*Journal of Electrical Engineering*
*The Institution of Engineers, Bangladesh*
*Vol. EE 31, No. I & II, December 2004*

# Evaluating the Web Search Engines Quality : Point of User View

**Mollah Muhammad Towhidul Hoque**

Bangladesh Atomic Energy Commission
Institute of Computer Science, AERE, Savar, Dhaka
E-mail: towhidh@yahoo.com

**ABSTRACT**

**This paper presents an evaluating method of the Web search engines quality. A large number of experts and organizations in the world are engaged in development and updating the methods for evaluating the quality of search engines. Unfortunately, there is no standardization for the methods of evaluating the quality of search engines, both nationally and internationally. The evaluation of the quality of Web search engines not only enables informed consumer choice but also assists and encourages search engine operators to improve their standard of service. Since many search engines claim to be using novel techniques, effectiveness comparisons between these engines and systems employing published methods are potentially of interest to the Information Retrieval research community. Experimental results demonstrate that the offered method is less expensive and is of less labour-consuming procedure, than all earlier known methods.**

## 1. INTRODUCTION

The Web search engines are very important and useful resources, and are playing a major role in the information age. Search information in the web resources is carried out by means of search engines. The search is performed with respect to keywords. For given keywords, different search engines give substantially different results. There are many search engines. All of them are developing dynamically. Their consumer qualities continuously vary. Therefore, there is an increased need for evaluation of their quality to define and select the more effective search engine(s) [1,2,3].

It is known that some search engine operators, experts and organizations already perform both internal and external evaluations. All methods on evaluation of search engines which are known at present [4, 5, 6] are rather labour consuming and comparatively expensive. They envisage involvement of high-paid experts. Moreover, in these methods only some first few documents are browsed, instead of all resource, covered by queries. According to theses evaluations it is impossible to judge whether there are relevant documents in the included Internet resource, except for the viewed documents, and there can be many of them.

On the basis of the analysis of research and development works on the quality of search engines, we offer more simple method for evaluating the quality of information search engines based on the formal quantitative analysis of the answers against the inquiries. In the given method, all assessments are carried out without viewing the lists and the documents in inquiry answer. This method justifies the search engines whether the search it fulfils the rules of information searching language or not, and evaluates the quality according to the generalized vectorial criterion.

## 2. THEORETICAL EVALUATIONS

For each subject some minimum set of terms (key words) can be set up, which can precisely define this subject. The set of keywords for each subject is selected by expert order. This set of keywords $\mathbf{S} = \{S_1, S_2, ....., S_n\}$ can contain $n$ of words, where $S_i$ is the $i^{th}$ keyword.

During the investigation, the search should be carried out for each probable value of a binary vector, defining the constituent unit of the completed disjunctive normal form of the function of a constant unit considered as Boolean function of $n$ variables.

The first constituent unit $\bigcap_{\substack{i=1\\j=0}}^{n} \overline{S}_{ij}$ is substituted by

$$\bigcup_{\substack{i=1\\j=0}}^{n} S_{ij} .$$

If total keywords are designated by n, then the total number of different values of a binary vector (**K**) will be

$$\mathbf{K} = 2^n . \tag{1}$$

These logical combinations can be written as follows in a general view:

$$\bigcup_{i=1,\,j=0}^{n} S_{ij} = S_{10} \cup S_{20} \cup \cdots \cup S_{n0} , \quad \text{here all keywords are combined by the "OR" operators.} \tag{2}$$

$$\bigcap_{i=1,\,j=1}^{n} S_{ij} = S_{11} \cap \overline{S}_{21} \cap \cdots \cap \overline{S}_{n1} \quad \left.\begin{array}{l}\\ \\ \end{array}\right\} \text{these combinations correspond} \tag{3}$$

$$\vdots$$

$$\bigcap_{i=1,\,j=r}^{n} S_{ij} = S_{1r} \cap S_{2r} \cap \cdots \cap \overline{S}_{nr} \quad \left.\begin{array}{l}\\ \\ \end{array}\right\} 2^n - 2 \text{ constituent units of completed disjunctive normal form} \tag{4}$$

$$\bigcap_{i=1,\,j=z}^{n} S_{ij} = S_{1z} \cap S_{2z} \cap \cdots \cap S_{nz} , \quad \text{here all keywords are combined by the "AND" operators.} \tag{5}$$

where $S_{ij}$ – $i^{\text{th}}$ keyword in $j^{\text{th}}$ logical combination;
$n$ – number of keywords;
$r$ – represents the combination of keywords, in which only last one word is applied with inversion;
$z$ – conjunction number, in which all words are applied without inversions.

During the investigation, it is necessary to register the following indices of the answers of the search engines on each inquiry:

$N$ – number of documents retrieved on each logical combination of the keywords;

$t$ – time, elapsed on deriving the results during searching.

For processing and subsequent evaluation of the results of the experiments, the following characteristics are selected:

– conditional relevance coefficient ( $R_c^H$ );
– distortion coefficient ( $\rho$ );
– coverage coefficient ( $C_c^H$ );
– generalized criterion ( $K_{ok}$ );
– average searching time ( $t_{avg}$ ).

Each of the assessed coefficients is normalized with respect to the best of the same type result, which is reflected by the superscript *H*.

For a set of keywords, $S_l = \{S_{l_1}, S_{l_2}, ..., S_{l_n}\}$ (6)

appropriate set of weighting coefficients can be assigned

$$W_l = \{W_{l_1}, W_{l_2}, ..., W_{l_n}\} , \tag{7}$$

where $S_{l_i}$ – $i^{th}$ element of $S_l$ set of keywords;
$W_{l_i}$ – element of the set of appropriate weighting coefficients, $l$ – group of keywords or subjects number.

The *total weighting coefficient* for a logical combination of keywords ( $W_{j\Sigma}$ ), which is defined, as

$$W_{i\Sigma} = \sum_{i=1}^{n} P_i W_i , \tag{8}$$

where $W_i$ – weighting coefficient of $S_i$ keywords;

$$P_i = \begin{cases} 0, \text{ if } S_i \text{ is used in a logical combination with } \overline{S}_i ; \\ 1, \text{ if } S_i \text{ is used without inversion } (S_i ). \end{cases}$$

In a logical combination

$$\bigcup_{i=1,\,j=0}^{n} S_{ij} = S_{10} \cup S_{20} \cup \cdots \cup S_{n0} , \tag{9}$$

only the number of documents is fixed against the answer to such inquiry.

In the correctly operating searching system, *the number of documents* ( $N_0$ ), *obtained on the inquiry (9)* should coincide or should be close to *the sum of the number of documents* ( $N_1$ ), *obtained on all other* $2^n - 1$ *combinations*, i.e.,

$$N_{j=0} = \sum_{j=1}^{z} N_j . \tag{10}$$

Such equality gives some guarantee that it is possible to entrust the search engine(s) in searching on the given subjects and the above search engine(s) may fulfill the rules included in its information searching language.

We understand a distortion as mismatch of the amount of the documents found in a combination "OR" and the amount of documents obtained in inquiries on all other combinations of keywords. The distortion coefficient ( $\rho$ ) is defined by the expression

$$\rho = \frac{\left| \sum_{j=1}^{z} N_j - N_{j=0} \right|}{\max( \sum_{j=1}^{z} N_j , N_{j=0} )} . \qquad (11)$$

If it turns out that $\rho = 0$, it is possible to state that the system works accurately, and the evaluation of the search engines can be continued. If it appears that $\rho \neq 0$, then it is necessary to evaluate the extent of distortion of search result. If $|\rho| >> 0$, then it is necessary to admit that the given system is unsuitable for searching on the above subjects. This is the first restriction of the considered list of the search engines.

For an approximate assessment of the relation of *relevant* and *irrelevant* documents in the thematic area covered by inquiry, we have incorporated conditional relevance criterion ( $C_{rc}$ ). The *conditional relevance criterion* is an extent of an assessment of the logical formula of inquiry and, accordingly, the extent of relevance of the answer. The value of conditional relevance criterion depends on the accepted number of keywords and their (keywords) weighting coefficients.

Conditionally the *relevant documents* are declared as the documents, which are retrieved on group of keywords related with specific logical formula with total weighting coefficient, equal or more *conditional relevance criterion*, i.e.,

$$W_{j\Sigma} \geq C_{rc} . \qquad (12)$$

Otherwise,

$$W_{j\Sigma} < C_{rc} . \qquad (13)$$

The conditional relevance of the results of works on searching system is evaluated by conditional relevance coefficient ( $R_c$ ), which is defined by the following of expression

$$R_c = \frac{N_{rd}}{N_1} , \qquad (14)$$

where $N_{rd}$ – number of relevant documents obtained on inquiry from any logical combination of keywords except for the combination "OR".

The assessment on relevance coefficient is necessary but insufficient, since the same value of conditional relevance coefficient can be obtained by different search engines. Moreover, it is possible to obtain higher value of conditional relevance coefficient in a very small coverage of the search engines. Therefore, it is necessary to use the relative coverage coefficient ( $C_c$ ), which can be written in the following expression

$$C_c = \frac{N_{1_k}}{N_{1_{\max}}} , \qquad (15)$$

where $N_{1_k}$ – the number of $N_1$ documents of the search engines, $N_{1_{\max}}$ – the number of $N_1$ documents of the search engines, which is certainly greatest among them.

The evaluation of the search engines is the multi-criteria problem, in which the generalized criterion ( $K_{ok}$ ) of quality can be represented as a vector

$$K_{ok} = (R_c^H , C_c^H , \rho^H) , \qquad (16)$$

where $R_c^H$ – normalized conditional relevance coefficient, $C_c^H$ – normalized coverage coefficient, $\rho^H$ – normalized undistortion coefficient.

All component of the generalized criterion are normalized with respect to the best result in appropriate measurement. Having determined the component of the generalized criterion ( $K_{ok}$ ), it is possible to rank the search engines and thus to reveal those systems from them, which most effectively work in the considered topics.

The ranking and selection of the group of search engines can be fulfilled by *methods of decision making* in the research problems of operations. The considered task relates to the task of decision making under the conditions of determinacy. These tasks are characterized by presence of several criteria, on which it is necessary to compare the results.

In this investigation three criteria's are considered: the normalized conditional relevance coefficient ( $R_c^H$ ), normalized coverage coefficient ( $C_c^H$ ) and the normalized undistortion coefficient ( $\rho^H$ ).

Usually, these coefficients, according to their importance in evaluating the search engines are formed in the following order: $R_c^H$ , $C_c^H$ and $\rho^H$ . All search engines, participating in the procedure, are rated separately on each index, $R_c^H$ , $C_c^H$ and $\rho^H$ , in decreasing order of these indices. For each index, rated histograms are created, where median systems are marked on the histograms. The indices are arranged on preferences $R_c^H \succ C_c^H \succ \rho^H$ .

Let us introduce two restrictions. First - for further reviewing, let us keep only median systems and those that are on more left of the histograms, i.e. better median. Second - in the group of systems, starting from median and more to the left, we should keep only systems, whose index is not lower than certain boundary. In this example, the threshold value for the boundary has been chosen as 0.6. The systems located on histograms more to the right of the median system, whose indices differ from median value not more than 10 %, we should equate according to appropriate quality with the median system.

In the rating histogram of conditional relevance coefficient $R_c^H$ (in descending order $R_c^H$ ), let us mark the searching system falling into the median. Let us refer this system and all systems located more to the left of the median as competitive. Then, let us consider the rating histogram of coverage coefficient $C_c^H$ . Let us define the system falling into the median, and consider this system and all systems located to the left of the median. For further reviewing let us keep the systems, which are present in both results. Further, let us consider rating histogram of undistortion coefficient $\rho^H$ , we should mark the search engines falling into the median and this and all systems located more to the left of this system, we shall compare with the results of the previous selection. The systems, which are present in all three results, are competitive within the framework of the given topic and in the given period.

The time elapsed on obtaining the results by searching through the same searching pattern significantly varies for various systems, but being reduced to one obtained document; it is always less and makes a fraction of a second. It allows, by evaluating the search engines as a first approximation at fast channels, to neglect this feature.

## 3. EXPERIMENTAL EVALUATIONS

The effectiveness of eighteen search engines is evaluated using offered method [7].

From numerous experiments it has been observed that at searching by one, two or three words, most of the search engines provide foreseeable number of documents with substantial low relevance, and at usage of five words in a combination of "AND", the number of found documents is almost zero. Therefore, for all experiments it is recommended to carry out inquiries for evaluating the quality of search engines with usage of four keywords.

Since for the group of keywords, the semantic significances of separate words are different, therefore, during the evaluation of the quality of search engines, the assignment of the maximum weighting coefficient has been chosen according to the following order: the value of the most important keyword is 5, for the second keyword 4 and so on, in the decreasing order in according to the importance of these keywords.

Formally, for conditional relevant documents with four keywords, the criterion of conditional relevance was equal to six ($C_{rc}$ = 6). Such weighting coefficient guarantees that the result of found documents, at least, on a conjunction of two keywords, excluding the combination from the last two words with the least weighting coefficients 3 and 2, is *relevant*.

With reference to railway subjects “Railway and railway facilities”, let us reduce the results of documents searched with the use of eleven search engines under the following conditions (table 1):

**Table 1:** Subject, keywords, weights and search engines involved in experiments.

| Topic | $S_i$ | $W_i$ | Search engines |
|---|---|---|---|
| Railway and railway facilities | rail | 5 | AltaVista, AOL, Euroseek, Galaxy, InfoSeek, Lycos, MSN, NorthernLight, Rambler, WebCrawler, Yandex |
| | sleeper | 4 | |
| | bracket | 3 | |
| | lining | 2 | |

Another seven search engines also checked in this experiment: Aport, Excite, MetaCrawler, Russia On the Net, Yahoo!, Google and Netscape. The results of the experiment have shown that these seven search engines very strongly distort the results at documents search in particular railway topics. The search engines “Russia On the Net” did

not found any documents at all, on any inquiry made on railway topics since August 1999 till December 2001. Therefore, further experiments and evaluations for these systems were not carried out.

The search engines Netscape, which during testing in 2000 was rejected, as inadmissible distorted results of searching; at repeated testing in January, 2002 on the topic "Railway and railway facilities" has shown quite satisfactory results.

In table 2, a set consisting of sixteen Boolean combinations of inquiries in information searching language of the system Alta Vista and results of searches are shown. The results of data processing of searching experiments obtained by offered method on the topic "Railway and railway facilities" are shown in tables 3 and 4.

**Table 2:** Checklist of searching experiments.

| № | n=4 | Variant of inquiry | $W_{j\Sigma}$ | $N_j$ | $t_j$ |
|---|---|---|---|---|---|
| 0 | 0000 | rail OR sleeper OR bracket OR lining | | 2527972 | 6 |
| 1 | 0001 | lining AND NOT rail AND NOT sleeper AND NOT bracket | 2 | 441914 | 5 |
| 2 | 0010 | bracket AND NOT rail AND NOT sleeper AND NOT lining | 3 | 439975 | 7 |
| 3 | 0011 | bracket AND lining AND NOT rail AND NOT sleeper | 5 | 2456 | 6 |
| 4 | 0100 | sleeper AND NOT rail AND NOT bracket AND NOT lining | 4 | 125614 | 5 |
| 5 | 0101 | sleeper AND lining AND NOT rail AND NOT bracket | 6 | 937 | 5 |
| 6 | 0110 | sleeper AND bracket AND NOT rail AND NOT lining | 7 | 630 | 5 |
| 7 | 0111 | sleeper AND bracket AND lining AND NOT rail | 9 | 27 | 6 |
| 8 | 1000 | rail AND NOT sleeper AND NOT bracket AND NOT lining | 5 | 1408835 | 6 |
| 9 | 1001 | rail AND NOT sleeper AND NOT bracket AND lining | 7 | 8305 | 6 |
| 10 | 1010 | rail AND NOT sleeper AND bracket AND NOT lining | 8 | 11209 | 6 |
| 11 | 1011 | rail AND NOT sleeper AND bracket AND lining | 10 | 417 | 6 |
| 12 | 1100 | rail AND sleeper AND NOT bracket AND NOT lining | 9 | 7512 | 5 |
| 13 | 1101 | rail AND sleeper AND NOT bracket AND lining | 11 | 224 | 5 |
| 14 | 1110 | rail AND sleeper AND bracket AND NOT lining | 12 | 110 | 5 |
| 15 | 1111 | rail AND sleeper AND bracket AND lining | 14 | 15 | 5 |

**Table 3:** The results of evaluating the searching of documents listed on conditional relevance.

| Search engines | Total weighting coefficient | | | | | | | | | | | | $N_1$ | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $N_{rd}$ | | | | $N_{nrd}$ | | | | | | | | | | |
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | | $N_{rd}$ | $N_{nrd}$ |
| Altavista | 441914 | 439975 | 125614 | 1411291 | 937 | 8935 | 11209 | 7539 | 417 | 224 | 110 | 15 | 2418794 | 29386 | 2448180 |

**Table 4:** The results of 11th search engines obtained by offered method.

| **Search engines** | $R_c^H$ | $C_c^H$ | $\rho^H$ |
|---|---|---|---|
| AltaVista | 0.4088 | 0.6702 | 0.9568 |
| AOL | 1 | 0.0509 | 0.8008 |
| Euroseek | 0.4190 | 0.0901 | 1 |
| Galaxy | 0.3694 | 0.0003 | 0.9963 |
| InfoSeek | 0.3597 | 0.1338 | 0.3810 |
| Lycos | 0.6498 | 1 | 0.9610 |
| MSN | 0.6185 | 0.1619 | 0.9684 |
| NorthernLight | 0.8948 | 0.9289 | 1 |
| Rambler | 0.7232 | 0.0026 | 1 |
| WebCrawler | 0.6823 | 0.0055 | 0.2076 |
| Yandex | 0.7737 | 0.2334 | 0.9998 |

The results of the searching experiments and evaluations of eleven search engines within the framework of the topic "Railway and railway facilities" are shown in figures 1, 2 and 3.

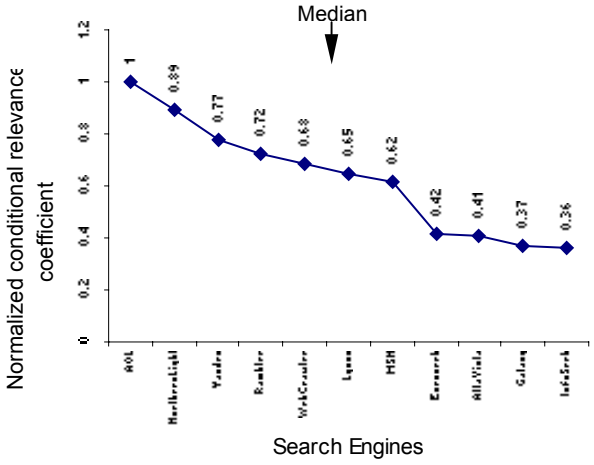


**Fig. 1**: Normalized conditional relevance coefficient.

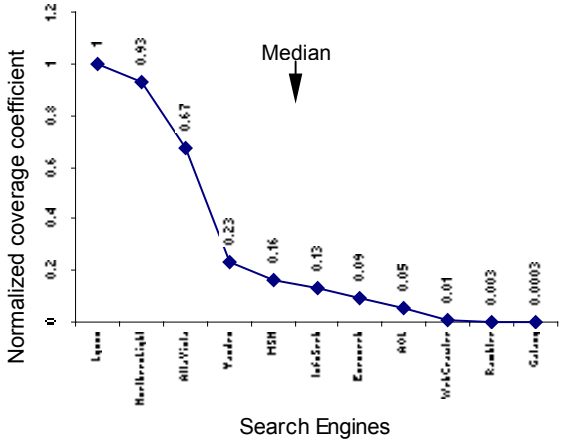


**Fig. 2**: Normalized coverage coefficient.

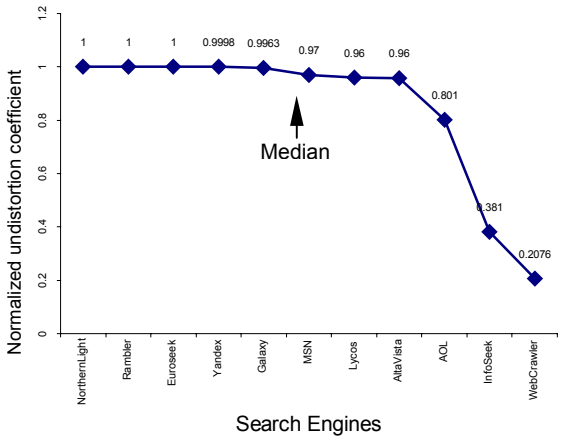


**Fig. 3**: Normalized undistortion coefficient.

## 4. CONCLUSIONS

A number of conclusions can be drawn from these experiments:

a) One of the main problems is that the search engines do not rank the relevance of results very well. Query-sensitive summaries can improve the efficiency of search. The query-sensitive summaries allowed users to perform relevance judgments more accurately and more rapidly.

b) Metasearch techniques can improve the efficiency of Web search by combining the results of multiple search engines, and by implementing functionality, which is not provided by the underlying engines (such as extracting queries term context and filtering dead links).

c) The experimental testing shown that with respect to specific topic, some search engines can be completely unsuitable due to distortion of results of searching.

d) The evaluation and rating of search engines shown that within the framework of one particular topic, one - two search engines have competitive characteristics. As a result of selection and ratings of search engines within the framework of the topic "Railway and railway facilities", only Northern Light and Lycos remain among recommended search engines.

e) Natural coefficients of conditional relevance ( $R_c^H$ ) for all search engines on all verifiable topics have appeared very low. The quantity of relevant documents in subject-matter area is from 0.4% to 10 %. This justifies the necessity to create search engines specialized on topics in a particular area. More comprehensive and more relevant results may be possible using such specialized search engine.

## ACKNOWLEDGEMENT

The author gratefully acknowledges Dr. Biryukov D.V., Associate Professor, Dept. of CSE, Moscow State University of Communication Means, for his assistance, encouragement and valuable suggestions during the work.

## REFERENCES

[1] Lawrence S. and Giles C. L., "Searching the World Wide Web", Science, Vol. 280, pp. 98-100, 1998.

[2] Lawrence S. and Giles C. L., "Searching the Web: General and Scientific Information Access", IEEE Communications, Vol. 37, No. 1, pp. 116-122, 1999.

[3] Lawrence S. and Giles C. L., "Accessibility of information on the web", Nature, Vol. 400, pp. 107-109, 1999.

[4] Hawking D., Craswell N., Bailey P. and Griffiths K., "Measuring the Quality of Public Search Engines", Search Engines Conference, Boston, USA, April 2000, Information Retrieval, Vol. 4, No. 1, pp. 33-59, 2001.

[5] Kharin N. and Ashmanov I., "Simplified method of comparative evaluation of technical effectiveness of searching engines in Internet", http://www.searchengines.ru/stories.php?story=01/12/10/2042905

[6] Nekrestyanov I., "Thematic-oriented methods of information searching", http://meta.math.spbu.ru/~igor/thesis/node1.html

[7] Mollah M.T.H., "Estimation of the Internet search engines quality applicable to efficiency of branch tasks solution (based on railway transport)", PhD Thesis, Moscow State University of Communication Means, 2002.