



Precision and Correlation among responses of Google Search Engine

Dr. Tauqeer Ahmad Usmani

Lecturer, Salalah College of Technology, Salalah, Sultanate of Oman

Abstract: This paper analyzed retrieval effectiveness in context of precision and correlation of Google search engine. The queries used are related to general and some computer science field. The queries are divided into three categories. The categories were simple one word, simple multi word and complex multi word queries. The results showed that the precision of Google was high for simple one word queries (0.76) and Bing had comparatively high precision of simple multi word queries (0.96) and Complex multi word queries (0.89).

Keywords: Google, Precision, search engines

I. Introduction

Web search is a key technology and one of the important purposes of the Web, since it is the primary way to access and read the content on the Web. Current standard Web which is not supporting Semantic Web technology, search is essentially based on a combination of textual keyword search with an importance ranking of the documents depending on the link structure of the Web. For this reason, it has many limitations, unwanted results, irrelevant results are coming in abundance. There are number of research activities towards more intelligent forms of searching and refinement of current searching technology on the Web, called Semantic search on the Web, or also Semantic Web search. With the outstanding development of information offered to end users through the Web, search engines approaches to play a major role. However, because of their common-purpose approach, it is always less unrelated that obtained result domain provide a group of useless pages. In this study, an analysis was made to assess the precision and correlation of Google.

II. Search Engines and Search Queries

Google was considered to examine the result of precision correlation for some selected queries. For getting relevant data from each search engine, the advance search features of search engines were used. The more sites were retrieved so it was decided to take first 100 sites for result evaluation. Total 15 queries of general and technical discipline were selected for study. The queries were classified into three categories according to search complexity: simple one-word queries, simple multi-word queries and complex multi- word queries (Appendix I)

(A) Precision of Search Engines

In a huge search results, 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. In the present study, the search results which were retrieved by the Google was categorized as 'more relevant', 'less relevant', 'irrelevant', 'links' and 'sites can't be accessed' on the basis of the following criteria:

- (i) If the web page is closely matched the query then it was categorize as 'more relevant' and given a score 2.
- (ii) If the web page is not closely related to the subject matter but consists of some relevant information related to the query then it was categorize as 'less relevant' and given a score 1.
- (iii) If the web page is not related to the search query then it was categorize as 'irrelevant' and given score 0.
- (iv) If web page consists of a whole series of links then it was categorize as 'links' and given a score 0.5 if the links are useful.
- (v) If the message like 'site can't be accessed' then it is categorize as 'site can't be accessed' if and only if same site tried later and gave same result then given a score 0.

(B) Precision of Google for Simple One-word Queries

Table 1 Showed that 39% of sites retrieved by Google search were less relevant followed by Irrelevant (34.2%) and more relevant sites (15.6%). It was observed that 10.6% sites are links and a small percentage of sites (0.6%) can't be accessed. The precision of Google was calculated using the above formula. The overall precision of the Google search was 0.76. In the case of search query 1.1 and 1.3 the precision was 0.86 and 0.78 respectively. The lowest precision was for query 1.5(0.69).

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
Q1.1	3,420,000,000	100	17	49	28	5	1	0.86
Q1.2	368,000,000	100	12	45	32	11	0	0.75
Q1.3	4,330,000,000	100	17	38	33	12	0	0.78
Q1.4	1,200,000,000	100	15	31	33	20	1	0.71
Q1.5	3,730,000,000	100	17	32	45	5	1	0.69
Total	13,048,000,000	500	78	195	171	53	3	0.76
%			15.6	39	34.2	10.6	0.6	

(C) Precision of Google for Simple Multi- Word Queries

Table 2 shows that the search results of Google for simple multi word queries. From the table it is clear that 33.2% sites are Irrelevant followed by 29.2% sites are less relevant while 18.4% sites are links. The percentage of site which is more relevant is 17%. A small number of percentage 2.2% sites are not accessed. The overall precision of Google is 0.72. The highest precision queries are 2.1(0.91) followed by query 2.2(0.85) and query 2.3(0.73).

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
Q2.1	32,400,000	100	24	28	18	29	1	0.91
Q2.2	241,000,000	100	18	44	28	10	0	0.85
Q2.3	578,000,000	100	17	33	37	11	2	0.73
Q2.4	41,400,000	100	17	30	44	7	2	0.68
Q2.5	21,100,000	100	9	11	39	35	6	0.47
Total	913,900,000	500	85	146	166	92	11	0.72
%			17	29.2	33.2	18.4	2.2	

(D) Precision of Google for Complex Multi- Word Queries

In the Table 3, The Google search for complex multi word queries was evaluated. 42.6% sites are Irrelevant followed by 27.6% was Less relevant. The percentage of more relevant sites is 20.8%. It was also observed that 6.8% sites are links and a small number of sites (2%) are cannot be accessed. The overall precision of Google search for complex multi word queries was found to be 0.73. The highest precision for the queries are 0.97 for queries 3.4 followed by precision 0.83 for queries 3.1.

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
Q3.1	659,000,000	100	25	31	37	4	3	0.83
Q3.2	24,800,000	100	19	29	46	6	0	0.70
Q3.3	244,000,000	100	15	28	39	17	1	0.67
Q3.4	365,000,000	100	29	37	28	3	2	0.97
Q3.5	32,600,000	100	16	13	63	4	4	0.47
Total	1,325,400,000	500	104	138	213	34	10	0.73
%			20.8	27.6	42.6	6.8	2	

(E) **Mean Precision of Google**

In Table 4, results show that the mean precision of Google was 0.74.

Table 4: Mean Precision of Google

Search engine	Simple one word query	Simple multi word query	Complex multi word queries	Mean Precision
Google	0.76	0.72	0.73	0.74

(F) **Correlation of Google Search**

Queries	Simple One word(A)	Simple Multi Word(B)	Complex Multi Word(C)	AB	AA	BB	BC	CC	AC
Q1.1	0.86	0.91	0.68	0.77	0.73	0.82	0.62	0.82	0.58
Q1.2	0.75	0.85	0.97	0.63	0.56	0.72	0.82	0.72	0.72
Q1.3	0.78	0.73	0.75	0.57	0.61	0.53	0.54	0.53	0.59
Q1.4	0.71	0.68	0.75	0.48	0.50	0.46	0.50	0.46	0.53
Q1.5	0.69	0.47	0.52	0.32	0.47	0.22	0.24	0.22	0.35
Total	3.78	3.62	3.66	2.77	2.87	2.74	2.73	2.74	2.77

$$\text{Correlation Coefficient } r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Correlation	r
AB	0.81
BC	0.91
AC	0.23

Figure 1: Correlation Between simple one word queries and simple multi word queries

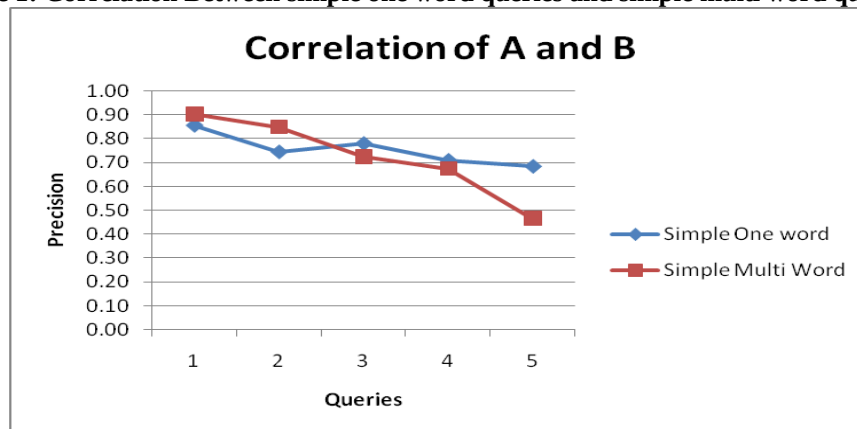


Figure 2: Correlation Between simple multi word queries and complex multi word queries

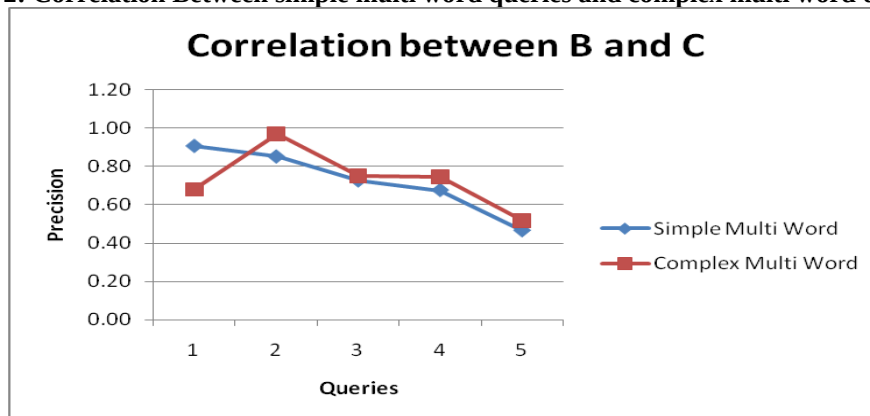
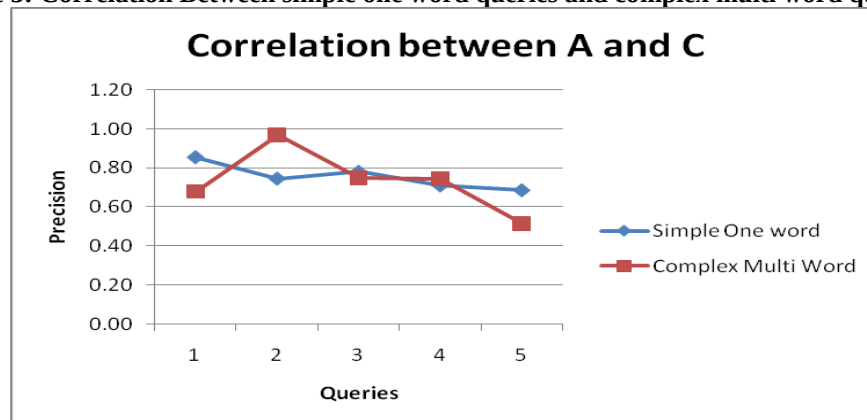


Figure 3: Correlation Between simple one word queries and complex multi word queries



Correlation between A and B is positive and near to 1 similarly correlation between B and C is also near to 1 and positive while correlation between A and C is near to 0 and positive.

III. CONCLUSION

The present study estimated the precision and relative recall of Google and Bing. The result of study showed that the precision of Google was high for simple one word queries, the precision of Bing was high for simple multi word queries and complex multi word queries both. Relative recall of Google was high for all simple one word queries, simple multi word queries and complex multi word queries. These two search engines gave more irrelevant results compare to relevant results. This comparison study showed that the Google gave better search results with more relative recall and precision for simple one word queries compare to Bing. Bing gave high precision for simple multi word queries and complex multi word queries. Over all precision was high for Bing but relative recall of Bing was less. This means that Google search was better for simple word while for complex words queries, Bing was better than Google. The correlation between simple one-word query and Complex one-word query of Google is weakly correlated so it should be improve to search all type of queries. The correlation between simple one-word query and Complex one-word query of Bing is negative and near to 0 so it should be improve to search all types of queries

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] Jiang Huiping, "Information Retrieval and the semantic web," *International Conference on Educational and Information Technology (ICEIT)*, Vol. 3, Pp. 461-463, 2010.
- 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.

APPENDIX I: SEARCH QUERIES

1. Simple one word queries

- Q 1.1: Program
- Q 1.2: Economics
- Q 1.3: History
- Q 1.4: Multimedia
- Q 1.5: Computer

2. Simple multi word queries

- Q 2.1 : Semantic Web
- Q 2.2: Search Engines
- Q 2.3: Operating System

Q 2.4: Office Automation

Q 2.5 Managerial Statistics

3. Complex multi word queries

- Q 3.1: Internet and its uses
- Q 3.2: Evaluation of computer world
- Q 3.3: System Analysis and Design
- Q 3.4: Policies and Planning of Indian Government
- Q 3.5: Evaluation of Indian History

AUTHORS PROFILE



Dr. Tauqeer Ahmad Usmani received PhD (Computer science) and Master (MCA) degrees in Computer Application from Kumaun University and Magadh University respectively. Presently working as Lecturer, Salalah College of Technology (Ministry of Manpower), Sultanate of Oman. Having 13 years of teaching experience in higher education including India and abroad. Published papers in International Journals. Members of various professional bodies of India and international repute. The research areas are Semantic Web, Intelligent Web.