Precision and Relative Recall of Search Engines: A Comparative Study of Google and Yahoo

B.T. Sampath Kumar
J.N. Prakash
Kuvempu University

Abstract

This paper compared the retrieval effectiveness of the Google and Yahoo. Both precision and relative recall were considered for evaluating the effectiveness of the search engines. Queries using concepts in the field of library and information science were tested and were divided into one-word queries, simple multi-word queries and complex multi-word queries. Results of the study showed that the precision of Google was high for simple multi-word queries (0.97) and Yahoo had comparatively high precision for complex multi-word queries (0.76). Relative recall of Google was high for simple one-word queries (0.92) while Yahoo had higher relative recall for complex multi-word queries (0.61).

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

Introduction

The Web can be used as a quick and direct reference to get any type of information all over the world. However, information found on the Web needs to be filtered and may include voluminous misinformation or non relevant information. The Internet surfer may not be aware of many search engines to get information on a topic quickly and may use different search strategies. Finding useful information quickly on the Internet poses a challenge to both the ordinary users and the information professionals. Though the performance of currently available search engines has been improving continuously with

powerful search capabilities of various types, the lack of comprehensive coverage, the inability to predict the quality of retrieved results, and the absence of controlled vocabularies make it difficult for users to use search engines effectively. The use of the Internet as an information resource needs to be carefully evaluated as no traditional quality standards or control have been applied to the Web. Librarians need to be able to provide informative recommendations to their clientele regarding the selection of search engines and their effective search strategies. In this study, an attempt was made to assess the precision and relative recall of Google and Yahoo.

Search Engines and Search Queries

Two search engines, Google and Yahoo were considered to examine the precision and relative recall for some selected search queries during July 2007 to November 2007. 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 100 sites for evaluation.

A total of 15 queries in the library and information science discipline were selected for the study. All search queries were classified into three categories by the level of search complexity; simple one-word queries, simple multi-word queries and complex multi-word queries (see Appendix 1).

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 the present study, the search results which were retrieved by the Google and Yahoo 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:

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

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. Advanced search options were used for retrieving sites. Only English pages were searched for each search query since the web pages in other languages would be difficult to assess for relevancy. It was specified that the search query must appear in the 'title of the web page'. Since the number of search results retrieved was large, only the first 100 sites were selected for analysis.

Precision of Google for Simple One-word Queries

Table 1 showed that 30.6% of the sites retrieved by Google were less relevant followed by links (29%) and irrelevant sites (22.2%). It was also observed that 14.2% sites were more relevant and only a small percentage of the sites (4%) "can't be accessed". The precision of the Google was calculated using the above formula. The overall precision of the Google was 0.73. In the case of search query 1.5 and 1.1 the precision was 0.82 and 0.8 respectively. The lowest precision was for search query 1.2 (0.65).

Table 1: Precision of Google for Simple One-word Queries

Search Query	Total no. of sites retrieved	No. of sites evaluated	More relevant	Less relevant	Irrelevant	Links	Can't be accessed	Precision
Q.1.1	81,100,000	100	14	26	7	52	1	0.80
Q.1.2	411,000,000	100	14	28	34	18	6	0.65
Q.1.3	279,000,000	100	9	29	22	40	0	0.67
Q.1.4	13,600,000	100	20	28	30	11	11	0.73
Q.1.5	366,000,000	100	14	42	18	24	2	0.82
Total	1,150,700,000	500	71	153	111	145	20	0.73
%			14.2	30.6	22.2	29.0	4.0	

Precision of Google for Simple Multi-word Queries

Table 2 illustrated the search results of Google for simple multi word queries. It is evident from the table that 41.6% of sites are less relevant while 25.4% of sites are more relevant. It is also observed that 18.8% and 9.2% of sites are irrelevant and links respectively. Only a few percent of sites (5%) "can't be accessed". The overall precision of the Google is 0.97 and the highest precision 1.45 is obtained for the search query 1 followed by search query 4 (1.06) and search query 3 (0.95) respectively.

Table 2: Precision of Google for Simple Multi-word Queries

Search Query	Total no. of sites retrieved	No. of sites evaluated	More relevant	Less relevant	Irrelevant	Links	Can't be accessed	Precision
Q.2.1	691,000	100	55	35	6	0	4	1.45
Q.2.2	24,200	100	12	35	33	12	8	0.65
Q.2.3	296,000	100	23	42	19	14	2	0.95
Q.2.4	83,300	100	26	49	12	10	3	1.06
Q.2.5	2,510	100	11	47	24	10	8	0.74
Total	1,097,010	500	127	208	94	46	25	0.97
%			25.4	41.6	18.8	9.2	5.0	

Precision of Google for Complex Multi-word Queries

Study also made an attempt to measure the relevancy of Google for complex multi word queries. In this case the option 'any where in the page' had been chosen since too few sites would be retrieved for the option 'only in the title of the page'. The data collected was presented in Table 3.

Table 3: Precision of Google for Complex Multi-word Queries

Search Query	Total no. of sites retrieved	No. of sites evaluated	More relevant	Less relevant	Irrelevant	Links	Can't be accessed	Precision
Q.3.1	499,000	100	12	17	26	37	8	0.59
Q.3.2	961,000	100	20	31	25	22	2	0.82
Q.3.3	1,520,000	100	25	31	13	27	4	0.94
Q.3.4	916,000	100	13	23	41	13	10	0.55
Q.3.5	1,040,000	100	9	45	38	3	5	0.64
Total	4,936,000	500	79	147	143	102	29	0.71
%			15.8	29.4	28.6	20.4	5.8	

As seen in Table 3, 29.4% sites were less relevant, 28.6% of the sites were irrelevant followed by links (20.4%). It was also observed that 15.8% of the sites were more relevant and only a small percentage of sites (5.8%) can't be accessed. The precision of Google for complex multi-word queries was found to be 0.71.

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.

Precision of Yahoo for Simple One-word Queries

The advanced search options were used for retrieving web pages in Yahoo. From Table 4, it can be seen that a total of 90,779,000 sites were retrieved from Yahoo and only 500 sites were selected for evaluation. The results of the study showed that 36.4% of sites were links followed by less relevant sites (27.8%). It was also observed that 20.8% of the sites were irrelevant and that only 13.4% of the sites were more relevant.

Table 4: Precision of Yahoo for Simple One-word Queries

Search Query	Total no. of sites retrieved	No. of sites evaluated	More relevant	Less relevant	Irrelevant	Links	Can't be accessed	Precision
Q.1.1	33,100,000	100	15	17	18	49	1	0.71
Q.1.2	31,200,000	100	10	26	30	32	2	0.62
Q.1.3	5,840,000	100	13	17	25	43	2	0.64
Q.1.4	139,000	100	18	48	9	23	2	0.95
Q.1.5	20,500,000	100	11	31	22	35	1	0.70
Total	90,779,000	500	67	139	104	182	8	0.72
%			13.7	27.8	20.8	36.4	1.6	

The highest precision (0.95) was for search query 1.4 and the least precision was for search query 1.2 (0.62) and the overall precision of Yahoo was 0.72.

Precision of Yahoo for Simple Multi-word Queries

From Table 5, it can be seen that 34.2% of sites were less relevant followed by irrelevant sites (29%) and more relevant sites (17.8%). It can also be seen that 11.4% of the sites were links and only 7.6% of sites were "can't be accessed". The precision of the Yahoo was 0.75. For search query 2.4 and 2.5 the precision was 0.91 and 0.79 respectively and the least precision was for search query 2.1(0.65).

Table 5: Precision of Yahoo for Simple Multi-word Queries

Search Query	Total no. of sites retrieved	No. of sites evaluated	More relevant	Less relevant	Irrelevant	Links	Can't be accessed	Precision
Q.2.1	459,000	100	13	37	39	4	7	0.65
Q.2.2	8,100	100	20	2	30	8	14	0.72
Q.2.3	328,000	100	15	23	25	35	2	0.70
Q.2.4	55,500	100	25	40	25	2	8	0.91
Q.2.5	741	100	16	43	26	8	7	0.79
Total	851,341	500	89	171	145	57	38	0.75
			17.8	34.2	29	11.4	7.6	

Precision of Yahoo for Complex Multi-word Queries

For Yahoo, the search for complex multi-word queries results showed that 34.6% of sites were less relevant while 26.8% of sites were irrelevant. It was also observed that 17.8% and 16.6% of sites were links and more relevant respectively as shown in Table 6. The overall precision of the Yahoo was 0.76, and the highest precision was obtained for search query 3.2 (0.88) and the least precision was for search query 3.4 (0.51).

Table 6: Precision of Yahoo for Complex multi-word queries

Search Query	Total no. of sites retrieved	No. of sites evaluated	More relevant	Less relevant	Irrelevant	Links	Can't be accessed	Precision
Q.3.1	263,000	100	18	30	29	18	5	0.75
Q.3.2	422,000	100	18	41	17	22	2	0.88
Q.3.3	6,020,000	100	19	30	13	30	8	0.83
Q.3.4	432,000	100	12	22	54	10	2	0.51
Q.3.5	627,000	100	16	50	21	9	4	0.86
Total	7,764,000	500	83	173	134	89	21	0.76
%			16.6	34.6	26.8	17.8	4.2	

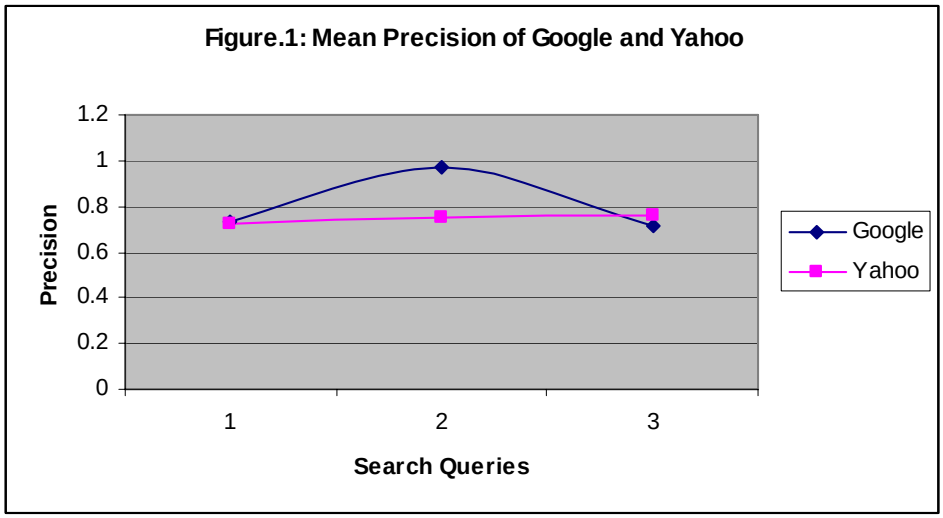
Mean Precision of Google and Yahoo

It is can be seen from Table 7 that the mean precision of Google was 0.80 and the mean precision of Yahoo was 0.74.

Table 7: Mean Precision of Google and Yahoo

Search Engine	Simple one-word Queries	Simple multi-word Queries	Complex multi-word Queries	Mean Precision
Google	0.73	0.97	0.71	0.80
Yahoo	0.72	0.75	0.76	0.74

Figure 1 showed the mean precision of Google and Yahoo for the three types of search queries.



Relative Recall of Google and Yahoo

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

$$\text{Relative recall} = \frac{\text{Total number of sites retrieved by a search engine}}{\text{Sum of sites retrieved by both Google and Yahoo}}$$

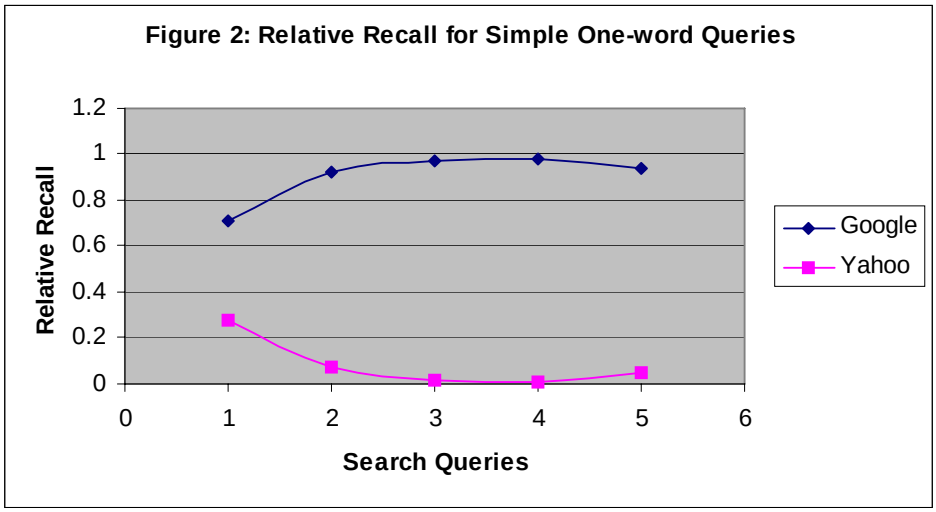
Relative Recall for Simple One-word Queries

The relative recall of the Google and Yahoo for simple one-word queries was calculated and presented in Table 8. The overall relative recall of the Google was 0.92 and Yahoo was 0.07.

Table 8: Relative Recall for Simple One-word Queries

Search Query	Google		Yahoo	
	Total no. of sites	Relative Recall	Total no. of sites	Relative Recall
Q.1.1	81,100,000	0.71	33,100,000	0.28
Q.1.2	411,000,000	0.92	31,200,000	0.07
Q.1.3	279,000,000	0.97	5,840,000	0.02
Q.1.4	13,600,000	0.98	139,000	0.01
Q.1.5	366,000,000	0.94	20,500,000	0.05
Total	1,150,700,000	0.92	90,779,000	0.07

Figure 2 showed the relative recall of Google and Yahoo for simple one-word search queries. In case of Google, the search query 1.4 had the highest relative recall value (0.98) followed by search query 1.3 (0.97) with the least relative recall for search query 1.1 (0.71). In case of Yahoo, the highest relative recall was for search query 1.1 (0.28) with the least relative recall for search query 1.4 (0.01).



Relative Recall for Simple Multi-word Queries

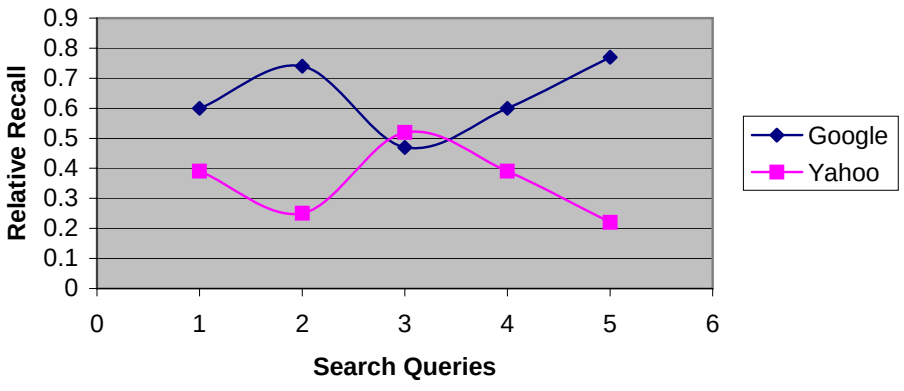
Table 9 illustrated that the relative recall of Google and Yahoo for all five simple multi-word queries. It was calculated that the overall relative recall of Google and Yahoo was 0.56 and 0.43 respectively.

Table 9: Relative Recall for Simple Multi-word Queries

Search Query	Google		Yahoo	
	Total no. of sites	Relative Recall	Total no. of sites	Relative Recall
Q.2.1	691,000	0.60	459,000	0.39
Q.2.2	24,200	0.74	8,100	0.25
Q.2.3	296,000	0.47	328,000	0.52
Q.2.4	83,300	0.60	55,500	0.39
Q.2.5	2,510	0.77	741	0.22
Total	1,097,010	0.56	851,341	0.43

The highest relative recall of Google was for search query 2.5 (0.77) while the highest relative recall of Yahoo was for search query 2.3 (0.52).

Figure 3: Relative Recall for Simple Multi-word Queries



Relative Recall for Simple Multi-word Queries

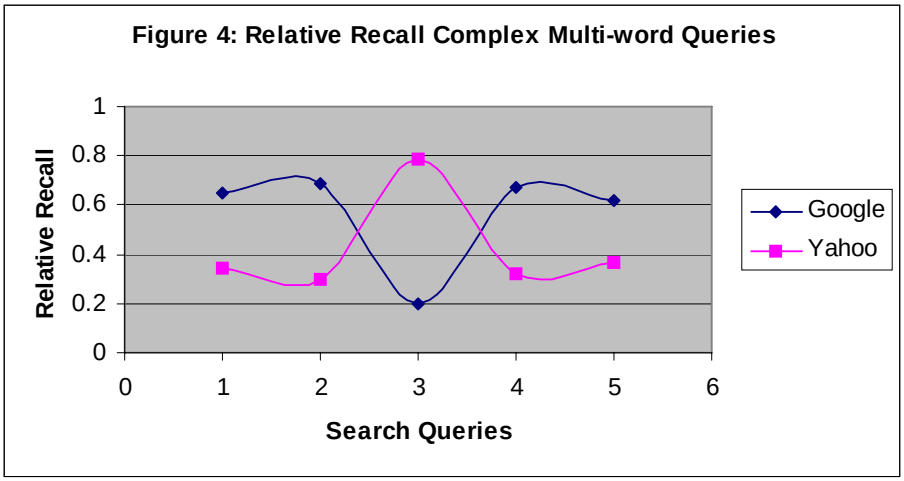
As seen in Table 10, the overall relative recall of Google and Yahoo for complex multi-word queries was 0.38 and 0.61 respectively.

Table 10: Relative Recall for Complex Multi-word Queries

Search Query	Google		Yahoo	
	Total no. of sites	Relative Recall	Total no. of sites	Relative Recall
Q.3.1	499,000	0.65	263,000	0.34
Q.3.2	961,000	0.69	422,000	0.30
Q.3.3	1,520,000	0.20	6,020,000	0.79
Q.3.4	916,000	0.67	432,000	0.32
Q.3.5	1,040,000	0.62	627,000	0.37
Total	4,936,000	0.38	7,764,000	0.61

In case of Google, the highest relative recall was for the search query 3.2 (0.69) followed by the search query 3.4 (0.67) with the least relative recall for search query 3.3 (0.20). In case of Yahoo, search query 3.3 received the highest relative recall (0.79) and the least relative recall was for search query 3.2 (0.30).

Figure 4: Relative Recall Complex Multi-word Queries



Mean Relative Recall of Google and Yahoo

The mean relative recall of Google and Yahoo was 0.62 and 0.37 respectively as seen in Table 11. Google had the highest precision (0.80) as well as the highest relative recall (0.62) as seen in Table 7.

Table 11: Mean Relative Recall of Google and Yahoo

Search Engine	Simple one-word Queries	Simple multi-word Queries	Complex multi-word Queries	Mean Relative Recall
Google	0.92	0.56	0.38	0.62
Yahoo	0.07	0.43	0.61	0.37

Conclusion

The World Wide Web with its short history has experienced significant changes. While the earlier search engines were established based on the traditional database and information retrieval methods, many other algorithms and methods have since been added to them to improve their results. The precision and relative recall value varies among the search engines depending on the database size. The gigantic size of the Web and vast variety of the users' needs and interests as well as the potential of the Web as a commercial market have brought about many changes and a great demand for the development of better search engines. The present study estimated the precision and relative recall of Google and Yahoo. The results of the study also showed that the precision of Google was high for simple multi-word queries and Yahoo had comparatively high precision for complex multi-word queries. Relative recall of Google was high for simple one-word queries while Yahoo had higher relative recall for complex multi-word queries. It was observed that Google and Yahoo showed diversity in their search capabilities, user interface and also in the quality of information. However these two search engines retrieved comparatively more irrelevant sites or links as compared to relevant sites. Google utilized the Web graph or link structure of the Web to become one of the most comprehensive and reliable search engines. This study provided evidence that the Google was able to give better search results with more precision and more relative recall as compared to Yahoo which would explain why it is the most widely used search engine for the Internet.

References

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

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.

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

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

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

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.

About the Authors

B.T. Sampath Kumar, Lecturer, Dept. of Library and Information Science, Kuvempu University, Karnataka, India
Email: sampathbt_2001@rediffmail.com
Web: www.freewebs.com/sampathkumar

J.N. Prakash, Dept. of Library and Information Science, Kuvempu University, Karnataka, India

Appendix 1: Search Queries

1. Simple one-word queries

Q 1.1: Encyclopedia

Q 1.2: Computer

Q 1.3: Multimedia

Q 1.4: Hypothesis

Q 1.5: Database

2. Simple multi-word queries

Q 2.1: Digital library

Q 2.2: Library automation

Q 2.3: Internet resources

Q 2.4: Intellectual property rights

Q 2.5: Search engine

3. Complex multi word queries

Q 3.1: Designing of library building

Q 3.2: Policies of collection development

Q 3.3: Evaluation of Web sites

Q 3.4: Internet and Web designing

Q 3.5 Evaluation of digital library