A Comprehensive Model for Web Search Evaluation

RASHID ALI* and M. M. SUFYAN BEG**

*Department of Computer Engineering, A. M. U., Aligarh 202002, India
**Department of Computer Science, University of California, Berkeley, CA 94720, USA

*rashidali@yahoo.com, **mmsbeg@cs.berkeley.edu

Abstract: – Web searching is such an activity that its importance can just not be ignored in the current scenario. But as there are a number of search engines available, there must be some procedure to evaluate them. In this paper, architecture of a comprehensive web search evaluation system is discussed. We use a number of objective evaluation techniques to supplement the subjective evaluation based on user feedback. The user feedback is obtained implicitly by watching the actions of user on search results in response to his query. These techniques are combined using Modified Shimura technique of Rank aggregation. The aggregated ranking is then compared with the original ranking given by the search engine. The correlation coefficient thus obtained is averaged for a set of queries. We show our experimental results pertaining to seven public search engines and fifteen queries.


1 Introduction

Web Searching is arguably the second most popular activity on Internet. A number of public search engines are available for this purpose. In the web based searching, a user queries the search engine for some query and gets the results in some order. Since different search engines employ different search algorithms and indexing techniques, the user gets different results in different order in response to the same query. So, naturally quality of search comes into picture. A general user is just interested in viewing the results on the first page and in that too, just the top few only. So, the search engine, which gives the results important to user in top few ones, should be voted as a better one. For this, we need to evaluate the search results.

In the present work, we propose architecture of a comprehensive web search evaluation system, which includes a number of evaluation techniques. Now, let the reason of using such a large number of evaluation techniques to be clarified. For subjective evaluation, we obtain the implicit feedback from the users by watching the actions of the user on search results presented before him in response to his query. For this, we assume the user to be an expert. Moreover, the user should be sincere and honest enough to select the most relevant documents in the proper order. With these assumptions, our subjective is poised to serve purpose solely. But, it is very much difficult to get the assumptions completely true. Every user may have some personal effect on his actions that in turn will affect his feedback. So, we feel to include objective techniques in the process to check the shortcomings of the subjective evaluation. But, selecting only one objective technique may favour a particular search engine and may be harsh to the rest. So, we use a number of objective techniques with the subjective evaluation to have an overall balanced judgment. These techniques give different rankings which are combined using Modified Shimura technique of Rank aggregation to get a "consensus" ranking. This "consensus" ranking when compared with the actual ranking provided by the search engine, gives a quantitative measure of search quality.

1.1 Related Work

In the past, some efforts have been made to evaluate search results from different search engines. In most of the cases, a uniform sample of the web pages is collected by carrying out random walks on the web. The size of indices is then measured using this uniform sample. A search engine having larger index size has higher probability to give good search results. In [1], [2] and [3], some attempts involving this is easily visible. In [4] also, the relative size and overlap of search engines is found but by using random
queries, which are generated from a lexicon of about 400,000 words, built from a broad crawl of roughly 300,000 documents in the Yahoo hierarchy. In [5] and [6], the search engines are compared using a standard query log like that of NEC research institute. In [7], a frozen 18.5 million page snapshots of part of the web is created for proper evaluation of web search systems. In [8], for two different sets of ad-hoc queries, the results from AltaVista, Google and InfoSeek are obtained. These results are automatically evaluated for relevance on the basis of vector space model. With the present effort, we aspire to get a comprehensive picture of web search evaluation.

2 Comprehensive Web Search Evaluation System

We propose a comprehensive web-search evaluation system as shown in fig. 1.

![Diagram](image)

**Figure 1: Comprehensive Search Quality Evaluation**

The user gives his query to the search engine and gets the search results ordered by search engine ranking say R_{SE}. Then, the user feedback is taken implicitly by watching the actions of the user on the search results and is characterized by a vector \( (V, T, P, S, B, E, C) \). Then, we get the four different ranking of the search results \( R_{VS}, R_{BS}, R_{PR} \) and \( R_{UF} \) using evaluations based on vector space model, Boolean Similarity measure, PageRank and user feedback respectively. These four rankings are then aggregated using Modified Shimura technique and we get a consensus ranking of the search results say \( R_{COMP} \). This ranking is then compared with search engine ranking and the correlation coefficient thus computed is a quantitative measure of web search quality of the search engine. The process may be repeated for different search engines and the search engines may be evaluated on the basis of the measure. The four evaluation techniques are discussed briefly in the following subsections.

2.1 Search Quality Measure using User Feedback Vector

We monitor the response of the user to the search results presented before him and characterize the feedback of the user by a vector \( (V, T, P, S, B, E, C) \) [9], which consists of the following.

(a) The sequence \( V \) in which the user visits the documents, \( V = (v_1, v_2, ..., v_N) \). If document \( i \) is the \( k \)th document visited by the user, then we set \( v_i = k \).

(b) The time \( t \) that a user spends examining the document \( i \). We denote the vector \( (t_1, t_2, ..., t_N) \) by \( T \).

(c) Whether or not the user prints the document \( i \). This is denoted by the Boolean \( p_i \). We denote the vector \( (p_1, p_2, ..., p_N) \) by \( P \).

(d) Whether or not the user saves the document \( i \). This is denoted by the Boolean \( s_i \). We denote the vector \( (s_1, s_2, ..., s_N) \) by \( S \).

(e) Whether or not the user book-marked the document \( i \). This is denoted by the Boolean \( b_i \). We denote the vector \( (b_1, b_2, ..., b_N) \) by \( B \).

(f) Whether or not the user e-mailed the document \( v \) to someone. This is denoted by the Boolean \( e_i \). We denote the vector \( (e_1, e_2, ..., e_N) \) by \( E \).

(g) The number of words that the user copied and pasted elsewhere. We denote the vector \( (c_1, c_2, ..., c_N) \) by \( C \).

When feedback recovery is complete, we compute the following weighted sum \( \sigma_j \) for each document \( j \) selected by the user.
There are a number of Boolean similarity measures [11] that can be used to compute the similarities of one document to another and documents to queries. We proposed a simplified Boolean similarity measure \( S^\circ \) based on \( Li \) Danzig measure \( S^\circ \) in [12]. If \( Q \) and \( C \) be the CDNF of the Boolean expressions \( Q \) and \( C \), the simplified individual similarity measure is defined as

\[
S^\circ (Q, C) = \begin{cases} 
0 & \text{if } T_Q^j \cap T_C^j = 0 \text{ or } \exists t \in T_Q^j, \neg t \in T_C^j \\
\frac{T_C^j - T_Q^j + T_Q^j - T_C^j + 1}{T_C^j - T_Q^j} & \text{otherwise}
\end{cases}
\]

Where \( \hat{Q}^i \) indicates the \( i \)th atomic descriptors of CDNF \( \hat{Q} \), \( \hat{C}^j \) indicates the \( j \)th compact atomic descriptor of CDNF \( \hat{C} \), \( T_Q^j \) and \( T_C^j \) are the set of descriptors in \( \hat{Q}^i \) and \( \hat{C}^j \) respectively. The similarity of two expressions, \( S^\circ \) is defined as the average value of the individual similarity measures (\( s^\circ \)). We obtain the ranking \( R_{BS} \) of the documents by sorting them in the decreasing order of their simplified Boolean similarity measures with the query.

### 2.3 Web Search Evaluation Using PageRank

Hyperlink structure of the web is found to be of great help in ranking of the documents. There are number of techniques available for hyperlink based ranking. In PageRank, which is adopted by Google[13], the quality measure of a page is its indegree. PageRanks form a probability distribution over web pages, so that the sum of all WebPages' PageRank will be one. PageRank can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web. Once we get the PageRanks of web pages corresponding to the documents in the search results, sorting the documents in the decreasing order of their PageRanks values would constitute the ranking \( R_{PR} \).

### 3 Rank Aggregation Using Modified Shimura Technique

Rank aggregation is the problem of generating a "consensus" ranking for a given set of rankings. We begin with the Shimura technique of fuzzy ordering [14], as it is well suited for non-transitive rankings, as in our case.
3.1 Shimura technique of fuzzy ordering
For variables $x_i$ and $x_j$ defined on universe $X$, a relativity function $f(x|x_i)$ is taken to be the membership of preferring $x_i$ over $x_j$. This function is given as

$$f(x|x_i) = \frac{f_x(x_i)}{\max(f_x(x_i), f_x(x_j))}$$

(3)

Where, $f_x(x_i)$ is the membership function of $x_i$ with respect to $x_j$ and $f_x(x_j)$ is the membership function of $x_j$ with respect to $x_i$. For $X = [x_1, x_2, \ldots, x_n]$, $f_x(x_i) = 1$. $C_i = \min_{j=1}^n f(x|x_i)$ is the membership ranking value for the $i^{th}$ variable. Now if a descending sort on $C_i$ (i=1 to n) is carried out, the sequence of $i$'s thus obtained would constitute the aggregated rank. For the lists $l_1$, $l_2$, $l_3$, $l_4$, from the N participating evaluation techniques, we can have

$$f_x(x_i) = \left[ \sum_{k=1}^{N} \left( \frac{\min(1, N) - \min(k(x_i) < k(x_j))}{N} \right) \right]$$

(4)

In our case $N=4$.

3.2 Modified Shimura technique
It is observed that classical Shimura technique gives worse performance in comparison to other rank aggregation techniques. We feel that the poor performance coming from the Shimura technique is primarily due to the employment of "min" function in finding $C_i = \min_{j=1}^n f(x|x_i)$. The "min" function results in many ties, when a descending order sort is applied on $C_i$. We, therefore, replace this "min" function by an OWA operator [15]. The OWA operators, in fact, provide a parameterised family of aggregation operators, which include many of the well-known operators such as the maximum, the minimum, the $k$-order statistics, the median and the arithmetic mean.

We will be using the relative fuzzy linguistic quantifier "at least half" with the pair $(a = 0.0, b = 0.5)$ for the purpose of finding the vector $C_i$ as follows-

$$C_i = \sum_j w_j z_j$$

(5)

Where $z_j$ is the $j^{th}$ largest element in the $i^{th}$ row of the matrix $f(x|x_i)$. $w_j$ is the weight of OWA based aggregation and is computed from the membership function $Q$ describing the quantifier. In the case of a relative quantifier, with $m$ criteria

$$w_j = Q(m) - Q((j-1)/m), j=0,1,2,\ldots, m$$

with $Q(0)=0$.

The membership function $Q$ of relative quantifier can be represented as

$$Q(r) = \begin{cases} 0 & \text{if } r < a \\ \frac{r-a}{b-a} & \text{if } b \leq r \leq a \\ 1 & \text{if } r > b \end{cases}$$

(6)

Now, as with the Shimura technique, if a descending sort on $C_i$ (i=1 to m) is carried out, the sequence of $i$'s thus obtained would constitute the aggregated rank.

We will be using the Modified Shimura technique for aggregating the four rankings $R_{UF}, R_{VS}, R_{BS}$, and $R_{PR}$. Let us say the aggregated ranking as $R_{COMP}$. Let the full list $R_{SE}$ be the sequence in which the documents were initially short-listed. Without loss of generality, it could be assumed that $R_{SE} = (1,2,3,\ldots,N_R)$, where $N_R$ is the total number of documents listed in the result. We compare the sequences and, and find Modified Spearman Rank Order Correlation Coefficient ($r'_s$)[12]. We repeat this procedure for a representative set of queries and take the average of $r'_s$. The resulting average value of $r'_s$ is the required measure of the search quality (SQM).

4 Experiments and Results
We experimented with the same 15 queries, which were used in [12] on seven popular search engines. For the sake of simplicity, we have obtained all our results for user feedback with the weights in equation (3) being $w_v=1$, $w_t=1$, $w_p=1$, $w_s=1$, $w_b=1$, $w_e=1$ and $w_c=1$. We get the user feedback vector \((V,T,P,S,B,E,C)\) and compute document weight($\sigma_i$) for each of the picked document. For vector space model, after the text pre-processing, normalized term vectors for the query and the documents are obtained. Then, dot product between the two is computed. For the Boolean Similarity measure based model, sets of descriptors in the compact atomic descriptors of query and in the compact atomic descriptors of all the documents picked up by user from the results of the query are obtained. Then, we compute the simplified Boolean similarity measure using equation (2). We also compute the PageRank corresponding to the urls of the documents. Sorting the document in decreasing order of these parameters give us four different rankings $R_{UF}, R_{VS}, R_{BS}$, and $R_{PR}$. These ranking are then aggregated using Modified Shimura technique with relative fuzzy linguistic quantifier "at least half" with the pair $(a = 0.0, b = 0.5)$. The aggregated
Modified Spearman Rank Order Correlation Coefficient obtained are also listed in the table 1. Modified Spearman Rank Order Correlation Coefficient ($r_s$) obtained using aggregated ranking ($R_{Comp}$) for the fifteen queries are given in Table 2. The results of Table 2 are pictorially represented in Figure 2. From Table 2 and Figure 2,
we observe that Google gives the best performance, followed by Yahoo, AltaVista, DirectHit, Excite, Lycos, and Hotbot, in that order.

5 Conclusions

We have tried to present the architecture of a comprehensive evaluation system for the public web search engines. The objective techniques are employed to supplement the subjective evaluation to realize a more accurate and efficient evaluation. We are aggregating the four different rankings of documents obtained from the four evaluation techniques using Modified Shimura Technique. Our results for 15 queries and 7 public web search engines show that Google gives the best performance, followed by Yahoo, AltaVista, DirectHit, Excite, Lycos, and Hotbot, in that order.

![Figure 2: Performance of Search Engines based on Aggregated Model](image)

References:


