IJSAR, 3(8), 2016; 117-123

International Journal of Sciences & Applied Research

www.ijsar.in

An analysis of search engine

E. Ramya* and C. Muruganantham

Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Tamil Nadu, India.

Correspondence Address: *E. Ramya, Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Tamil Nadu, India.

Abstract

Searching useful information and locating appropriate ontology from WWW or semantic Web is an important task in ontology research domain. The most difference between ontology information and common information is that ontology has semantic structure. The algorithm firstly parses input messages and preliminary keywords based results into concepts sets. Then it decides matching rule according to influence of concepts set on ontology semantic and creates weight vector by matching rule. General Queries were tested. Results of the study showed that the precision of Google was high as compared to other two search engines and Yahoo has better precision than Bing.

Keywords: Internet, Search engines, Google, Yahoo, Bing, Precision, Relative recall

Introduction

Search engines in the World? Besides Google and Bing there are other search engines that many not be so well known but still serve millions of search queries per day. It may be a shocking surprise for many people but Google is not the only search engine available today on the Internet! In fact there are a number of search engines that try to remove Google from its throne but none of them is ready (yet) to even pose a threat. Nevertheless, there are search engines that are worth considering and the top 10 are presented below after the break.

Search engine in the United States market, with a market share of 65.6%. Google indexes billions of web pages, so that users can search for the information they desire through the use of keywords and operators.

In 2003, The New York Times complained about Google's indexing, claiming that Google's caching of content on its site infringed its copyright for the content. In this case, the United States District Court of Nevada ruled in favor of Google in Field v. Google and Parker v. Google. The publication 2600: The Hacker Quarterly has compiled a list of words that the web giant's new instant search feature will not search.

The "Hummingbird" update to the Google search engine was announced in September 2013. The update was introduced over the month prior to the announcement and allows users ask the search engine a question in natural language rather than entering keywords into the search box. Search Engines is a web based computer program that find documents, Images, Videos and HTML pages for specified keyword and

returns the high quality results of any webpage containing those keywords. Search engines crawled and indexed the billions of pages and show the results depending upon their importance. World most popular safe search engine is Google end of 2016 62+ billion pages indexed by Google. Top 5 safe search engines is Google, Bing, Yahoo, Ask.com, AOL.

In this information and technology age online business, internet marketing and advertising have created a significant impact. Huge amount of revenues is generated by making and advertising good websites. However, there are billions of websites with vast range of categories and topics worldwide with distinct languages, areas and content types. To find information or website about a particular topic, a user uses search engine by writing search query in words, keywords, or phrases shape. In this era, most advanced search engines include Google, Yahoo, Bing MSN, AOL etc. The goal uses a query evaluation process for searching.

Search engines and search queries

Three search engines namely Bing, Yahoo and Google (BYG) were considered to examine the precision for some selected search queries during March 10, 2013 to March 17, 2013. In order to retrieve relevant data from each search engine, the advanced search features of the search engines were used. Since, more sites were retrieved from the search engines for each query; it was decided to select only the first 30 sites as user hardly goes beyond three to four pages of the search results. Results from India only were selected for evaluation. A total of 15 queries from various discipline were selected for the study.

Precision of search engines

After a search, the user is sometimes able to retrieve relevant information and sometimes able to retrieve irrelevant information. The quality of searching the right information accurately would be the precision value of the search engine (Shafi & Rather, 2005). In this paper, the search results which were retrieved by the Google, Yahoo and Bing were categorized as „more relevant", „less relevant", „irrelevant", „links" and „sites can't be accessed" on the basis of the following criteria (Chu & Rosenthal, 1996; Leighton, 1996; Ding & Marchionini, 1996; Clarke & Willett, 1997):

- If the web page is closely matched to the subject matter of the search query then it was categorized as „more relevant" and given a score of 2.
- If the web page is not closely related to the subject matter but consists of some relevant concepts to the subject matter of the search query then it was categorized as „less relevant" and given a score of 1.
- If the web page is not related to the subject matter of the search query then it was categorized as „irrelevant" and given a score of 0.
- If a web page consists of a whole series of links, rather than the information required, then it was categorized as „links" and given a score of 0.5 if inspection of one or two of the links proved to be useful.
- If a message appears "site can"t be accessed" for a particular URL the page was checked again later. If the message occurs repeatedly the page was categorized as „site can't be accessed" and given a score of 0.

These criteria enabled the calculation of the precision of the search engines for each of the search queries by using the formula:

Precision = Sum of the scores of sites retrieved by a search engine/
Total number of sites selected for evaluation

Materials and methods

Precision of Google

Google, being one of the most popular search engines on the Internet, was selected

as one of the search engines for comparison. Google focuses on the link structure of the Web to determine relevant results and is representative of the variety of easy-to-use search engines. This study would measure the relevance of the web sites retrieved for each search query. Only English pages were searched for each search query since the web pages in other languages would be difficult to assess for relevancy. Since the number of search results retrieved was large, only the first 30 sites were selected for analysis. Around 70% results are found to be more relevant, 13.77% results are less relevant and only 7.77% results are found to be irrelevant. 7.5% results are useful but do not contain any direct information but useful information is found only by clicking on links provided in the results and only 0.44% results were either shown but deleted or pages not available.

Precision of Yahoo

Yahoo is another popular and well-known Internet search engine. The same set of search queries and the same methodology were used in Yahoo. Yahoo is the seco but deleted or pages not available in largest search directory on the web by query volume, at 6.42%, after its competitor Google at 85.35%. Around 67% results are found to be more relevant, 16.88% results are less relevant and only 8.0% results are found to be irrelevant. 6.8% results are useful but do not contain any direct information but useful information is found only by clicking on links provided in the results and only 1.0% results were either shown but deleted or pages not available.

Table 1: Precision of Google.

S.No.	Total Results	Result Selected	More Relevant	Less Relevant	Irrelevant	Link	Site can't be accessed	Precision
1	1,550,000	30	15	8	4	1	2	1.283333
2	12,300,00	30	21	4	4	1	0	1.55
3	2,780,00	30	19	6	0	5	0	1.55
4	1,950,000	30	20	5	4	1	0	1.516667
5	6,260,000	30	20	8	1	1	0	1.616667
6	2,040,000	30	24	3	1	2	0	1.733333
7	534,000	30	23	2	2	3	0	1.65
8	93,100	30	15	10	4	1	0	1.35
9	163,000	30	25	3	0	2	0	1.8
10	652,000	30	20	2	3	5	0	1.483333
11	1,360,000	30	21	5	4	2	0	1.6
12	2,810,000	30	21	3	2	4	0	1.566667
13	2,220,000	30	24	0	5	1	0	1.616667
14	1,500,000	30	26	2	0	2	0	1.833333
15	136,000	30	25	1	1	3	0	1.75
Total	21,268,100	450	319	62	35	34	2	23.9
Percentage			70.88889	13.77778	7.777778	7.555556	0.444444	
Mean								1.59333333

Table 2: Precision of Yahoo.

S. No.	Total Results	Result Selected	More Relevant	Less Relevant	Irrelevant	Link	Site can't be accessed	Precision
1	89,400	30	16	9	3	1	1	1.383333
2	170,000	30	19	6	2	2	1	1.5
3	305,000	30	17	6	4	2	1	1.366667
4	138,000	30	20	7	1	2	0	1.6
5	31,800	30	19	6	3	2	0	1.5
6	75,000	30	27	1	0	1	1	1.85
7	37,100	30	24	5	1	0	0	1.766667
8	24,100	30	20	6	1	3	0	1.583333
9	6,710	30	26	3	0	1	0	1.85
10	16,200	30	24	4	0	2	0	1.766667
11	256,000	30	18	7	2	3	0	1.483333
12	34,000	30	17	4	5	3	1	1.316667
13	24,700	30	19	2	2	7	0	1.45
14	175,000	30	20	4	4	2	0	1.5
15	27,300	30	16	6	8	0	0	1.266667
Total	1,410,310	450	302	76	36	31	5	23.18333
Percentage			67.11111	16.88889	8	6.888889	1.111111	
Mean								1.54555556

Table 3: Precision of Bing.

S. No.	Total Results	Result Selected	More Relevant	Less Relevant	Irrelevant	Link	Site can't be accessed	Precision
1	89,300	30	17	6	4	2	1	1.366667
2	1,72,000	30	19	5	3	3	0	1.483333
3	3,11,000	30	16	5	5	3	1	1.283333
4	1,37,000	30	19	6	3	2	0	1.5
5	32,800	30	20	5	2	2	1	1.533333
6	77,900	30	21	4	3	2	0	1.566667
7	36,400	30	22	5	2	1	0	1.65
8	28,300	30	23	4	3	0	0	1.666667
9	6,720	30	20	6	2	2	0	1.566667
10	16,900	30	21	4	4	1	0	1.55
11	2,53,000	30	19	3	2	6	0	1.466667
12	35,000	30	15	2	3	10	0	1.233333
13	25,000	30	19	3	2	6	0	1.466667
14	1,75,000	30	18	6	2	4	0	1.466667
15	27,300	30	19	7	1	3	0	1.55
Total	375,620	450	288	71	41	47	3	22.35
Percentage			64	15.77778	9.111111	10.44444	0.666667	
Mean Precision								1.49

Precision of Bing

Bing is yet another popular and well-known Internet search engine. The same set of search queries and the same methodology were used in Bing. Bing was unveiled by Microsoft CEO Steve Ballmer on May 28, 2009 at the All Things Digital conference in San Diego for release on June 1. Notable changes include the listing of search suggestions while queries are entered and a list of related searches (called "Explore pane") based on semantic technology from Powerset which Microsoft purchased in 2008. Around 64% results are found to be more relevant, 15.77% results are less relevant and only 9.0% results are found to be irrelevant. 10.0% results are useful but do not contain any direct information but useful information is found only by clicking on links provided in the results and only 0.66% results were either shown but deleted or pages not available.

Results and discussions

Relative recall of Bing, Yahoo and Google

Recall is the ability of a system to retrieve all or most of the relevant documents in the collection (Shafi & Rather, 2005). The relative recall can be calculated using following the formula:

Relative recall = Total number of sites retrieved by a search engine / Sum of sites retrieved by all Search Engines (BYG) The relative recall of the Bing, Yahoo and google (BYG).for general queries was calculated and presented in Table 4. The overall relative recall of the Google was 0.922, for Yahoo was 0.061 and for Bing it was 0.016.

Table 4: Relative recall of Bing, Yahoo and Google.

S. No.	Google	Yahoo	Bing	Total	Relative Recall (Google)	Relative Recall (Yahoo)	Relative Recall (Bing)
1	1,550,000	89,400	89,300	1,728,700	0.896627524	0.051715162	0.051657315
2	1,230,000	170,000	172,000	1,572,000	0.782442748	0.108142494	0.109414758
3	278,000	305,000	311,000	894,000	0.310961969	0.341163311	0.34787472
4	1,950,000	138,000	137,000	2,225,000	0.876404494	0.062022472	0.061573034
5	6,260,000	31,800	32,800	6,324,600	0.989785915	0.005027986	0.005186099
6	2,040,000	75,000	77,900	2,192,900	0.930274978	0.034201286	0.035523736
7	534,000	37,100	36,400	607,500	0.879012346	0.061069959	0.059917695
8	93,100	24,100	28,300	145,500	0.639862543	0.165635739	0.194501718
9	163,000	6,710	6,720	176,430	0.923879159	0.038032081	0.03808876
10	652,000	16,200	16,900	685,100	0.951685885	0.023646183	0.024667932
11	1,360,000	256,000	253,000	1,869,000	0.727661851	0.136971643	0.135366506
12	2,810,000	34,000	35,000	2,879,000	0.976033345	0.011809656	0.012156999
13	2,220,000	24,700	25,000	2,269,700	0.978102833	0.010882495	0.011014672
14	1,500,000	175,000	175,000	1,850,000	0.810810811	0.094594595	0.094594595
15	136,000	27,300	27,300	190,600	0.713536201	0.143231899	0.143231899
Total	21,268,100	1,410,310	1,423,620	24,102,030			
Recall	0.882419448	0.058514158	0.059066394				

Figure 1 shows the relative recall of Bing, Yahoo and Google (BYG) for general queries. In case of Google, the search query 5 had the highest relative recall value of 0.98 and least relative recall for search query 3 of value 0.31. In case of Yahoo, the highest relative recall was for search query 3 (0.34) with the least relative recall for search query 5 (0.005). Similarly, the highest relative recall value of 0.34 for search query and lowest value of 0.005 for search query 5.

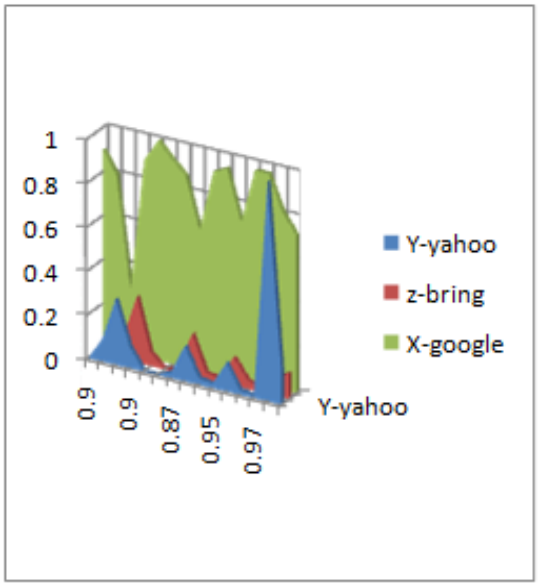

Figure 1: Relative Recall of Google, Yahoo and Bing Search Engines.

X-google	Y-yahoo	z-bring
0.9	0	0
0.8	0.1	0.05
0.3	0.3	0.1
0.9	0.1	0.3
1	0	0.06
0.93	0	0
0.87	0.03	0.03
0.64	0.16	0.19
0.92	0.03	0.03
0.95	0.02	0.02
0.72	0.13	0.13
0.97	0.02	0.04
0.97	0.01	0.02
0.81	1	0.08

Conclusion

While nobody can guarantee top level positioning in search engine organic results, proper search engine optimization can help. Because the search engines, such as Google, Yahoo!, and Bing, are so important today it is necessary to make each page in a Web site conform to the principles of good SEO as much as possible.

To do this it is necessary to:

1. Understand the basics of how search engines rate sites
2. Use proper keywords and phrases throughout the Web site
3. Avoid giving the appearance of spamming the search engines
4. Write all text for real people, not just for search engines
5. Use well-formed alternate attributes on images
6. Make sure that the necessary meta tags (and title tag) are installed in the head of each Web page
7. Have good incoming links to establish popularity
8. Make sure the Web site is regularly updated so the content is fresh.

To get the best search engine visibility, web designers should follow the Five Basic Rules of Web Design, which state that a web site should be:

- Easy to read
- Easy to navigate
- Easy to find
- Consistent in layout and design
- Quick to download

By following these rules, you are building your web site to satisfy your target audience. The added benefit of following these rules is that both directory editors and search engines are looking for these same characteristics.

References

[1] Clarke, S., & Willett, P. (1997). Estimating the recall performance of search engines. ASLIB Proceedings, 49 (7), 184-189.

[2] 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.

[3] Ding, W., & Marchionini, G. (1996). A Comparative study of the Web search service performance. Proceedings of the ASIS 1996 Annual Conference, 33, 136-142

[4] Leighton, H. (1996). Performance of four WWW index services, Lycos, Infoseek, Webcrawler and WWW Worm. Retrieved from http://www.winona.edu/library/webind.htm

[5] 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

[6] 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.