Performance Evaluation of search engines via user efforts measures

Rajesh Kumar Goutam[1] and Sanjay K. Dwivedi[2]

[1] Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

[2] Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

Abstract

Many metrics exist to perform the task of search engine evaluation that are either looking for the experts judgments or believe in searchers decisions about the relevancy of the web documents. However, search logs can provide us information about how real users search. This paper explains, our attempts to incorporate the users searching behavior in formulation of user efforts centric evaluation metric. We also incorporate two dimensional users traversing approach in the ERR metric. After the formulation of the evaluation metric, authors judge its goodness and found that presented metric fulfills all the requirements that are needed for a metric to be mathematically accurate. The findings obtained from experiments, present a complete description for search engine evaluation procedure.

Keywords: *Information retrieval, Search engine performance, Search engine evaluation, Correlation based Ranking.*

1. Introduction

The size of World Wide Web is continuously expanding rapidly. This is because of world wide move to migrate the information from online resources. To retrieve some information from the web, search engines are essentially required. When these search engines receive the queries, return a list of documents which are ranked on the basis of their quality. Normally, search engine presents thousands of pages in response of a single query. Practically, this is not possible to access all these documents at all. With the help of our literature survey, we conclude that a normal searcher browses approximately first ten results so it is essential for a relevant document to get a place in top ten positions. Search engines prepare ranking with the help of their evaluation algorithm. Each search engine uses its own algorithm. As web is open to all, holds no restrictions to upload the documents, results expansion in web size. It seems impossible for a search engine to crawl all the web pages as quickly as these are getting uploaded. So it is a quick requirement to develop an evaluation metric that can evaluate the web pages in fastest way.

2. Background and Related work

Information is as vital as it was thousands years ago. A number of researchers contributed with their valuable and unforgettable efforts to convert the slow traditional sources of information to vast and fast resources of information. Now, there are various sources of information are available. From these resources of information, web has been accepted as very fast and primary resource of information. It has amazing power to satisfy its users with all kinds of information instantly.

Search engines are essentially required tools, used to migrate the information from the web. Various organizations have launched their search engines with different functionalities. Now, the situation is very critical as thousands of information retrieval systems are existing and each is claiming for its superiority and accurateness. So, the evaluation of the search engines performance is done to decide their efficiency and accurateness. Chu and Rosenthal [1] evaluated the capabilities of AltaVista, Excite and Lycos search engines on the basis of their performance. They used five criteria to perform the task of evaluation. These criteria were composition of web indexes (Coverage), Search Capability, Retrieval performance, output option (presentation) and users efforts. Although, the authors planned an effective strategy to perform the evaluation task but their evaluation process was slow as experts judgments were required. Suri [2] presented a search engine evaluation metric in which users traversing approach among the citations has been used. In this paper, the requirements of a good metric have also been discussed. O. Chapelle et. al. [3] used the cascade model and used a metric ERR (Expected Reciprocal Rank). In this metric the documents are judged for relevancy with probability of relevance. This approach seems to be some un-appropriate because same information can be irrelevant for a person which is relevant to some other person while both submit the same query. Cleverdon [4] suggested six criteria for search engine evaluation. These criteria are web coverage, dwell

time, recall, precision, presentation and user efforts. We explored the user efforts in the form of session duration, Ranked Precision and Clicks hits. We do not use the precision and recall as evaluation criteria because of some problems. Evaluations based on precision and recall is not difficult to compute but these measures are considered bit incomplete. Precision assumes that probability of randomly selected and retrieved web-pages becomes relevant. It also assumes that frequently search engines present the most relevant results in the top positions in ranking system. Precision computes the exactness of the retrieving relevant documents in the information retrieval process. It also computes how many documents are relevant in total retrieved web documents. It does not care if we are not retrieving all the relevant web-pages but we suffer if we are retrieving non-relevant web documents.

3. Metric Formulization

To measure the performance of the search engines, We have derived a metric named Ranked Precision (RP) which is based on two dimensional users traversing approach [6,7] among the retrieved citations. This metric returns a number between 0 and 1. In this metric, we divided all the web documents in four categories: Most Relevant, Partially relevant, somewhat relevant and completely relevant. Different relevance scores are assigned to different categories of documents.

Initially, we divided the relevance score for web documents in two parts S_z and W_j. where S_z is relevance score for sub-links and W_j is the score for root-links. For the calculation of total relevance score of sub-links, we shall sum up the total relevance scores of all sub-links. During the calculation of total sub-links score, we have presumed that users can visit up to m^{th} link. It is not necessary at all that each searcher will have to visit m^{th} link. If the searcher finds the satisfactory information in intermediate links then he/she can exit. One notable issue with the calculation of relevance score of sub-links, is its decrement in successive way, as the search length increases. In this way, total sub-links relevance numeric score for single root link is $\sum_{z=1}^{z=m} \frac{s_z}{z}$

Adding the relevance numeric score (w_j) for the root link to $\sum_{z=1}^{z=m} \frac{s_z}{z}$ we got the term

$$A_1 = (\sum_{z=1}^{z=m} \frac{s_z}{z} + w_j) \quad (1)$$

After inclusion of dead links b_j, we get the term A_2 as follows:

$$A_2 = (\sum_{z=1}^{z=m} \frac{s_z}{z} + w_j) * b_j \quad (2)$$

In the equation (2) b_j is the variable that holds only two numeric values 1 and 0. If the suggested citation (by search engines) is not alive then b_j holds the value 0 otherwise 1.As the searcher is viewing the root links one by one from the top of list to the bottom of list. So we shall multiplying the A_2 with the term $((n+1) - r_j)$. To concrete the concept, we suppose that there are n root links and the rank of the j^{th} document is r_j. The term $((n+1) - r_j)$ helps in reducing the relevance score of the root links gradually as the search length increases. So

$$A_3 = ((n+1) - r_j) * (\sum_{z=1}^{z=m} \frac{s_z}{z} + w_j) * b_j \quad (3)$$

A_3 is the relevance score of single root link (j^{th}) and its sub-links. Extending the equation (3) for the n number of root links. The equation (3) takes the form

$$A_4 = \sum_{j=1}^{j=n} [((n+1) - r_j) * (\sum_{z=1}^{z=m} \frac{s_z}{z} + w_j) * b_j] \quad (4)$$

We divide the equation (4) by the term $\frac{n(n+1)}{2}$ to find the Ranked Precision (RP). This term is used to calculate the best case and worst case of the RP metric.

$$RP = \frac{A_4}{\frac{n(n+1)}{2}} \quad (5)$$

Finally, Putting the value of the A_4 from the equation (4) into the equation (5) we found the metric

$$RP = \frac{\sum_{j=1}^{j=n} [\{((n+1) - r_j)\} * (\sum_{z=1}^{z=m} \frac{s_z}{z} + w_j) * b_j]}{\frac{n(n+1)}{2}} \quad (6)$$

Table 1. Score for root links.

Root links Relevance Score (w_j = 0.50 maximum)	
0.41-0.50	The most Relevant
0.31-0.40	Partly Relevant
0.10-0.30	Somewhat Relevant
0	Not Relevant at all

Table 2. Score for Sub-links.

Sub-links Relevance Score (s_z = 0.50 maximum)	
0.41-0.50	The most Relevant
0.31-0.40	Partly Relevant
0.10-0.30	Somewhat Relevant
0	Not Relevant at all

In our metric as shown in equation (6), the different numeric scores (S_z and W_j) are assigned by searchers to web pages that depend upon the quality of information published on it. In the table 1 and table 2, the ranges for relevancy about the documents are described.

Searchers normally prefer to search only few top citations to find the desired information. Silverstein et. al. [5] presented an study in which it was highlighted that approximately 85% of the searchers visit only top ten results. We considered this fact in our consideration and derived a metric in which user can fix the top ranges for the documents selection. In the equation (6), n is number of documents existing on top positions, are required to be examined for relevancy. Although, equation (6) is capable to evaluate the search engines and differentiate them but its working depends upon searchers judgments. Searcher can assign the highest relevance score to the documents which are irrelevant.

Cranfield style of evaluation has gained much popularity in past two decades. According to this method, the relevancy of the results decreases from top to bottom gradually. The principle of the cascade model considers this approach. The cascade model considers that the relevancy of retrieved documents becomes in descending order. It also considers that searcher stops the searching as he/she finds the results. Olivier Chapelle et. al. [3] used the cascade model and used an ERR (Expected Reciprocal Rank) metric for search engine evaluation. For this evaluation metric, the authors suggested some extensions to improve the results.

The Olivier chapelle et. al. [3] used the ERR as follows.

$$ERR = \sum_{r=1}^{n} \frac{R_r}{r} \qquad (7)$$

In the Equation (7), the R_r is defined as the probability or relevance and r is the rank of document.

In our method of search engine evaluation, we used the same metric (Equation 7) as it was used by authors (Olivier chapelle et. al., 2009). The main difference lies in computing the probability of relevance. In our method of search engine evaluation, R_r is calculated with the help of correlation between six parameters: session duration, dwell time, Ranked Precision (RP), Clicks Hits, user satisfaction with quality of results and user satisfaction for presentation of results.

Correlation CR1 is calculated between Session duration and Dwell time as these are positively correlated. In other words, variation in session duration time results the corresponding increment or decrement in the dwell time of WebPages. It is important to know here that how we organize the results according to session duration and dwell time. The document for which the session duration is minimum, is kept on the top in the furnished list while the documents for which the session duration is maximum is kept on the lowest position in the list. Conversely, the document is positioned at the top for which the dwell time is maximum while the document holds minimum dwell time is kept on the lowest position in the list. In both the cases the documents positions may change or identical.

Correlation CR2 is established between the Ranked precision (RP) and Clicks Hits because these two parameters are indirectly correlated. It could be easily concluded that Ranked Precision (RP) is directly dependent to search length [6]. Variation in the depth of relevant result will increase or decrease the corresponding clicks Hits. During the CR2 calculation, we form the first list in such a way that the maximum RP is positioned on the top thereafter the successive decrement begins. In the second list, maximum clicks hits corresponding a query are kept on the top in the list after that the successive decrement starts until the organization of all results gets completed.

Correlation CR3 is calculated between user's satisfaction with the presentation of results and user's satisfaction with the quality of results. We organize the numeric scores about user's satisfaction with the presentation and quality of results in descending order. Suppose, a search engine is presenting the low quality results in the top while the relevant results are positioned in bottom of the list. It is also possible that search engine can present ambiguous results corresponding particular query. In both the cases, the users satisfaction with the presentation will degrade.

Some other correlation pairs are still possible with the help of these six parameters. Session duration can be correlated with Ranked Precision (RP) as the increment in the search length results the corresponding increment in the session duration and vice versa. Similarly, the session duration can be correlated with clicks hits but it is very difficult to correlate session duration with the user's satisfaction with the quality and presentation of results because these two parameters are not dependent on session duration.

In our opinion, Dwell time cannot establish the correlation with the Ranked Precision (RP) because it is concerned with the time which is spent on a single document. It is not dependent on the search length. Correlation between the dwell time and clicks hits can be formed as expansion in the quality assessment time normally invite more clicks hits.

4. Metric Characteristics

For the validation of the evaluation task, the authors (P.K. Suri et. al. 2005) realized all the requirements for a good metric. We also validate our metric with same

requirements and found that our evaluation metric meets all the requirements that are needed to decide a metric as mathematically good.

(1) Empirically and intuitively persuasive: The metric results should rise and fall appropriately under various situations. Both metrics extended ERR and RP returns a value between 0 and 1. It can be easily seen that the value of RP becomes 1 when all the results retrieved are highly relevant and RP becomes 0 if all the retrieved results are irrelevant.

(2) Consistent and Objective: Both the metric RP and extended ERR are capable to yield relevant results. It is always essentially required that if a person derive some results with a metric, it should always be possible to derive same results in same situations by another person. For this purpose, we include three user efforts based signals such as session duration, dwell time and clicks hits so that the decision of a particular searcher could not affect the end results of the metric.

(3) Programming language independent: our metric of search engine evaluation is not derived for any particular language or particular platform so it can be programmed in any language for evaluation task.

(4) An effective mechanism for quality feedback: Number of clicks-hits help the search engines' developers to collect information that can be used by them to evaluate the effectiveness of their products and subsequently make easier development of a higher quality product.

(5) Possibility for extensions: extended ERR metric is extensible with some other search engines evaluation parameters such as query formulation time and web coverage as well.

As we discussed, our evaluation metric fulfills all the requirements that are necessary for the goodness of any metric. So on the basis of these six reasons, we can say that ERR is a good metric for search engine evaluation.

5. Experimental Results

We test the efficiency of our extended metric ERR with the 150 TREC pattern queries. We used *Mousotrom 5.0* software to record the session duration in minutes and dwell time in seconds. With this software we count the total number of click-hits, web documents are receiving. Besides of *Mousotrom 5.0*, we also used macromedia Dreamweaver CS5.5 software to validate the HTML web-pages. The relevance score for the web-pages, which is decided with users interactions with browsers is further integrated with searchers own judgments about the quality and presentation of results. This is done because the relevance judgments, collected automatically can produce the bias results as few web-sites incorporate the attractive advertisements on which few searchers make hits un-necessarily. To reduce the impact of this biasness, we combined the results derived automatically with searchers own judgments derived manually for quality of results and presentation of results. We apply our newly derived metric over a set of 150 TREC queries. The findings of the testing are shown in the table 3.

In the table 3, on the basis of six users' efforts measures and three correlation pairs, we computed the average correlation values for all three selected search engines. On the basis of these correlation values, all the selected search engines are compared. In our results, we found that 'Google' is most efficient search engine than rest two search engines. Our statistics decide 'MSN' as less significant search engine than 'Google' and 'Yahoo' systems. From the testing of results, we can conclude that approximately all the search engines consider all these six parameters because none of the correlation pair attains a value near to zero or zero. if a correlation pair attains a numeric value zero it means the positions organized for the queries for first list, are assigned positions in exactly reverse order in second list. For the first correlation pair CR1, approximately seventy three queries changed their positions in second list in 'Google' search engine. Similarly, in the correlation pair CR2 in 'Google' search engine, approximately seventy eight queries changed their positions from the first list and in the correlation pair CR3

Table 3. Search Engines comparison

Search Engine	Session Duration (Average)	Dwell Time (Average)	RP (Average)	Clicks-Hits (Average)	Score for Quality (Average)	Score for Presentation (Average)	CR1	CR2	CR3	Average* (ERR)
Google	13.12	224.09	0.67	21.19	0.67	0.61	0.51	0.49	0.45	0.48
Yahoo	16.86	194.15	0.61	15.67	0.55	0.57	0.44	0.42	0.44	0.43
MSN	18.78	137.96	0.52	11.56	0.44	0.53	0.47	0.37	0.35	0.39

Table 4. Ranges for skipped Citations.

Search Engine	Correlation Pair	Ranges for skipped Citations (0-30)	Ranges for skipped Citations (31-60)	Ranges for skipped Citations (61-90)	Ranges for skipped Citations (91-120)	Ranges for skipped Citations (121-150)	Total variations in queries' Positions
Google	CR1	43	11	10	7	2	73
	CR2	54	9	14	1	0	78
	CR3	38	12	16	17	0	83
Yahoo	CR1	25	23	26	10	6	90
	CR2	35	19	17	3	10	84
	CR3	42	33	10	5	8	98
MSN	CR1	11	23	26	35	16	111
	CR2	9	41	40	10	21	121
	CR3	20	12	8	19	29	88

eighty three queries changed their positions. The numeric value for the correlation is not dependent only on the queries' positions varying in both lists but also depends upon the number of citations being skipped. In other words, the correlation value is conversely proportional to the number of citations that are being skipped in query organization in second list in any pair. The findings in table 4, shows variation ranges in queries' positions in all the correlation pairs in all selected search engines. In the table 4, maximum small variations in queries' positions for 'Google' search engine are found in all the selected correlation pairs. Therefore, the average correlation value in table 3 for 'Google' search engine becomes large. For the 'Yahoo' search engine, comparatively some large variations in queries' positions are found than 'Google' so the average correlation value in the table 3 becomes small for 'Yahoo' than 'Google' search engine. In our results, extremely large variations are found in the queries' positions in the all the correlation pairs for the 'MSN' search engines so comparatively small correlation value is found for 'MSN' search engine than rest two search engines.

6. Conclusions

Most of the evaluation metrics for search engine evaluation are based upon unrealistic assumption that the user visits only the root links. However, the authors use the users' two dimensional searching approach and believe that searchers not only visit the root-links but also hits to sub-links to find the desired and satisfactory information. In this paper, we present the extended ERR metric that incorporates the six users action dependent ranking parameters to evaluate the search engines. Furthermore, we focused on the characteristics of newly formed metric. The authors validate their metric with the characteristics which are required to be judged for the goodness of the evaluation metric. Finally, we test the performance of our method for evaluation with 150 TREC pattern queries. On the basis of the average relevance score, we selected the 'Google' as most efficient search engine from a set of three search engines.

References

[1] Chu, H., Rosenthal, M., " Search engines for the world wide web: a comparative study and evaluation methodology",. Proceedings of the Annual Conference for the American Society for Information Science, 1996, pp. 127-135.

[2] P.K Suri, Rakesh Kumar, R.K Chauhan , "Search Engines Evaluation ", DESIDOC Bulletin of Information Technology, 2005, pp. 3-10.

[3] O. Chapelle, D. Metzler, Y. Zhang and P. Grinspan, "Expected Reciprocal Rank for Graded Relevance", In Proceeding of the 18th ACM conference on Information and knowledge management. 2009. New York, USA, pp. 621-630.

[4] Cleverdon, C.W., Mills, J., and Keen, E.M.., "An inquiry in testing of information retrieval systems", Cranfileld, U.K.: Aslib Cranfield Research Project, College of Aeronautics, 1966, pp. 230-232.

[5] Silverstein, C., Henzinger, M., Marais, J. & Moricz, M., "Analysis of a very large Alta Vista query log", Technical Report 1998-014, COMPAQ Systems Research Center, Palo Alto, Ca, USA, 1998.

[6] Rajesh Kumar Goutam and Sanjay K. Dwivedi, "Search Engines Evaluation using users efforts" In Proceedings of the 2nd International Conference on Computer and Communication Technology (ICCCT). 2011, Allahabad, India, pp. 589-594.

[7] Sanjay K. Dwivedi and Rajesh Kumar Goutam , "Evaluation of Search Engines using Search Length," In Proceedings of the International Conference of computer Modeling and Simulatione, 2011, Mumbai, India, pp. 502-505.

Sanjay K. Dwivedi Associate professor, Department of computer science at Babasaheb Bhimrao Ambedkar University, Lucknow 226025 (U.P.) India. His research interest is in Artificial Intelligence, web Mining, NLP and sense disambiguation etc. He has 16 years of experience of teaching and research and has handled/involved in some government funded research projects.

He has published a large number of research papers in reputed international journals and conferences.

Rajesh Kumar Goutam Research Scholar in Department of computer Science at Babasaheb Bhimrao Ambedkar University, Lucknow - 226025 (U.P.) India. His research interest is search engines and its performance evaluation, and web technology.