*DESIDOC Bulletin of Information Technology*, Vol. 25, No. 2, March 2005, pp. 3-10


# Search Engines Evaluation

**Rakesh Kumar, P.K. Suri, & R.K. Chauhan**

**Abstract**

The volume of world wide web (WWW) is increasing enormously due to a world wide move to migrate information to online sources. To search some information on WWW, search engines are used, which when presented with queries, return a list of web pages ranked on the basis of estimation of relevance. Generally the search engines due to the abundance of information available on the web return millions of pages. But user studies indicate that a common user browses through top 10 or 20 documents only. So it's all-important to get into those top 10 documents. To achieve this web authors are increasingly beginning to rely on underhand techniques to ensure their sites get seen, in turn affecting the performance of search engines. The existing measures to evaluate these systems' performance are not adequate in the current world of highly interactive end-user systems. In this study a metric 'Ranked Precision' is proposed to evaluate the performance of search engines.

## 1. INTRODUCTION

The goal of an information retrieval system (IRS) is to locate relevant documents in response to a user's query. Documents are typically retrieved as a ranked list, where the ranking is based on estimation of relevance[1]. Today, while searching for information on world wide web (WWW), one usually performs a query through a term-based search engine. The WWW is a rapidly expanding hyperlinked collection of unstructured information. The friendly user interface and the hypermedia features have been attracting a number of information providers. As a result, the web has become a sea of all kinds of data. The lack of the structure and the enormous volume of the WWW poses tremendous challenge on the search engines.

It is observed that a number of companies have developed various search engines such as Google, Lycos, Hotbot, Yahoo, Excite, AltaVista, etc. How do these search engines differ from one another in performance? How to evaluate and measure their effectiveness? What are the existing measures? What are their limitations? These questions invite a research. In this study, we attempted to find the answer of some of these questions and proposed a metric to measure the performance of search engines.

### 1.1 Existing Measures

The terms 'measure', 'measurement', and 'metrics' are often used interchangeably, but according to Pressman,[2] a measure provides a quantitative indication of the extent, amount, dimensions, capacity, or size of some attribute of a product or process. Measurement is the act of determining a measure. The IEEE Standard Glossary of Software Engineering Terms[3] defines metrics as "a quantitative measure of the degree to which a system, component, or process possesses a given attribute."

Cleverdon[4] listed six main measurable quantities of IRS:

(1) Coverage of the collection

(2) Time lag, i.e., the average interval between the time the search request is made and the time an answer is given

(3) Form of presentation of the output

(4) Effort involved on the part of the user in obtaining answers to his search requests

(5) Recall of the system

(6) Precision of the system.

It is claimed that (1)-(4) are readily assessed. It is recall and precision, which attempt to measure the effectiveness of the IRS. Precision and recall are metrics that traditionally define the 'quality' of the retrieved documents set. As shown in the figure 1, these measures assume: (a) There is a set of records in the database, which is relevant to the search topic, (b) Records are assumed to be either relevant or irrelevant, (c) Actual retrieval set may not perfectly match the set of relevant records.

### *1.1.1 Recall*

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. Trying to accomplish a high recall rate on the internet is difficult, due to the enormous volume of information, which must be searched, and it is impossible to know the total number of relevant documents in a collection not found by a search. Recall is usually expressed as percentage.

The formula to calculate recall is:

Recall = (B/(A+B))*100

### *1.1.2 Precision*

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. The formula to calculate precision is:

Precision = (B/(B+C))*100

The 100% precision is an obtainable goal, since the system could be programmed to return just one completely relevant document. However, returning just one document might not resolve the seeker's query. Therefore, a search system attempts to maximize both precision and recall simultaneously.

### *1.1.3 Fallout*

Fallout is all the junk that came up in search that was irrelevant. If one retrieves 100 documents and 20 are relevant, then fallout is 80%. Fallout becomes a bigger issue (and bigger problem) as the size of database grows and retrieval gets larger. Scanning 80 irrelevant documents to find 20 relevant ones may not be so bad, but scanning 1000 documents to find 200 relevant would be painful.

## 2. RELATED WORK

The use of precision and recall goes back to 1955 when Kent, *et al.,* proposed them as the primary measures of performance. Ever since, those two criteria have been used in many search engine evaluations. Raghavan *et al*[5] question the usefulness of precision and recall as measures of IR systems performance and argue that in real life, items are more or less relevant, and that precision requires a definition of what is to be

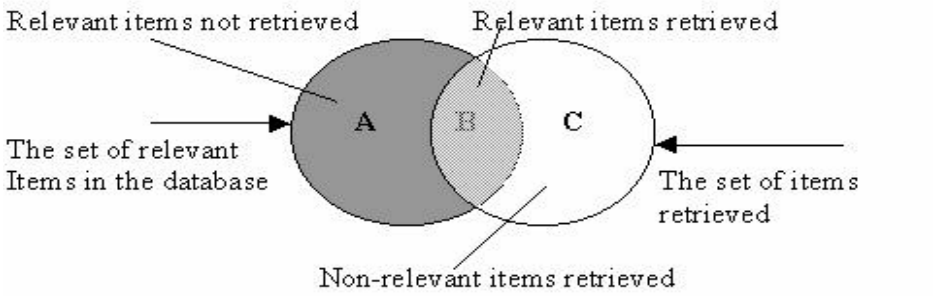


**Figure 1. Precision and recall**

considered as relevant. Saracevic[6] notices that relevance is a complex human cognitive and social phenomenon, and argues against binary yes-no decisions. Nielsen[7] points out that precision and recall assumes that the users want a complete set of relevant documents. This might have been true in traditional information retrieval, Nielsen argues, but on the web nobody will have time to read all relevant documents, so it is more important to present a small but useful sample. Chu and Rosenthal[8] point out that recall, as an evaluation criterion for web searching is problematic, since it is almost impossible to know exactly how many relevant documents there are for a specific query. It may however still be useful to do precision tests after more than one product have been selected and installed, since one can compare relative recall between systems. They tested first ten precisions, had enough queries for statistical comparisons, recorded crucial information about how the searching was conducted, and performed some statistical tests. They compared three search services (AltaVista, Excite and Lycos) on the basis of precision calculated for the first ten hits. Their study also proposed a methodology for evaluating WWW search engines in terms of the following five aspects:

(1) Composition of web indexes

(2) Search capability

(3) Retrieval performance (precision, recall, time lag)

(4) Output option

(5) User effort.

Ding and Marchionini,[9] the best-modeled study to date, studied first twenty precisions, but used only five queries. Westera[10] used five queries, all dealing with wine. Leighton's[11] 1995 study evaluated the performance of four index services: Infoseek, Lycos Webcrawler, and WWW Worm. Employed measures included: average top 10 precisions and response time. Gauch and Wang[12] had twelve queries, studied almost all of the major search services (and even the major metasearch services) and reported first twenty precisions. Tomaiuolo and Packer[13] studied first ten precisions on two hundred queries. They did list the query topics that they searched, but they used structured search expressions (using operators) and did not list the exact expression entered for each service. Scoville[14] used first ten precisions and gave exact mean scores. Meghabghab and Meghabghab[15] examined the effectiveness of five WWW search engines (Yahoo, Webcrawler, Infoseek, Excite, and Lycos) by measuring precision on five queries. They found that Yahoo obtained the best performance, followed by Infoseek and Lycos.

## 3. PROBLEM STATEMENT

As observed from all the above-mentioned studies, generally precision is used as a measure of effectiveness, but it has certain limitations. In order to identify the limitations of precision with respect to WWW, one has to focus on the two aspects: (a) Behaviour of search engine user, and (b) Keyword spamming.

### 3.1 Behaviour of Search Engine Users

There are generally millions of the sites pertaining in some manner to the queries submitted by the user, search engine will retrieve them, rank them on the basis of their relevance, and present a huge list of documents. But studies show that approximately 80% of web searchers never view more than first 10 results in a result's list.[16, 17, 18] A search engine typically displays results page by page, perhaps 10 hits at a time. Many users tend to look at only the first page of results they get from a search engine, rather than scrolling through pages and pages of hits. Several studies have analysed the queries that users submit to search engines, and the length of search sessions.[19, 20] The finding that these studies share is that users browse through very few result pages. These studies differ in the reported distribution of page views, but agree that at least 58% of the users view only the first page (the top-10 results), and that no more than 12% of users browse through more than 3 results pages. While the above describes the behaviour of users as they browse through multiple result

pages, statistics have also been gathered on the browsing patterns of users as they view a single page of results. It has been observed that users are reluctant to scroll beyond the visible part of the page, and so search results that are 'above the fold' are viewed (and clicked on) by more users than results at the bottom of the page.[21]

## 3.2 Keyword Spamming

The search engines face a web-specific obstacle, called search engine persuasion.[22] With the growing economic impact of the WWW, it is crucial for businesses to have their sites ranked high by the major search engines. In the increasingly market-driven environment, it is becoming even more important to ensure that your site gets seen by as many people as possible. In particular, corporate and organisations who are trying to attract advertising revenue to their web sites need to ensure that they can show potential advertisers that placing an advertisement on their pages is likely to achieve maximum exposure. There are quite a few companies who sell this kind of expertise; they design web sites, which are tailored to rank high with specific queries on the major search engines. Under such competitive circumstances, web authors are increasingly beginning to rely on underhand techniques to ensure their sites get seen. Emerging practices are popularly known as keyword spamming and site optimization. These companies research the ranking algorithms and heuristics of term-based search engines and know how many keywords to place (and where) in a web page so as to improve the page's ranking. This directly affects the page's visibility.[23] A less sophisticated techniques used by some site creators, is called keyword spamming.[24] Keyword spamming is doctoring the content of a web site to ensure that it hits the top of a list of results retrieved from a web search engine. Keyword spamming can be done by following ways:

### *3.2.1 Keyword Stuffing META Tags*

Search engine spiders index documents using Meta tags in documents - tags in the head of a document, which can be used to provide information on ownership and content. Alta Vista uses Meta tags for indexing if these are present, in preference to indexing the first hundred lines of the document. Meta tags are used quite legitimately for indexing of web documents. Web authors, however, can exploit and abuse the use of Meta tags in order to push their sites up the relevancy ratings.

### *3.2.2 Hidden Text*

Another method used was that of hidden text. Black text was used against a black background near the bottom of the web page. This didn't show up on screen as the background effectively masked it. However, it was visible in the HTML source of the document, where it was picked up and indexed by search engines. A huge number of keywords were slipped into the web site by stealth using this method.

### *3.2.3 Keyword Stuffing Text*

A slightly more worrying development is the tendency for web authors to include keywords in their documents, which bear no relation to the subject of the document.

Now it can be summarized that: (a) Non relevant documents may appear in the list of documents displayed by the search engine and even on the very first page, (b) How many relevant documents retrieved out of total documents retrieved is not sufficient to measure the effectiveness. Where the relevant documents are ranked is also equally important. The precision measure does not take care of this. If out of N documents retrieved, M documents are relevant then precision will remain same i.e., M/N irrespective of the fact that non-relevant documents retrieved (N-M) are ranked high or low.

## 4. RANKED PRECISION

To analyse the performance of the search engines, we have devised a metric, Ranked Precision (RP), which returns a number between 0 and 1. The RP measures the effectiveness of a search engine in terms of: (a) First n documents, retrieved and ranked according to their relevance by search

engine, (where n represents some top ranked documents generally browsed by the user i.e., 20 to 30.), and (b) Number of relevant documents retrieved and their rank in first n documents.

In order to use this metric to evaluate the performance of a search engine, queries are to be unambiguous. Because the term based search engines faces the following problems when handling broad-topic queries:[25]

(1) Synonymy: Retrieving documents containing the term car when given the query automobile

(2) Polysemy/ambiguity: When given the query Jordan, should the engine retrieve pages pertaining to the Hashemite Kingdom of Jordan, or pages pertaining to the basketball legend Michael Jordan.

The formula to compute RP metric is given below:

$$RP = \left(\left(\sum_{j=1}^{j=n} ((n+1) - r_j)\right) * w_j / (n(n+1)/2)\right)$$

Where n = first n documents displayed under consideration.

k = Number of relevant documents in n documents.

$r_j$ = rank of jth document.

$w_j$ = relevance score of jth document.

RP = Ranked precision

The model for calculating relevance score is given in table 1.

**Table 1. Subjective relevance score**

| Relevance Score | Description |
|---|---|
| 1 | The most relevant |
| .75 | Partly relevant |
| .5 | Somewhat relevant |
| 0 | Not relevant |

The RP metric can be easily used to compute objective ranked precision, useful ranked precision, and best ranked precision by simply assigning the relevance score to the documents as given in table 2, table 3, and table 4.

**Table 2. Relevance score for objective ranked precision**

| Relevance score | Description |
|---|---|
| 1 | Relevant (Most, partly or somewhat) |
| 0 | Not relevant |

**Table 3. Relevance score for useful ranked precision**

| Relevance score | Description |
|---|---|
| 1 | Most relevant or partly relevant |
| 0 | Somewhat relevant or Not relevant |

**Table 4. Relevance score for best ranked precision**

| Relevance score | Description |
|---|---|
| 1 | Most relevant |
| 0 | Partly relevant, somewhat relevant or not relevant |

## 4.1 Metric Characteristics

The RP metric realizes all the requirements of a good metric as proposed by Ejiogu.[26] He suggested that a metric should possess the following characteristics:

(1) Simple and computable: It should be easy to learn how to derive the metric and its computation should not be effort and time consuming. RP is simple and computable because it does not demand inordinate efforts.

(2) Empirically and intuitively persuasive: The metric should satisfy the engineer's intuitive notion about the product under consideration. The metric should behave in certain ways, rising and falling appropriately under various conditions. RP returns a value between 0 and 1. It increases in value as the level of

effectiveness of search engine increases. It is evident from the following table 5.

In table 5, k represents the number of non-relevant documents, and r represents the rank of non-relevant document, and n is taken 10 (i.e., number of documents displayed on the first page by search engine). In this case we have assumed that there is only one spammed document per observations. Although the results can be verified for k>1. It is clear from it if the very first ranked document is a non-relevant document then the value of RP is minimum and it increases progressively as the rank of non-relevant document increases. From the Table 5 it is also observed that RP is a better measure of effectiveness in comparison to precision, which does not reflect the effect of rank of non-relevant document. As observed from the table 5 the precision in every case is 0.9. The figure 2 is the graphical comparison of RP and precision.

(3) Consistent and objective. The metric should always yield results that are unambiguous. The third party would be able to derive the same metric value using the same information. This is true in case of RP.

(4) Consistent in its use of units and dimensions It uses only those measures that do not lead to bizarre combinations of units.

(5) Programming language independent

(6) An effective mechanism for quality feedback. Ranked Precision provides search engine developers with information that can be used by them to evaluate the effectiveness of their products and subsequently lead to development of a higher quality product.

**Table 5. Precision vs. ranked precision**

| n | k | R (Rank) | ORP | Precision |
|---|---|---|---|---|
| 10 | 1 | 1 | 0.818182 | 0.9 |
| 10 | 1 | 2 | 0.836364 | 0.9 |
| 10 | 1 | 3 | 0.854545 | 0.9 |
| 10 | 1 | 4 | 0.872727 | 0.9 |
| 10 | 1 | 5 | 0.890909 | 0.9 |
| 10 | 1 | 6 | 0.909091 | 0.9 |
| 10 | 1 | 7 | 0.927273 | 0.9 |
| 10 | 1 | 8 | 0.945455 | 0.9 |
| 10 | 1 | 9 | 0.963636 | 0.9 |
| 10 | 1 | 10 | 0.981818 | 0.9 |

In addition to the above-mentioned characteristics, Roche[27] suggests that metric should be defined in an unambiguous manner. According to Basili,[28] metrics should be tailored to best accommodate specific products and processes. If we compare RP with recall, recall requires the how many relevant documents are available for a specific query, which is almost impossible to find in case of web. Recall can be good metric to evaluate traditional IRSs, but it cannot be used to measure the effectiveness of search engines. The RP can be used easily and unambiguously to measure search engines. So it can be inferred that RP caters all these requirements.

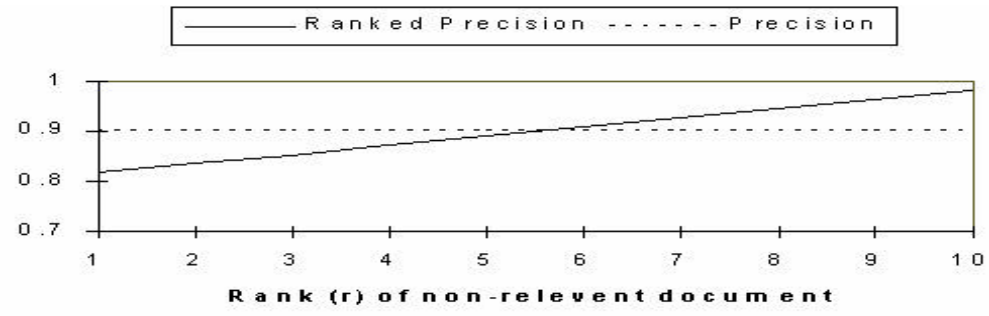


**Figure 2. Precision vs. ranked precision**

## 5. CONCLUSION

Lord Kelvin once said that when you can measure what you are speaking about and express it in numbers, you know something about it. Measurement is fundamental to any engineering discipline and IRS is no exception. Most evaluations of these systems (search engines) use methods that are no longer adequate in the current world of highly interactive end-user systems. It was observed that a number of underhand practices such as keyword spamming are used by web site developers to promote their sites so that as many people see it as possible. A search engine should be able to identify these types of sites and penalize them by ranking them at bottom. It is always better for a good page to be higher in the ranked list. The user will see and evaluate the pages listed first, so pages located there should be given more weight. The RP metric yield a lower value if the non-relevant documents are ranked higher and penalize them. The proposed RP metric may help the designers of search engines to measure the effectiveness of their products. This feedback may be used by them to improve into a higher quality product that can satisfy the requirements of the end users.

## REFERENCES


1. Belkin, N., and Croft, W.B. Retrieval techniques. *Ann. Rev. Inf. Sci. and Tech. (ARIST)*, 1987, **22**, 109-45.
2. Pressman, S. Roger. Software engineering: A practitioner's approach. McGraw-Hill, 1997.
3. IEEE Software Engineering Standards, Std. 610.12-1990, pp.47-48,1993.
4. Cleverdon, C.W.; Mills, J.; & Keen, E.M, An inquiry in testing of information retrieval systems. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK, 1966.
5. Raghavan, V.; Bollman, P.; & Jung, G.S. A critical investigation of recall and precision as measures of retrieval system performance. *Communication of the ACM,* 1989, **7**(3), 205-29.
6. Saracevic, T. Evaluation of evaluation in information retrieval. *In* Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. *Special issue of SIGIR Forum*, 1995, 138-46.
7. Nielsen, J. User interface directions for the web. *Communications of the ACM,* 1999, **42** (1), 65-72.
8. Chu, H., & Rosenthal, M. Search engine for the world wide web: A comparative study and evaluation methodology. *In* Proceedings of the Annual Conference for the American Society for Information Science, 19-24 October 1996, 127-35.
9. Ding, Wei and Gary Marchionini. A comparative study of web search service performance. *In* Proceedings of the Annual Conference for the American Society for Information Science, 19-24 October 1996, 136-42.
10. Westera, Gillian. Search engine comparison: Testing retrieval and accuracy. http://WWW.curtin.edu.au/curtin/library/staffpages/ gwpersonal/senginestudy/ results.htm (7 February 1997).
11. Leighton, H. Vernon. Performance of four world wide web (WWW) index services: Infoseek, Lycos, WebCrawler, and WWWWorm. 1995. http://WWW.winona.edulibrary//webind.htm (1 July 1996).
12. Gauch, Susan and Wang, Guijun. Information fusion with profusion. Webnet 96 Conference, San Francisco, CA, October 15-19, 1996. http://WWW.csbs.utsa.edu:80/info/webnet96/html/155.htm [22 February 1997].
13. Tomaiuolo, Nicholas G. & Packer, Joan G. An analysis of Internet search engines: Assessment of over 200 search queries. *Computers in Libraries,* 1996, **16**(6), 58. The list of queries used is in: Quantitative Analysis of Five WWW 'Search Engines' http://neal. ctstateu. edu: 2001/ htdocs/ websearch.html (7 February 1997).

14. Scoville, Richard. Special report: Find it on the net! *PC World, 1996,* **14**(1), 125-26. http://WWW. pcworld. com/ reprints/ lycos.htm (1 February 1997).

15. Meghabghab, D.B. and Meghabghab, G.V. Information retrieval in cyberspace. Proceedings of ASIS Mid-Year Meeting, 1996.

16. Holscher, C. & Strube, G. Web search behaviour of internet experts and newbies. *International Journal of Computer and Telecommunication Networking*, 2000,**33**, 337-46.

17. Silverstein C.; Henzinger, M.; Marais, H.; & Moricz, M. Analysis of a very large web search engine query log. *SIGIR Forum,* 1999, **33**, 6-12.

18. Jansen, B.J.; Spink, A.; & Saracevic, T. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 2000, **36**, 207-27.

19. Markatos, E.P. On caching search engine query results. *In* Proceedings of the 5th Inter-national Web Caching and Content Delivery Workshop, May 2000.

20. Lempel, R.; & Moran, S. Predictive caching and prefetching of query results in search engines. *In* Proceedings of the 12th World Wide Web Conference (WWW2003), Budapest, Hungary, 2003.

21. Broder, A. Web search. Invited talk at the 9th Text Retrieval Conference (TREC-9), Gaithersburg, Maryland, November 2000.

22. Marchiori, M., The quest for correct information on the web: Hyper search engines. *In* Proceedings of 6th International Conference on WWW, 1997.

23. Lempel, R. & Moran, S. SALSA: Stochastic approach for link structure analysis. *ACM Transactions on Information Systems,* 2001, **19**(2), 131-60.

24. Chakrabarti, S., *et al.* Hypersearching the web. *Sci Am,* 1999.

25. Papadimitriou, C.H.; Tamaki H.; Raghavan, P.; & Vempala S. Latent semantic indexing: a probabilistic analysis. *In* Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 1998.

26. Ejiogu, L. Software engineering with formal metrics, QED Publishing, 1991.

27. Roche, J.M. Software metrics and measurement principle. ACM *Software Engineering Notes,* 1994, **19**(1), 76-85.

28. Basili, V.R., and Weiss, D.M. A methodology for collecting valid software engineering data. *IEEE Trans. Software Engineering,* 1984, **SE-10**, 728-38.

**Contributors:** **Sh. Rakesh Kumar**, Dept. of Computer Sc. and Applications, Kurukshetra University, Kurukshetra. Tele: 01744-294145, e-mail: rsagwal@rediffmail.com

**Sh. P.K.Suri,** Dept. of Computer Sc. and Applications, Kurukshetra University Kurukshetra. Tele: 01744-239595, e-mail: pksuritf25@rediffmail.com

**Sh. R.K.Chauhan**, Dept. of Computer Sc. and Applications, Kurukshetra University Kurukshetra. Tele: 01744-239588, e-mail: rkckuk@yahoo.com