Article can be accessed online at http://www.publishingindia.com

# Comparative Study of the Search Engines on the Basis of the Relevant Links on the First Web Page

**Nripendra Dwivedi*, Preeti Sirohi****

**Abstract**

Web search engines are the keys to the huge knowledge treasure of information and are used to extract query specific information from the complete database. Every search engine uses its own algorithm to rank the relevant links returned by the search engine. It is therefore essential for the user to understand the difference between the search engines in order to attain higher satisfaction level in terms of the result retrieved on the basis of users query. There is great assortment of search engines which offers various options to the web user. Thus it is significant to evaluate and compare the search engines in the quest of the best search engine which will provide the best result in the form of more number of informative links with the relevant result description on the first web page. The purpose of this paper is to compare four major search engines (Yahoo, Google, Ask, and Bing) for their retrieval efficiency on the query topics given from different fields (Computer Science, Physics, Chemistry, and Mathematics). The parameter which is taken to judge the best search engine providing no of relevant link on the first web page is "Number of relevant links returned for the query on the first page" and finding the best search result for the random query topics taken from the different fields (Computer Science, Physics, Mathematics). The research involves real life queries which are used frequently by researchers and academicians on regular basis and the result of the queries are analyzed. Based on the above results, tables are created and analysis is done for evaluating the performance of the four selected search engines. Performance is measured on the quality of the result returned on the first page of the search engine. The analysis of the result is done using the statistical tool-ANOVA (Analysis of Variance). By this analysis we evaluate the relative performance of these search engines.

**Keywords:** Search Engines, Google, Yahoo, World Wide Web

## 1. Introduction

Search engines have forever changed the way people access and discover knowledge, allowing information almost on any subject to be quickly and easily retrieved within seconds. Past history has witnessed that there was no way to search on the Internet. The only way to gather knowledge was through books or words of mouth. Searching for any information was a strenuous and difficult process. Invention of the search engine has brought revolution in the World Wide Web. With the increasing demand of the searching from the Internet and the advancement of the technology various search engines were developed. Today things have changed. "Do we really need books?" is the question that frequently comes in our mind. World Wide Web has come out as an explosion containing hundreds of millions of electronic collection of linked data for various fields. Search engines have completely changed the life of the major crowd. The World Wide Web is one of the largest publicly available databases of documents on the Internet.

* Associate Professor, Department of Computer Science, Institute of Management Sciences, Ghaziabad, Uttar Pradesh, India. E-mail: Nripendra.Dwivedi@imsgzb.com

** Assistant Professor, Department of Computer Science, Institute of Management Sciences, Ghaziabad, Uttar Pradesh, India. E-mail: Preeti.sirohi@imsgzb.com

Search engine uses automated software applications specially designed to retrieve information or user queries by travelling from links to links and pages to return the best result of the query. The information or query requested by the user can be of any particular object or subject thereby providing latest and historical information. Search engine has become one of the important applications of the Internet. Search engines use unique algorithms and formula to find the search result. Search engines design has made the search very easy, user-friendly and helpful. End user wants to find the result with minimum effort and load while attaining maximum result. A survey done by iProspect and Jupiter Research on the behaviour of the search engine users in January 2006 shows that 62% of the search engine users click on a search result within the first page of the result. Therefore, the first page of the result for the query should be ranked in accordance with the importance carrying maximum informative links related to the query topics. The journey of search engine started with the development of the first search engine.

This paper focuses on doing the analysis and comparative study among four popular search engines (Google, Yahoo, Ask, and Bing). The proposed research shall be able to define the best search engine in which we can find the maximum number of informative links of the first we search page. The end result will help the user to know the search engine which can give more relevant information in the form of relevant links returned to the user for the query topic. For the analysis and the comparative study, we have selected the keywords randomly from different subject areas such as Physics, Computer Science, Mathematics, and Chemistry. These keywords or the query topics are searched in different search engines. For analysis we need to record number of relevant links associated with the keywords on different search engines and make a count. We can do a comparative research using "ANOVA" and find the result.

The rest of the paper is organised as follows. The second section describes quality parameter on which search engines are evaluated. The third section describes the methodology which is followed in carrying out the research. The next section describes the results of the analysis, and the last section gives the conclusion.

## Quality Parameter

Number of relevant links returned for the query on the first web page corresponding to given a query topic is taken from the different fields.

## Methodology

To do a comparative study of the search engines, four specialised areas are taken. From each of these specialised areas random query topics are taken. These query topics are selected such that they cannot be further decomposed into sub-query topics. Table 1 shows various specialised areas and the query topics.

**Table 1: The Various Specialised Areas, and Query Topics Selected for Their Grade of Four Search Engines**

| *SPECIALISED AREAS* | *TOPICS* |
|---|---|
| Computer Science | Software Crises |
| | Unit Testing |
| | Software Quality Control |
| | Switch |
| | Enumerated Types |
| | Data Types |
| | Paging |
| | Last IN First Out |
| | Stack Overflow |
| | Concurrency Control |
| Physics | Force |
| | Acceleration |
| | Light |
| | Atom |
| | Photon |
| | Mass |
| | Electron |
| | Voltage |
| | Temperature |
| | Pressure |
| Chemistry | Molecule |
| | Boyle's Law |
| | Calorimeter |
| | Carbon |
| | Charles law |
| | Chromatography |
| | Dyes |
| | Entropy |
| | Gels |
| | Glycolysis |

| | |
|---|---|
| Mathematics | Integer |
| | Polygon |
| | Addition |
| | Union of sets |
| | Matrix |
| | Ordered Pair |
| | Multiplication |
| | Injective |
| | Binary Relations |
| | Venn Diagram |

A modified sampling method is used in this research using Query Based Sampling and the result displayed in the search engine displayed on the first taken. In this method, query sampling terms are carefully selected from ontology on four subject areas and sent to the search engine. The terms for retrieval are not repeated. The topics selected do not represent broader concepts but are decomposed in such a way that sub-query cannot be generated from the query topics.

Query topics selected above from different specialised areas are then used in the entire four selected search engine. Each query topic when used in the search engine returns number of links on the first web page. These links returned are then opened and checked for the desired information related with the query topics which are then counted by the experts of those specialised fields. Each link on the first web page of the result is opened and checked for the relevant information. The numbers of relevant links are then counted corresponding to the topics for all the four search engines. Table 2 shows number of relevant links corresponding to the query topics

With the analysis of all the four search engines on the query topics number of relevant links are counted. The study is to find whether the search engines are significantly different or not. To do the analysis we apply "Analysis of Variance" test. Table 2 shows the mean grade calculated for all the four search engines (Google, Yahoo, Ask, and Bing) on total of 40 query topics from specialised fields of Computer Science, Physics, Chemistry, and Maths. Table 3 shows the grand mean calculation by using formulae and also calculating between Column population variance.

**Table 2: Number of Links Corresponding to Query Topic for Each Search Engine**

| *Area* | *Topic* | *No. of Links* | | | |
|---|---|---|---|---|---|
| | | *Google* | *Ask* | *Yahoo* | *Bing* |
| Computer Science | Software Crises | 6 | 4 | 4 | 4 |
| | Unit Testing | 5 | 3 | 6 | 6 |
| | Software Quality Control | 4 | 4 | 5 | 3 |
| | Switch Case | 3 | 4 | 4 | 4 |
| | Enumerated Types | 9 | 7 | 9 | 8 |
| | Data Types | 6 | 6 | 8 | 7 |
| | Paging | 4 | 2 | 3 | 4 |
| | Last IN First Out | 3 | 3 | 2 | 2 |
| | Stack Overflow | 1 | 1 | 1 | 1 |
| | Concurrency Control | 9 | 6 | 9 | 9 |
| Physics | Force | 1 | 2 | 2 | 2 |
| | Acceleration | 7 | 3 | 7 | 8 |
| | Light | 3 | 2 | 8 | 5 |
| | Atom | 5 | 1 | 10 | 8 |
| | Photon | 2 | 1 | 8 | 4 |
| | Mass | 2 | 1 | 6 | 5 |
| | Electron | 3 | 3 | 7 | 8 |
| | Voltage | 5 | 1 | 4 | 4 |
| | Temperature | 1 | 1 | 5 | 5 |
| | Pressure | 4 | 2 | 4 | 3 |

| | | | | | |
|---|---|---|---|---|---|
| Chemistry | Molecule | 3 | 2 | 7 | 7 |
| | Boyle's Law | 2 | 2 | 3 | 3 |
| | Calorimetry | 6 | 3 | 4 | 6 |
| | Carbon | 1 | 1 | 4 | 4 |
| | Charles law | 6 | 3 | 3 | 3 |
| | Chromatography | 4 | 3 | 3 | 4 |
| | Dyes | 2 | 1 | 4 | 4 |
| | Entropy | 3 | 1 | 5 | 4 |
| | Gels | 1 | 0 | 2 | 2 |
| | Glycolysis | 2 | 1 | 8 | 6 |
| Maths | Integer | 4 | 4 | 6 | 5 |
| | Polygon | 4 | 3 | 7 | 6 |
| | Addition | 3 | 1 | 6 | 4 |
| | Union of sets | 6 | 4 | 7 | 4 |
| | Matrix | 1 | 0 | 0 | 7 |
| | Ordered Pair | 3 | 4 | 3 | 4 |
| | Multiplication | 2 | 1 | 1 | 2 |
| | Injective | 4 | 3 | 4 | 3 |
| | Binary Relations | 8 | 8 | 6 | 6 |
| | Venn Diagram | 3 | 2 | 1 | 6 |
| | | Mean(x1) = 5.0 | Mean(x2) = 4.0 | Mean(x3) = 5.1 | Mean(x4) = 4.8 |
| | GRAND MEAN = (40/160)*5 + (40/160)*4 + (40/160)*5.1 + (40/160)*4.8 = 4.725 | | | | |

**Table 3: Shows Population Variance-Between Column Variance for All Search Engines**

| *n (size of each sample i.e. number of search topics taken as query topic)* | *$\bar{a}$ = Mean of grade of all topics through every search engine* | *Grand mean of mean grade through all these three search engines* | *$y2 = (\bar{a} - \text{Grand mean})^2$* | *$n*y^2$* |
|---|---|---|---|---|
| 40(by Google) | 5(by Google) | 4.725 | $(0.275)^2 = 0756$ (of Google) | 40*0.0756 = 3.024 (of Google) |
| 40( by Yahoo) | 4( by Yahoo) | 4.725 | $(-725)^2 = 5256$ (of Yahoo) | 40*.5256 = 21.024 (of Yahoo) |
| 40( byAsk) | 5.1( byAsk) | 4.725 | $(0.375)^2 = 0.1406$ (of Ask) | 40*.1406 = 5.524 (of Ask) |
| 40( by Bing) | 4.8( by Bing) | 4.725 | $(0.075)^2 = 0.0056$ ( of Bing) | 40*0.0056 = 0.224( of Bing) |
| | | | | $\sum n_j y^2 = \sum n_j$ ($\bar{a}$ – Grand mean )$^2$ = 29.896 |
| Between column variance = $\sum n_j(\bar{a} - \text{Grand mean})^2/(k-1) = 29.896/(4-1) = 9.96533$ Where k = number of sample nj = size of jth sample | | | | |

Sample mean of all the search engines is almost similar. ANOVA is used for the analysis of variance for a single factor in our research. Null hypothesis is tested for the entire search engine. If $\mu 1$ is the mean of Google, $\mu 2$ is the mean for Yahoo, $\mu 3$ is the mean for Ask and $\mu 4$ is the mean for Bing then according to the null hypothesis $\mu 1 = \mu 2 = \mu 3 = \mu 4$. In analysis of variance, data values may also be scaled by multiplying or dividing by constants without affecting the value of the F ratio.

F ratio compares population variance among the sample size between column variance to the population variance within column variance. Table 4 calculates the population variance within column variance.

**Table 4: Population Variance-Within Column Variance Through All Search Engines**

| *No. of links for search result against Google Search Engine (Method 1) (x1)* | *(x1-mean (x1))2* | *No. of links for search result against Yahoo Search Engine) (Method-2) (x2)* | *(x2-mean (x2))$^2$* | *No. of links for search result against Ask Search Engine) (Method-3) (x3* | *(x3-mean(x3))$^2$* | *No. of links for search result against Bing Engine) (Method-4) (x4)* | *(x4-mean(x4))$^2$* |
|---|---|---|---|---|---|---|---|
| 6 | (6.0-5)2 = 1 | 4 | (4.0-4)2 = 0 | 4 | (4.0-5.1)2 = 1.21 | 4 | (4.0-4.8)2 = 0.64 |
| 5 | 0 | 6 | 1 | 3 | .81 | 6 | 1.44 |
| 4 | 1 | 5 | 0 | 4 | .01 | 3 | 3.24 |
| 3 | 4 | 4 | 0 | 4 | 1.21 | 4 | 0.64 |
| 9 | 16 | 9 | 9 | 7 | 15.21 | 8 | 10.24 |
| 6 | 1 | 8 | 4 | 6 | 8.41 | 7 | 4.84 |
| 4 | 1 | 3 | 4 | 2 | 4.41 | 4 | 0.64 |
| 3 | 1 | 2 | 1 | 3 | 9.61 | 2 | 7.84 |
| 1 | 16 | 1 | 9 | 1 | 16.81 | 1 | 14.44 |
| 9 | 16 | 9 | 4 | 6 | 15.21 | 9 | 17.64 |
| Mean(x1) = 50/10 = 5.0 | $\sum$(x1-mean(x1))$^2$ = 5.7 | Mean(x2) = 40/10 = 4.0 | $\sum$(x2-mean(x2))$^2$ = 3.2 | Mean(x3) = 51/10 = 5.1 | $\sum$(x3-mean(x3))$^2$ = 7.29 | Mean(x4) = 48/10 = 4.8 | $\sum$(x4-mean(x4))$^2$ = 6.16 |
| | Sample variance = s12 = ($\sum$(x1-mean(x1))$^2$)/(40 -1) = 5.7/39 = 0.1461 | | Sample variance = S2$^2$ = ($\sum$(x2-mean(x2))$^2$)/(40 -1) = 3.2/39 = 0.0820 | | Sample variance = S3$^2$ = ($\sum$(x3-mean(x3))$^2$)/(40 -1) = 7.29/39 = 0.1869 | | Sample variance = S4$^2$ = ($\sum$(x4-mean(x4))$^2$)/(40-1) = 6.16/39 = 0.1579 |
| Population variance(within column variance) = $\sigma$2 = $\sum((n_j-1)/(n_t-k))sj^2$ = ((40-1)/(160-4))*0.1461 + (39/156)*0.0820 + (39/156)*0.1869 + (39/156)* 0.1579 = 0.143225 | | | | | | | |
| Where $n_j$ = size of jth sample, $n_t$ = Total sample size, k = number of sample, sj = sample variance through jth search engine | | | | | | | |

**Table 5: Analysis of Variance (ANOVA) Table for F Ratio**

| *Source of variation* | *Sum of Squares* | *Degree of freedom* | *Mean Square* | *Variance Ratio(F) (Calculated)* |
|---|---|---|---|---|
| Between column | 29.896 | 3 | 9.96533 | 9.96533/0.143225 = 69.5781 |
| Within column | 22.343 | 156 | 0.143225 | |

Using Table 3 and Table 4, ANOVA Table is derived which is shown in Table 5.

ANOVA Table shows "Sum of squares", "Degree of freedom", "Mean Square" for their source of variations viz. between columns and within columns. F ratio is also shown in fifth column of this table.

At 5% level of significance the tabular value of F for (3,156) is 2.68. Since, the computed value of F = 69.5781 is greater than the tabular value of F = 2.68, therefore, we do not accept our null hypothesis. So we can say through the selected search engines viz. Google, Ask, Yahoo, Bing search results differ significantly.

## Key Findings

On the basis of F ratio obtained from analysis of variance, we can say that the four considered search engines statistically differ significantly. Therefore, it is concluded that in view of grades of quality parameter, statistically these search engines (Google, Ask, Yahoo, and Bing) cannot be considered equivalent.

## Conclusion

If we evaluate Google, Ask, Yahoo, and Bing search engines, on the parameter "Number of relevant links returned for the query on the first page", we find these four search engines differ significantly. Hence, from users' perspective, these search engines do not return same quality result on specified parameter. Thus, search engines are different from users' point of view on this parameter.

## References

1. Bharat, K. & Henzinger, M. R. (1998). Improved Algorithms for Topic Distillation in a Hyperlinked Environment.21st ACM SIGIR Conference.
2. Courtois, M. P., Baer, W. M. & Stark, M. (2005). *Cool tools for searching the Web: A performance evaluation*. Online, 19(6), 14-32.
3. Fazli, R. & Ayisigi, B. (2004). Automatic performance evaluation of web search engines. *Information Processing and Management*, January, 40(3), 495-514.
4. Haveliwala, T. H. (2002). *Topic Sensitive Page Rank*. In Proceedings of the Eleventh International World Wide Web Conference.
5. iProspect Search Engine User Behavior Study. (2006). A report byiProspect and Jupiter Research, January. www.iprospect.com
6. Marchionini, G. (1992). Interfaces for end-user information seeking. Journal of the American Society for Information Science, 43(2), 156-163.
7. Martin, P. C., William, M. B. & Marcella, S. (2004). Cool tools for searching the web: Performance Evaluation. Online, 19(6), 14-32.
8. Page, L., Brin, S., Motwani, R. & Winograd, T. (2002). The Page Rank Citation Ranking: Bringing order to the web. Stanford Digital Libraries Working Paper, 1998.