:

- *Sponsored results* predictably at the top of the search results page.
- *Results Snapshot option:* The default for the Results Snapshot is to be "off", but you have to turn this on and ZapMeta will then load thumbnail-sized snapshots of each Web search result.
- *Quick View option:* Most of the search results returned had the text "Quick View", which is kind of like the Results Snapshot option, but even better. What this does is open up a screen within your screen of whatever search result you've clicked on - it's a super preview. You can decide quickly whether or not that the site you're looking at is what you want, without leaving the search results page.
- *Search results sorting options:* At the top of your search results, you are given the option to sort your results by relevance, popularity, title, source, or domain.
- *Alexa and Internet Archive:* To the immediate left of the search results Web page titles are three little page icons. The first one leads you to the actual Web page. The second icon with a yellow circle brings you to Alexa's data for that search result. The third little page icon (it has a tenny tiny little blue "i" on it) leads you to a hard-link of Internet Archive's repository for that web site.

d. Google

Google Inc. is an American multinational corporation specializing in Internet-related services and products. These include search, cloud computing, software, and online advertising technologies. Most of its profits are derived from AdWords. Google was founded by Larry Page and Sergey Brin while they were Ph.D. students at Stanford University. Together they own about 16 percent of its shares. They incorporated Google as a privately held company on September 4, 1998. An initial public offering followed on August 19, 2004. Its mission statement from the outset was "to organize the world's information and make it universally accessible and useful", and its unofficial slogan was "Don't be evil". In 2006 Google moved to headquarters in Mountain View, California, nicknamed the Googleplex.

Rapid growth since incorporation has triggered a chain of products, acquisitions and partnerships beyond Google's core search engine. It offers online productivity software including email (Gmail), an office suite (Google Drive), and social networking (Google+).Desktop products include applications for web browsing, organizing and editing photos, and instant messaging. The company leads the development of the Android mobile operating system and the browser- only Chrome OS for a netbook known as a Chrome book. Google has moved increasingly into communications hardware: it partners with major electronics manufacturers in production of its high-end Nexus devices and acquired Motorola Mobility in May 2012. In 2012, a fiber-optic infrastructure was installed in Kansas City to facilitate a Google Fiber broadband service. The corporation has been estimated to run more than one million servers in data centers around the world and to process over one billion search requests and about 24 peta bytes of user-generated data each day. In December 2012 Alexa listed google.com as the most visited website in the world. Numerous Google sites in other languages figure in the top one hundred, as do several other Google-owned sites such as YouTube and Blogger. Its market dominance has led to criticism over issues including copyright, censorship, and privacy.

e. Yahoo

Yahoo! Inc. is an American multinational Internet corporation headquartered in Sunnyvale, California. It is globally known for its Web portal, search engine Yahoo Search, and related services, including Yahoo Directory, Yahoo Mail, Yahoo News, Yahoo Finance, Yahoo Groups, Yahoo Answers, advertising, online mapping, video sharing, fantasy sports and its social media website. It is one of the most popular sites in the United States. According to news sources, roughly 700 million people visit Yahoo websites every month. Yahoo itself claims it attracts "more than half a billion consumers every month in more than 30 languages."

Yahoo was founded by Jerry Yang and David Filo in January 1994 and was incorporated on March 1, 1995. On July 16, 2012, former Google executive Marissa Mayer was named as Yahoo CEO and President, effective July 17, 2012.

According to comScore, Yahoo during July 2013 surpassed Google on the number of United States visitors to its Web sites for the first time since May 2011, set at 196 million United States visitors, having increased by 21 percent in a year.

f. Search

Occasionally Search.com will highlight specialized results that are based on the context of your query. Examples of specialized results include specific links to news, images, or video. Top Matching Results may

highlight information from other Search.com pages, content from the CNET Network of sites, or third party content. The listings are based purely on relevance. Search.com does not receive payment for listings in this section but our partners that provide this data may get paid for listing these products.

This section contains paid listings which have been purchased by companies that want to have their sites appear for specific search terms and related content. These listings are administered, sorted and maintained by a third party and are not endorsed by Search.com.

Search.com sends your search query to several search engines at one time and integrates the results into one list which has been sorted by relevance using Search.com's proprietary algorithm. You can customize the list of search engines included in your metasearch from the preferences. The search engines that are used in your metasearch may allow companies to pay to have their Web sites included within the results. To view the Paid Inclusion policy for a specific search engine, please visit their Web site. Search.com does not accept payment or share revenue with any search engine partner for listings in this section.

III. PROPOSED ALGORITHM

In this section, we proposed an algorithm to evaluate the performance of Meta search engines based on Precision and Relative Recall.

Algorithm to Calculate Precision:

Step 1: Set more_relevant = 0, relevant = 0, irrelevant = 0, links = 0, link_count = 0, userquery_count = 0
Step 2: Select and open a Web Browser.
Step 3: Enter User Query and click on Search option.
Step 4: Record number of sites retrieved by a search engine to calculate Relative Recall
Step 5: Consider each webpage link.
Step 6: If User Query appears in link then
Add 1 to more_relevant
else
If User Query appears in link description
Add 1 to relevant
else
If links appear other than User Query
Add 1 to links
else
Add 1 to irrelevant
Step 7: Add 1 to count.
Step 8: If count ≤ 50 then repeat steps 5 through Step 8 until count becomes 50.
Step 9: Record more_relevant, relevant, irrelevant and links count.
Step 10: Calculate Precision as (more_relevant * 2 + relevant * 1 + irrelevant * 0 + links* 0.5) / 50.
Step 11: Add 1 to userquery_count.
Step 12: Repeat step 3 through step 12 until userquery_count = 10.

Algorithm to Calculate Relative Recall

Step 1: Set query_no = 0, searchengine_no = 0.
Step 2: Define no_of_queries as **n**.
Step 3: Define no_of_searchengines as **s**.
Step 4: Read **n** and **s**.
Step 5: Add 1 to query_no.
Step 6: Read number of sites retrieved by each search engine.
Step 7: Add 1 to searchengine_no.
Step 8: Calculate Relative Recall by dividing number of sites retrieved by a search engine with sum of sites retrieved by 's' search engines.
Step 9: Record Relative Recall value.
Step 10: Repeat Step 7 through Step 10 until searchengine_no ≤ **s**.
Step 11: Repeat Step 5 through Step 11 until query_no = 'n'.

IV. EXPERIMENTAL APPROACH

A total of 10 queries in the technology and research discipline were selected for the study. All the search queries were classified into two categories by the level of search capabilities like simple one-word queries and simple multi-word queries. (See Appendix 1)

In the present study, the search results which are retrieved by 06 search engines are categorized as "more relevant", "relevant", "irrelevant" and "links" on the basis of following criteria.

- If the user query appears in web page link then it is categorized as "more relevant" and given a score of 2.
- If the web page is not closely matched to the user query but consists of some relevant content, then it is categorized as "relevant" and given a score of 1.
- If a message appears as "site can't be accessed" for a particular web page link then it was categorized as "links" and given a score of 0.5.
- If the web page is not at all related to the user query then it was categorized as "irrelevant" and given a score of 0.

This study would measure the relevance of the websites retrieved for each search query. Advanced search options were used for retrieving sites. Only English language pages were searched for each query since the web pages in other languages would be difficult to assess the relevancy. It was specified that the search query must appear in the "title of the web page" or in its short description displayed below the web page link. Since the number of search results retrieved was large, only the first 50 sites were selected for analysis.

Precision of Meta Search Engines

Precision is the ratio of the number of relevant documents retrieved and the total number of irrelevant & relevant documents retrieved.

$$\text{Precision} = \frac{\text{Sum of scores of sites retrieved by search engines}}{\text{Total number of sites selected for evaluation}}$$

Table 1: Precision of Meta Search Engines for Simple One-Word Queries

Search Query	Search Engine	No. of Sites Evaluated	MR	R	IR	L	P
Q 1.1	Dogpile	50	49	0	0	1	1.97
	Clusty	50	42	6	2	0	1.8
	Zapmeta	50	45	2	3	0	1.84
	Google	50	45	2	2	1	1.85
	Yahoo	50	48	1	0	1	1.95
	Search	50	48	1	1	0	1.94
Q 1.1 Total		300	277	12	8	3	11.35
Q 1.2	Dogpile	50	48	2	0	0	1.96
	Clusty	50	32	16	0	8	1.68
	Zapmeta	50	46	0	4	0	1.84
	Google	50	47	2	1	0	1.92
	Yahoo	50	48	1	1	0	1.94
	Search	50	44	1	5	0	1.78
Q 1.2 Total		300	265	22	11	8	11.12
Q 1.3	Dogpile	50	49	0	0	1	1.97
	Clusty	50	30	18	2	0	1.56
	Zapmeta	50	49	1	0	0	1.98
	Google	50	49	1	0	0	1.98
	Yahoo	50	48	2	0	0	1.96
	Search	50	49	1	0	0	1.98
Q 1.3 Total		300	274	23	2	1	11.43
Q 1.4	Dogpile	50	48	0	1	1	1.93
	Clusty	50	42	7	1	0	1.82
	Zapmeta	50	49	1	0	0	1.98
	Google	50	45	4	1	0	1.88
	Yahoo	50	48	2	0	0	1.96
	Search	50	41	5	4	0	1.74
Q 1.4 Total		300	273	19	7	1	11.31
Q 1.5	Dogpile	50	48	1	1	0	1.94
	Clusty	50	32	13	5	0	1.54
	Zapmeta	50	47	3	0	0	1.94
	Google	50	46	2	1	1	1.89
	Yahoo	50	47	2	0	1	1.93
	Search	50	44	1	5	0	1.78
Q 1.5 Total		300	264	22	12	2	11.02

Table 1 (Simple One-word Queries) showed that 80.6% of the sites retrieved by Dogpile are more relevant followed by Yahoo with 79.6%. Clusty retrieved only 59.3 % of more relevant documents which is the least among considered search engines for evaluation. Search retrieved 5% of irrelevant sites which is the maximum value followed by Clusty with 3.3% among Dogpile (0.6%), Google (1.6%), Zapmeta (2.3%). Yahoo retrieved only 0.3% of irrelevant sites.

The Precision values for the search engines were calculated using the above formula. The precision value of Dogpile is high with 1.97 for search queries Q1.1 and Q1.3 and Clusty search engine is low with 1.54 for search query Q1.5.

Table 2: Precision Of Meta Search Engines For Simple Multi-Word Queires

Search Query	Search Engine	No. of Sites Evaluated	MR	R	I	L	P
Q 2.1	Dogpile	50	43	6	1	0	1.84
	Clusty	50	23	10	17	0	1.12
	Zapmeta	50	38	12	0	0	1.76
	Google	50	42	6	1	1	1.81
	Yahoo	50	43	6	0	1	1.85
	Search	50	48	2	0	0	1.96
Q 2.1 Total		300	237	42	19	2	10.34
Q 2.2	Dogpile	50	43	7	0	0	1.86
	Clusty	50	19	26	5	0	1.28
	Zapmeta	50	37	3	0	0	1.54
	Google	50	47	2	0	1	1.93
	Yahoo	50	45	3	1	1	1.87
	Search	50	44	6	0	0	1.88
Q 2.2 Total		300	235	47	6	2	10.36
Q 2.3	Dogpile	50	28	21	1	0	1.54
	Clusty	50	10	22	18	0	0.84
	Zapmeta	50	37	10	3	0	1.68
	Google	50	41	8	0	1	1.81
	Yahoo	50	37	13	0	0	1.74
	Search	50	39	11	0	0	1.78

Q 2.3 Total		300	192	85	22	1	9.39
Q 2.4	Dogpile	50	46	4	0	0	1.92
	Clusty	50	30	14	6	0	1.48
	Zapmeta	50	49	0	1	0	1.96
	Google	50	44	5	0	1	1.87
	Yahoo	50	50	0	0	0	2
	Search	50	48	2	0	0	1.96
Q 2.4 Total		300	267	25	7	1	11.19
Q 2.5	Dogpile	50	41	8	1	0	1.8
	Clusty	50	17	23	10	0	1.14
	Zapmeta	50	41	8	1	0	1.8
	Google	50	47	2	0	1	1.93
	Yahoo	50	40	10	0	0	1.8
	Search	50	44	6	0	0	1.88
Q 2.5 Total		300	230	57	12	1	10.35

Table 2 (Simple Multi-word Queries) showed that 74.3% of the sites retrieved by Search are more relevant followed by Dogpile (67%), Zapmeta (67.3%), Google (73.6%), Yahoo (71.6%). Clusty retrieved only 33% of more relevant sites which is low among other considered search engines for evaluation. Clusty retrieved 18.6% of irrelevant sites where as Search retrieved 0% of irrelevant sites followed by Google and Yahoo with 0.3%, Zapmeta 1.6% and Dogpile 1%.

Based on the precision values, the overall precision of Dogpile, Clusty, Zapmeta, Google, Yahoo and Search are 1.95, 1.68, 1.92, 1.9, 1.95 and 1.84 respectively for simple one-word queries. Search recorded the highest overall precision value with 1.89 followed by Google (1.87), Yahoo (1.85), Dogpile (1.79), Zapmeta (1.74). Clusty recorded the lowest overall precision value of 1.37.

Mean Precision of Meta Search Engines

The mean precision of Dogpile, Clusty, Zapmeta, Google and Search are 9.37, 7.13, 9.16, 9.44 and 9.39. Yahoo had the highest mean precision of 9.50 which is just 0.6 higher than Google as shown in the table 3.

Table 3: Mean Precision of Dogpile, Clusty, Zapmeta, Google, Yahoo and Search

Search Engine	Simple one-word Queries	Simple multi-word Queries	Mean Precision
Dogpile	9.77	8.96	9.37
Clusty	8.4	5.86	7.13
Zapmeta	9.58	8.74	9.16
Google	9.52	9.35	9.44
Yahoo	9.74	9.26	9.50
Search	9.22	9.56	9.39

Relative Recall of Meta Search Engines

Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the database.

$$\text{Relative Recall} = \frac{\text{Total no. of sites retrieved by search engine}}{\text{Sum of sites retrieved by all search engines}}$$

For calculating the relative recall, we considered only Zapmeta, Google, Yahoo and Search Meta search engines. We skipped Dogpile for some technical issue raised while counting the search results and Clusty as it relative recall is almost equal to 0.

Mean Relative Recall of Meta Search Engines

The mean relative recall of Zapmeta, Yahoo and Search are 1.46, 3 and 0.42. Google had the highest relative recall of 6.06 as shown in the table 6.

Table 6: Mean Relative Recall of Zapmeta, Google, Yahoo and Search

Search Engine	Simple One-Word Queries	Simple Multi-Word Queries	Mean Relative Recall
Zapmeta	1.03	1.89	1.46
Google	5.75	6.37	6.06
Yahoo	4.94	1.06	3
Search	0.33	0.51	0.42

V. Future Enhancements

We are planning to extend this study by identifying top 10 Meta search engines performance analysis with sample of at least 100 sites for analysis. The present study recorded the precision values just if the user query appears in the web page link or in its description. But further we continue the recording only if the content displayed in the web page.

VI. Conclusion

The present study estimated the precision and relative recall values of the most frequently used Meta search engines. The results provided evidence that the Google is able to give better search results with more relative recall as compared to other search engines and why it is most widely used search engine on the internet. Dogpile is also given better search results with more precision value because it extracts and filters the documents from most effective multiple search engines like Google and Yahoo.

References

[1] Nagaraju Mamillapally, Trivikram Mulukutla – Performance Evaluation of Meta Search Engines from User's Prespective, Proceedings of National Conference on Advances in Signal Processing, Communications and Networking, Vol 1, Issue 1, p.p.1-5, 29th & 30th August 2013.

[2] B.T.Sampath Kumar, J.N. Prakash –Precision and Relative Recall of Search Engines: A Comparative Study of Google and Yahoo , Singapore Journal of Library & Information Management, Volume 38, 2009.
[3] Tauqeer Ahmed Usmani –A Comparative Study of Google and Bing Search Engines in Context of Precision and Relative Recall Parameter, International Journal on Computer Science and Engineering, Vol 4 No.1 January 2012.
[4] Clarke, S., & Willett, P. (1997). Estimating the recall performance of search engines. ASLIB Proceedings, 49 (7), 184-189.
[5] Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: A comparative study and evaluation methodology. Proceedings of the ASIS 1996 Annual Conference, 33, 127-35.
[6] Ding, W., & Marchionini, G. (1996). A Comparative study of the Web search service performance. Proceedings of the ASIS 1996 Annual Conference, 33, 136-142.
[7] Leighton, H. (1996). Performance of four WWW index services, Lycos, Infoseek, Webcrawler and WWW Worm. Retrieved from http://www.winona. edu /library/webind.htm.
[8] Shafi, S. M., & Rather, R. A. (2005). Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology. Webology, 2 (2), Retrieved from http://www.webology.ir /2005/v2n2/a12.html
[9] Wu, G., & Li, J. (1999). Comparing Web search engine performance in searching consumer health information: Evaluation and recommendations. Bulletin of the Medical Library Association, 87 (4), 456-461.

APPENDIX 1: SEARCH QUERIES

1. Simple One-Word Queries
Q 1.1 Database
Q 1.2 Internet
Q 1.3 Precision
Q 1.4 Recall
Q 1.5 Intranet

2. Simple Multi-Word Queries
Q 2.1 Information Technology
Q 2.2 Research Methodology
Q 2.3 Information Retrieval System
Q 2.4 Search Engines
Q 2.5 Web Usability