A Comparative Study over Search Engine Optimization on Precision and Recall Ratio

C. Sunitha, B. Meena Preethi and M. Akshay

Abstract--- *Search engines were created to assist users to find information by employing indexing techniques and suggest appropriate alternatives to browse. These search engines have in- efficiencies and are not focused enough to the needs of individual users and little has been done to ensure that the information presented is of a high recall and precision standard. 'Recall' measures how efficient the system is at retrieving the relevant documents from the WWW, while 'precision' measures the relevance of the retrieved set of documents to the users' requirements. This paper presents evaluation of first ten results of search pertaining to 'Computer software ' for estimation of precision and recall. It shows that Yahoo is most comprehensive in retrieving ' Computer software ' followed by Google and HotBot. It also reveals that the search engines (google) perform well on structured queries while Yahoo performs better on unstructured queries.*

Keywords--- *Search Engine, Precision and Recall, Structured and Unstructured Queries, World Wide Web*

I. INTRODUCTION

SEO considers how search engines work, what people search for, the actual search terms or keywords typed into search engines and which search engines are preferred by their targeted audience. Search engines and indices were created to help people find information amongst the rapidly increasing number of World Wide Web (WWW) pages. The Web is growing as the fastest communication medium.

Optimizing a website may involve editing its content, HTML and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of back links, or inbound links, is another SEO tactic.

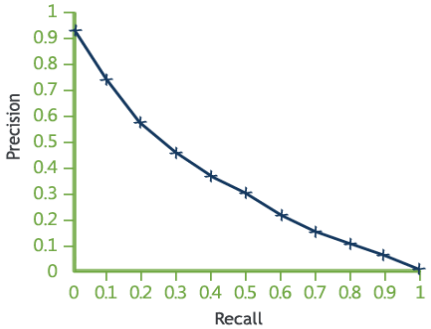
It shows the process of affecting the visibility of a website or a web page in a search engine's "natural" or un-paid ("organic") search results. In general, the earlier (or higher ranked on the search results page), and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users.

SEO may target different kinds of search, including image search, local search, video search, academic search, news search and industry-specific vertical search engines. The reasons include their comprehensive databases having information on different magnitude like media, marketing, entertainment, advertisement etc.

- *Precision*
  - Precision is an important measure of search effectiveness. It is the ability to filter out irrelevant hits and focus on potentially useful information
  - Precision can be difficult to measure in absolute terms because people have subjective opinions of search results
  - Tuning precision starts with an assessment of the organization's content and search requirements

Precision doesn't just happen. It should be a key element in the process of implementing a search solution. Poor precision damages the reputation of a search system and discourages its use. High precision generally impresses search users.

- *Recall*
  - Recall measures how well a search finds every possible document that could be of interest to the searcher
  - Depending on circumstances, many documents may be relevant to a query
  - Recall is particularly important in applications where the user cannot afford to miss information (for example research, security or compliance applications)
  - Recall has less influence on user satisfaction than precision. Many searchers, especially on the Web, are satisfied by precise results, even where recall is low



C. Sunitha, Head of the Department, Department of Computer Applications & Software Systems, Sri Krishna Arts and Science College, Kuniamuthur, Coimbatore - 641008, India.

B. Meena Preethi, Assistant Professor, Department of Computer Applications & Software Systems, Sri Krishna Arts and Science College, Kuniamuthur, Coimbatore - 641008, India. E-mail: preethi.balasundaram@gmail.com

M. Akshay, II M.Sc Software Systems Student (PG), Department of Computer Applications & Software Systems, Sri Krishna Arts and Science College, Kuniamuthur, Coimbatore - 641008, India. E-mail: akshaaysachin@gmail.com

ISBN 978-93-82338-79-6

II. METHODS

SEO techniques can be classified into two broad categories: techniques that search engines recommend as part of good design, and those techniques of which search engines do not approve.

- *Getting Indexed*

The leading search engines, such as Google, Bing and Yahoo, use crawlers to find pages for their algorithmic search results. Pages that are linked from other search engine indexed pages do not need to be submitted because they are found automatically.

Some search engines, notably Yahoo!, operate a paid submission service that guarantee crawling for either a set fee or cost per click. Such programs usually guarantee inclusion in the database, but do not guarantee specific ranking within the search results. Two major directories, the Yahoo Directory and the Open Directory Project both require manual submission and human editorial review.

- *Preventing Crawling*

To avoid undesirable content in the search indexes, webmasters can instruct spiders not to crawl certain files or directories through the standard robots.txt file in the root directory of the domain. Additionally, a page can be explicitly excluded from a search engine's database by using a meta tag specific to robots. When a search engine visits a site, the robots.txt located in the root directory is the first file crawled. The robots.txt file is then parsed, and will instruct the robot as to which pages are not to be crawled. As a search engine crawler may keep a cached copy of this file, it may on occasion crawl pages a webmaster does not wish crawled. Pages typically prevented from being crawled include login specific pages such as shopping carts and user-specific content such as search results from internal searches.

- *Increasing Prominence*

A variety of methods can increase the prominence of a webpage within the search results. Cross linking between pages of the same website to provide more links to most important pages may improve its visibility.

Writing content that includes frequently searched keyword phrase, so as to be relevant to a wide variety of search queries will tend to increase traffic. Updating content so as to keep search engines crawling back frequently can give additional weight to a site. Adding relevant keywords to a web page's meta data, including the title tag and meta description, will tend to improve the relevancy of a site's search listings, thus increasing traffic. URL normalization of web pages accessible via multiple urls, using the canonical link element[37] or via 301 redirects can help make sure links to different versions of the url all count towards the page's link popularity score.

III. RELATED LITERATURE

The growing body of literature on web search engine evaluation is purely descriptive in nature and has little consistency. Scoville (1996) surveyed a wide range of web search engines for examining the relevance of documents retrievable through them. The first ten hits evaluated for precision have shown Excite, Infoseek and Lycos superior. Leighton (1996) evaluated the precision of Infoseek, Lycos, WebCrawler and WWW Worm using eight reference questions and rated Lycos and Infoseek higher. Ding and Marchionini (1996) investigated Infoseek, Lycos and Open Text for precision, duplication and degree of overlap using five complex queries.

The first twenty hits assessed for precision show that the best results are obtained from Lycos and Open Text. Leighton and Srivastava (1997) searched fifteen queries on AltaVista, Excite, HotBot, Infoseek and Lycos taking the first twenty hits for evaluation of precision. Chu and Rosenthal (1996) have investigated AltaVista, Excite and Lycos for their search capabilities and precision. The authors have used ten search queries of varying complexity by evaluating the first ten results for relevance assessment and revealed that AltaVista outperformed Excite and Lycos both in search facilities and retrieval performance.

Clarke and Willett (1997) searched thirty queries of varying nature on AltaVista, Excite and Lycos and obtained best results in terms of precision, recall and coverage from AltaVista. Bar-Ilan (1998) investigated six search engines using a single query "Erdos". All 6,681 retrieved documents examined for precision, overlap and an estimated recall report that no search engine has high recall.

IV. OBJECTIVES

The following objectives are laid down for the study:

- Identification of search engines for retrieval of scholarly information in the field of Computer software.
- Assessment of recall and precision of the select search engines.
- Understanding the effect of nature and types of queries on precision and recall of the select search engines.

The data was analyzed for results.

V. SEARCH ENGINES FOR THE STUDY

The search engines investigated are:

- AltaVista (General)
- Google (General)
- HotBot (General)
- AskJeeves (Science & Technology)
- Yahoo (Computer software)

VI. SAMPLE SEARCH QUERIES

Twenty search terms were drawn out of a sample of 140 terms compiled with the help of "*LC List of Subject Headings*" (LCSH, 2003). These were classified under three groups: single, compound and complex terms for investigating how search engines control and handle single and phrased terms. Single terms were submitted in natural form, compound terms

ISBN 978-93-82338-79-6

as suggested by respective search engines and complex terms with suitable Boolean Operators 'AND' and 'OR' between the terms to perform special searches. Five separate queries were constructed for each term in accordance with the syntax of the select search engine.

VII. Test Environment

The select search engines offer two modes of searching i.e. simple and advanced mode. The study has chosen the advanced mode of search throughout the study to make use of available features for refining and producing precise number of results. In case of AltaVista and Google "match all of the words" was chosen for single and complex terms and "exact phrase" for compound queries. HotBot and AskJeeves offer these options through pull down menus.

Each search was carried out by choosing title field (i.e. all of the words in title) and limiting age of documents published from 2002 to 2010. All the search engines (except AskJeeves and Yahoo) were controlled to retrieve the results in English language. Yahoo on the other hand offered relatively different limiting options among which "relevance then date" and hidden Boolean 'OR' were preferred during search.

Each query was submitted to the select engines which retrieved a large number of results but only the first ten results were evaluated to limit the study in view of the fact that most of the users usually look up under the first ten hits of a query.

Each query was run on all the five select search engines on the same day in order to avoid variation that may be caused due to system updating (Clarke & Willet, 1997). These first ten hits retrieved for each query were classified as scholarly documents and other categories.

VIII. Estimation of Precision and Recall

Precision is the fraction of a search output that is relevant for a particular query. Its calculation, hence, requires knowledge of the relevant and non-relevant hits in the evaluated set of documents (Clarke & Willet, 1997). Thus it is possible to calculate absolute precision of search engines which provide an indication of the relevance of the system. In the context of the present study precision is defined as:

$$\text{Precision} = \frac{\text{Sum of the scores of scholarly documents retrieved by a search engine}}{\text{Total number of results evaluated}}$$

To determine the relevance of each page, a four-point scale was used which enabled us to calculate precision. The criteria employed for the purpose is as under:

- A page representing full text of research paper, seminar/conference proceedings or a patent is given a score of three.
- A page corresponding to an abstract of a research paper, seminar/conference proceedings or a patent is given a score of two.
- A page corresponding to a book or a database is given a score of one.
- A page representing other than the above (i.e. company web pages, dictionaries, encyclopedia, organization, etc.) is given a score of zero.
- A page occurring more than once under different URL is assigned a score of zero.
- A non response of the server for subsequent three searches is assigned a score of zero.

The recall on the other hand is the ability of a retrieval system to obtain all or most of the relevant documents in the collection. Thus it requires knowledge not just of the relevant and retrieved but also those not retrieved (Clarke & Willet, 1997).

There is no proper method of calculating absolute recall of search engines as it is impossible to know the total number of relevant in huge databases.

However, Clark and Willett (1997) have adapted the traditional recall measurement for use in the Web environment by giving it a relative flavour. This study also followed the method used by Clark and Willett by pooling the relevant results (corresponding here to scholarly documents) of individual searches to form the denominator of the calculations.

The relative recall value is thus defined as:

$$\text{Relative Recall} = \frac{\text{Total number of scholarly documents retrieved by a search engine}}{\text{Sum of scholarly documents retrieved by all five search engines}}$$

However, in the case of overlapping between search engines results, only the overlapped results are included for the pooling by taking five search engines (say a, b, c, d and e) into consideration which retrieve a1, b1, c1, d1 and e1 scholarly documents respectively. Further, where there is no overlap between search engines (i.e. $a \cap b$, $a \cap c$, $a \cap d$ and $a \cap e$ is zero) then the relative recall of search engine 'a' is calculated as a1/(a1+b1+c1+d1+e1).

Again if overlapping exists between search engines i.e. $a \cap b = b2$, $a \cap c = c2$, $a \cap d = d2$ and $a \cap e = e2$ then the relative recall of engine 'a' is a1/(a1+b2+c2+d2+e2). The relative recall is more in case of overlapping between search engines. The mean values for precision and relative recall is obtained by micro-averaging (Clarke & Willet, 1997; Tague, 1992) i.e. average score for each engine against a query is summed over all the twenty queries and mean value calculated from these totals for single, compound and complex terms separately.

IX. Engines Revisited

Two search engines namely AltaVista and HotBot were revisited during June 2005 to investigate the effect of their changing algorithm policy on precision and recall. The mean precision and recall of the observations in AltaVista show a slight increase while as HotBot shows marginal increase in precision but decrease in its recall value (Table 2).

ISBN 978-93-82338-79-6

X. SIGNIFICANT FEATURES INFERENCE

Search engines keep their ranking algorithms and the features that are used to determine the relevance of a page secret. However, to be able to understand which features might be abused by spammers and malware authors to push their pages, a more detailed understanding of the page ranking techniques is necessary.

Thus, the goal of the first step of our work is to determine the features of a web page that have the most-pronounced influence on the ranking of this page.

Table 1:Feature Set used for Inferring Web

1 Keyword(s) in title tag
2 Keyword(s) in body section
3 Keyword(s) in H1 tag
4 Amount of indexable text
5 Keyword(s) in URL file path
6 Number of inbound links
7 Anchor text of inbound links
8 External links to low quality sites
9 External links to high quality sites
10 Inlinks Vs Outlinks distribution

A feature is a property of a web page, such as the number of links pointing to other pages, the number of words in the text, or the presence of keywords in the title tag. To infer the importance of the individual features, we perform "black-box testing" of search engines. More precisely, we create a set of different test pages with different combinations of features and observe their rankings. This allows us to deduce which features have a positive effect on the ranking and which contribute only a little. Once the features were selected, the next step was to create a large set of test pages, each with a different combination and different values of these features.

For these test pages, we had to select a combination of search terms (a query) for which no search engine would produce any search results. We arbitrarily chose "gerridae plasmatron" as the key phrase to optimize the pages for inferring the importance. For features whose possible values exceed the Boolean values (i.e., present or absent), such as keyword frequencies, we selected representative values that correspond to one of the following four classes.

- The feature is not present at all.
- The feature is present in *normal* quantities.
- The feature is present in *elevated* quantities.
- The feature is present in *spam* quantities.

The first five features are termed to be δ, representing the content quality and the second five features are termed as ϕ and the CQ represents the content quality assessment and LQ represents the link quality assessment and thus it could be derived as follows in eq 1 and eq 2.

$$CQ(w) = \sum_{w_i:\text{pointstoW}} \delta \cdot \frac{CQ(w_i)}{ON(w_i)} + (1-\delta).\frac{LQ(w_i)}{ON(w_i)} \qquad (1)$$

$$LQ(w) = \sum_{x_i:\text{point by w}} \phi \cdot \frac{CQ(x_i)}{IN(x_i)} + (1-\phi).\frac{LQ(x_i)}{IN(x_i)} \qquad (2)$$

Given G is the graph denoting the candidate, p denotes the pages present in each node of G, Thr is the threshold

1 G := set of pages
2 for each page p in G do
3 p.auth = 1 // p.auth is the authority score of the page p
4 p.hub = 1 // p.hub is the hub score of the page p
5 function HubsAndAuthorities(G)
6 for step from 1 to k do // run the algorithm for k steps
7 for each page p in G do // update all authority values first
8 for each page q in p.incomingNeighbors do // p.incomingNeighbors is the set of pages that link to p
9 p.auth += q.hub
10 for each page p in G do // then update all hub values
11 for each page r in p.outgoing Neighbors do // p.outgoing Neighbors is the set of pages that p links to
12 p.hub += r.auth
13 calculate the intersection of p.hub and p.auth, if the number of elements in the intersection set is equal to or bigger than the threshold Thr, mark p as a bad page.
14 for each bad page bp
15 calculate content and link quality using eq(1), eq(2)
16 repeat for every website
15 end

XI. RESULTS AND DISCUSSION

The mean precision and relative recall of select search engines for retrieving scholarly information are presented in Table 1.

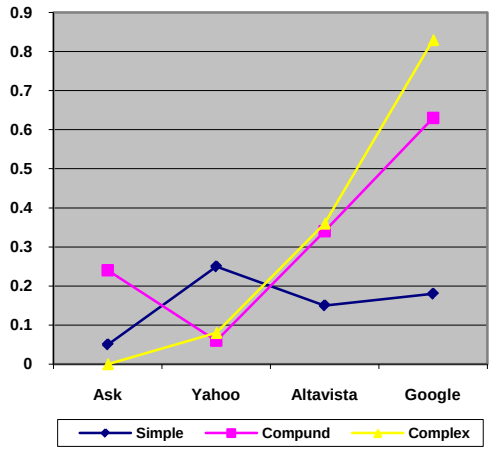
Table 2: Mean Precision and Relative Recall of Search Engines during 2010

	Altavista	Google	HotBot	AskJeeves	Yahoo
Precision	0.27	0.29	0.28	0.57	0.14
Recall	0.18	0.20	0.29	0.32	0.05

Table 3: Comparison of Mean Precision and Mean Recall of AltaVista and HotBot Search Engines between 2009 and 2010

Search Engine	Mean Precision 2010	Mean Precision 2009	Mean Recall 2010	Mean Recall 2009
Altavista	0.27	0.29	0.18	0.21
HotBot	0.28	0.33	0.29	0.27

ISBN 978-93-82338-79-6



Comparing the mean precision, AskJeeves scored the highest rank (0.57) followed by Google (0.29) and HotBot (0.28). AltaVista obtained (0.27) while Yahoo received the lowest precision (0.14). The mean precision obtained for single, compound and complex queries of the respective search engines show AskJeeves as having the highest precision (0.83) for complex queries followed by compound queries (0.63). AltaVista scored the highest precision (0.50) for complex queries followed by compound quires (0.24). Google and HotBot performed better with complex and compound queries while Yahoo performed better with single queries.

XII. Conclusion

Comparing the corresponding mean relative recall values, AskJeeves has the highest recall (0.32) followed by HotBot (0.29) and Google (0.20). AltaVista scored a relative recall of 0.18 and Yahoo the least (0.05). While AskJeeves performed better on complex queries (0.39) followed by compound queries (0.37). HotBot did better in single and compound queries (0.31). Google attained highest recall on compound queries (0.22) followed by complex queries (0.21). AltaVista's performance is better on complex queries (0.28) where as Yahoo performed better on single queries (0.11).

The results depict better performance of AskJeeves in retrieving scholarly documents and it is the best choice for those who have access to various online journals or databases. Google is the best alternative for getting web-based scholarly documents and its recent introduction of 'Google Scholar' in its beta test for accessing scholarly information offers better dividends for researchers.

AskJeeves acquired the highest recall and precision due to the induction of its journal citations along with web resources; otherwise Google would rank the first. HotBot offers a good combination of recall and precision but has a larger overlap with other search engines which enhance its relative recall over Google search engine. AltaVista once prominent on the Web has lagged behind and the Yahoo is the weakest among the select search engines in all respects.

Further, the results reveal that structured queries (i.e. phrased and Boolean) contribute in achieving better precision and recall. The findings also establish the case that precision is inversely proportional to recall i.e. if precision increases recall decreases and vice versa.

References

[1] Bar-Ilan, J. (1998). On the overlap, the precision and estimated recall of search engines: A case study of the query "Erdos". *Scientometrics*, 42 (2), 207-208.

[2] Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology. In: *Proceedings of the ASIS 1996 Annual Conference*, October, 33, 127-35. Retrieved August 19, 2003 from http://www.asis.org/annual-96/ElectronicProceedings/chu.html

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

[4] Ding, W., & Marchionini, G. (1996). A comparative study of the Web search service performance. In: *Proceedings of the ASIS 1996 Annual Conference*, October, 33, 136-142.

[5] Leighton, H. (1996, June 25). Performance of four WWW index services, Lycos, Infoseek, Webcrawler and WWW Worm. Retrieved June 10, 2005 from http://www.winona.edu/library/webind.htm

[6] Leighton, H., & Srivastava, J. (1997). Precision among WWW search services (search engines): AltaVista, Excite, HotBot, Infoseek and Lycos. Retrieved June 11, 2005 from http://www.winona.edu/library/webind2.htm

[7] Library of Congress (2003). *Library of Congress Subject Headings* (vol.s 1-5). Washington: Library of Congress, Cataloging Distribution Service.

[8] Modi, G. (1996). Searching the Web for gigabucks. *New Scientist*, 150 (2024), 36-40.

[9] Oppenheiem, C., Moris, A, Mcknight, C., & Lowley, S. (2000). The evaluation of WWW search engines. *Journal of documentation*, 56 (2), 190-211.

ISBN 978-93-82338-79-6