SIEM: An Integrated Evaluation Metric for Measuring Search Engine's Performance

Sojdeh Lotfipour
Department of Software Engineerig, Islamic
Azad University, Rasht Branch, Rasht, Iran

Fatemeh Ahmadi-Abkenari
Department of Software Enginnering &
Information System, Payame Nour University
(PNU), Iran

ABSTRACT

Search engines as Web-based information retrieval applications traverse a database consisting of millions of Web documents upon receiving a user issued query. In order to evaluate the accuracy and strength of a search engine regarding its robustness in finding relevant Web documents, a set of metrics have been proposed by researchers that each of them evaluates one aspect of a search engine's performance. One of the existing challenges in this area is the lack of one measurement that could state a search engines performance from different perspectives.

Some of developed metrics so far are general information retrieval evaluation measures that are not designed as specialized tools for search engine's evaluation. Some other metrics measure the system ability in finding accurate data while other metrics measures the speed of the system for performing the search process. In this paper different evaluation metrics such as precision, recall, f-measure, MAP, MRR, DCG and NDCG will be discussed. Then according to the conducted experiment and an analytical solution, a hybrid evaluation metric is proposed that based on it the overall strength of a search engine could be measured.

Keywords

Evaluation Metrics, Search Engine Evaluation, Web Information Retrieval.

1. INTRODUCTION

A search engine is a Web-based application that searches the use specified topics in the format of keywords, contents and existing data on WWW and provides the obtained results in the format of addresses of some saved locations. Some of search engines are limited to search the contents in a special Website. On the other hand, general-purpose search engines search the contents of Web throughout WWW and save an abstract of them in an indexed database. The users' queries will be searched within a pre constructed database of Web documents. Search engines have two main operations: providing answers through a list of relevant retrieved pages and result ranking [1], [2], [3], [11].

The process of evaluation search engines' performance is for measuring the accuracy of answers set, the number of accurate items within answers set and also the accuracy of ranking process. For evaluating the accuracy and strength of a search engine in finding relevant documents, a set of metrics are proposed by researches to analyze the operation of search engines that are categorized in two classes: set-oriented metrics and rank-oriented metrics. The first group of metrics has no attention to the accurate rank of each link in the answers set while the metrics in second group seek the accurate ranking strength of the tested search engine. Some of evaluation metrics are effectiveness and efficiency metrics. The effectiveness metrics measure the ability of search engines in finding accurate data while the efficiency metrics assess the speed in finding accurate data [16], [17], [19].

In this paper first, different evaluation metrics such as precision, recall, f-measure, MAP, DCG and NDCG will be discussed. Then the conducted experiment will be reviewed. Finally the proposed metric of this paper is introduced and elaborated [7], [8], [9].

2. EVALUATION METRICS

As mentioned above, the quality of search engines could be verified through employing evaluation metrics. A group of evaluation metrics for this purpose is barrowed from the evaluation metrics in the field of information retrieval. For this reason first, we review this group of metrics with the aim of identifying the aspect of a search engine they evaluate. In continue a number of widely used metrics are reviewed meanwhile table 1 provides a classification of evaluation metrics that could be implemented in search engine evaluation field by categorizing them into two groups: Precision/Recall-based and system metrics [4], [5], [6].

Table1. Evaluation metrics for search engines

Group	Metric
Precision/Recall Based and derivatives	Precision Recall F-measure F_1-measure Mean Average Precision (MAP)
System Based	Mean Reciprocal Rank (MRR) Discounted Cumulative Gain (DCG) Discounted Cumulative Gain (DCG) Normalized Discounted Cumulative Gain (NDCG)

2.1 Precision/Recall based Metrics

Precision and recall are metrics that are not only used in evaluating search engines' results but also in the evaluation of other information retrieval systems. These metrics measure the general ability of an information retrieval system in retrieving the relevant results according to user demand and produce a result between zero and one. In recall metric, positive rate of accuracy or sensitivity is equal to the proportion of number of relevant retrieved records to the general number of relevant records (retried records and non-retrieved relevant records). Precision metric is equivalent to the proportion of number of retrieved relevant records to the general number of retrieved records. In general, there is an inverse relationship between recall and precision. It means that an increase in recall results a decrease in precision. Figures 1 and 2 illustrate the concepts of recall and precision metrics [10], [12], [13], [16].

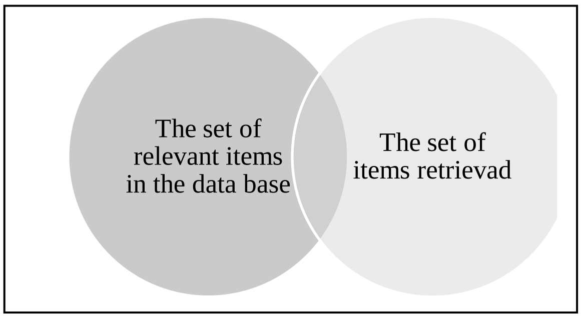
The set of relevant items in the data base
The set of items retrievad

Fig 1: Concept of recall metric

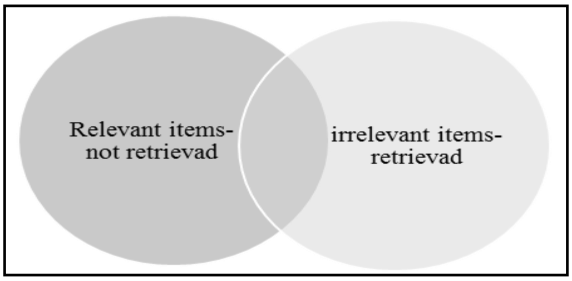

Fig 2: Concept of precision metric

2.2 F-Measure

F-measure metric is considered as a kind of effectiveness metrics carries a combination of recall and precision concepts and is in a form of a harmonic mean of these two metrics. The advantage of this metric is expressing the efficiency of information retrieval system in a numeric format. The reason for using the concept of harmonic mean instead of math mean is that the harmonic mean is emphasizing on small amounts while the math mean is affected by exceptionally large amounts. In F-measure metric, the β weight constant is using as shown in equation (1) below. In the case that this weight constant was equivalent to one, the formula will be changed to F_1-measure.

$$F\text{-measure} = (\beta^2+1)\ Precision*Recall\ /\ (\beta 2*Precision + Recall) \qquad (1)$$

2.3 MAP

The MAP metric is the average of precision from multiple queries. This metric measures the quality in all the levels of Recall. MAP metric as shown below in equation (2) is a ranking-oriented method and considers a damping factor, inconspicuous the importance of the found results in the answers list. This metric has been employed in separate research and projects such as TREC [14], [15], [18].

$$MAP = \frac{1}{relevant}\sum_{k=1}^{relevant}(Precision\ at\ rank\ of\ k\ relevant\ document) \quad (2)$$

2.4 MRR

MRR metric stands for Mean Reciprocal Rank suppose that the user is only interested in one relevant answer among the answers list. The nature of this metric is a simple process but with a strict look. MRR as is shown in equation (3) is equivalent to zero in the case of retrieving irrelevant results and otherwise, will be equivalent to 1/r [14], [15].

$$RR = \begin{cases} 0 & if\ no\ relevant\ results \\ \frac{1}{r} & else \end{cases} \qquad (3)$$

2.5 DCG

One of the most popular metrics is DGG which is based on the concept of cumulated gain. This metric stands for Discounted Cumulated Gain, has a ranking-oriented look and considers a discount function that decreases the grade of documents their ranking are increased to let user to test more relevant different resources which have wrongly labeled with a lower ranking. DCG metric as shown in equation (4) addresses the quality of retrieval independent from the quality of existing results [14], [15].

$$\sum_{r=1}^{R}\frac{(Gain\ result\ @r)}{\log_b(r+1)} \qquad (4)$$

2.6 NDCG

NDCG metric is resulted from DCG metric. This metric is a normalized DCG and has an unlimited range that is normalized through dividing DCG on the list of ideal results list (iDCG) which is existed in all results as shown in equation (5). NDCG runs well in most conditions and works through integrating the results list of multiple search engines [14], [15].

$$NDCG = \frac{DCG_r}{iDCG} \qquad (5)$$

3. PROBLEM STATEMENT

Evaluation measures for information retrieval systems and especially for search engines are diverse metrics that each of them determines the performance level of a tested system from a different perspective. Precision, recall and F_1-measure metrics are set-based metrics that determines the ability of a search engine to fetch relevant Web documents in the result set regardless of their rank position. Other discussed metrics have a rank-oriented approach and measures the quality of answer set regarding the position of relevant links not only their existence in answer set. There is a lack of an integrated metric that combines different features of the existing metrics into one unified solution in a way that could determines the quality of a search engine's performance through a robust approach. The aim of this paper is reaching to such an approach based on the conducted experiments presented in the next section.

4. EXPERIMENTAL RESULTS

In order to propose a unified evaluation metric for search engines first several queries are issued to various search engines and the engines' performance are evaluated through an exact measuring of aforementioned metrics. Due to the fact that in employing evaluation metrics, identifying the total of relevant documents is required, so the obtained results from *Google* search engine with 100 first links in the answers set of each query has been used in this test as a the basis. The search engines used in this experiment are *Ask*, *Bing*, *Excite*, *Lycos* and *DogPile*. The search queries are listed in table 2. In this article, the search engines such as *Look Smart* and *Alta Vista* are not used regarding their limited number of results in the queries of this experiment.

The crawls have been performed in two durations from 20th of June 2014 to 30th of July 2014 and from 30th of August 2014 to 6th of September 2014. From *Google* as the base search engine of this research, the first 100 pages have been used and for other five search engines of this study, the first 150 pages have been investigated. So for nine queries that are presented in this paper, the number of 675000 comparisons is carried out through the experiment of this research.

Table 2. Selected queries and search engines

Issued Queries	Search Engines
Q1: "Software Engineering Course"	*ASK* *Bing* *Excite* *Lycos* *DogPile*
Q2: "Wireless Sensor Network"	
Q3: "Social Network Analysis"	
Q4: "Real time System"	
Q5: "Object-oriented Database"	
Q6: "Virtual Memory"	
Q7: "Distributed Database"	
Q8: "Local Area Network"	
Q9: "ADSL Technology"	

Tables 3 to 11 show the fetched results obtained from measuring the performance of *ASK*, *Bing*, *Excite*, *Lycos* and *DogPile* through issuing the queries listed in table 2 against the metrics of recall, precision, F_1-measure, MAP, MRR, DCG and NDCG. Figures 3 to 11 represent the performance of search engines against the mentioned evaluation metrics. For better illustration of the chart, all the relevant amounts to DCG metric are multiplied in 0.1.

Table 3- Search engines evaluation (based on 1st query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.23	0.15	0.18	0.72	1	7.75	1
Bing	0.12	0.08	0.095	0.18	0.20	2.80	0.36
Excite	0.12	0.08	0.095	0.13	0.14	2.41	0.31
Lycos	0.17	0.11	0.13	0.22	1	4	0.52
Dog Pile	0.16	0.11	0.13	0.16	0.25	3.29	0.42

Table 4- Search engines evaluation (based on 2nd query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.34	0.23	0.27	0.65	1	8.83	1
Bing	0.24	0.16	0.19	0.39	1	6.46	0.73
Excite	0.09	0.06	0.07	0.24	0.17	2.31	0.26
Lycos	0.28	0.19	0.22	0.44	1	7.43	0.84
Dog Pile	0.15	0.10	0.12	0.24	0.25	3.42	0.39

Table 5- Search engines evaluation (based on 3rd query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.24	0.16	0.19	0.48	0.25	5.87	0.69
Bing	0.27	0.18	0.21	0.27	0.50	6.28	0.73
Excite	0.24	0.16	0.19	0.30	0.17	5.14	0.60
Lycos	0.31	0.21	0.25	0.48	1	8.59	1
Dog Pile	0.24	0.16	0.19	0.23	0.25	5.01	0.62

Table 6- Search engines evaluation (based on 4th query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.30	0.20	0.24	0.34	1	6.99	1
Bing	0.13	0.09	0.10	0.43	1	4.74	0.68
Excite	0.16	0.11	0.13	0.49	1	5.50	0.79
Lycos	0.12	0.08	0.095	0.26	0.50	3.64	0.52
Dog Pile	0.05	0.03	0.04	0.22	0.25	1.56	0.22

Table 7- Search engines evaluation (based on 5th query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.23	0.15	0.18	0.61	1	6.78	0.93
Bing	0.20	0.13	0.16	0.66	1	6.94	0.95
Excite	0.21	0.14	0.17	0.26	0.17	4.52	0.62
Lycos	0.23	0.15	0.18	0.58	1	7.31	1
Dog Pile	0.15	0.10	0.12	0.39	0.25	4.07	0.56

Table 8- Search engines evaluation (based on 6th query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.25	0.17	0.20	0.43	1	6.32	0.92
Bing	0.27	0.18	0.22	0.26	0.50	6.23	0.90
Excite	0.12	0.08	0.096	0.17	0.17	2.73	0.39
Lycos	0.28	0.19	0.22	0.37	1	6.88	1
Dog Pile	0.11	0.073	0.09	0.13	0.20	2.29	0.33

Table 9- Search engines evaluation (based on 7th query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.12	0.08	0.095	0.30	0.20	3.12	0.46
Bing	0.09	0.06	0.07	0.37	1	3.49	0.52
Excite	0.03	0.02	0.02	0.09	0.17	0.72	0.11
Lycos	0.25	0.17	0.20	0.31	1	6.75	1
Dog Pile	0.03	0.06	0.02	0.12	0.17	0.87	0.13

Table 10- Search engines evaluation (based on 8th query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.37	0.25	0.30	0.56	0.33	8.87	1
Bing	0.28	0.19	0.23	0.48	1	7.97	0.90
Excite	0.24	0.16	0.19	0.37	0.17	5.45	0.61
Lycos	0.29	0.19	0.23	0.65	1	8.87	0.99
Dog Pile	0.23	0.15	0.18	0.68	0.25	5.54	0.62

Table 11- Search engines evaluation (based on 9th query)

	Recall	Precision	F_1	MAP	MRR	DCG	NDCG
Ask	0.92	0.61	0.73	0.60	1	18.18	1
Bing	0.26	0.17	0.20	0.34	0.33	6.03	0.33
Excite	0.36	0.24	0.29	0.32	1	7.52	0.41
Lycos	0.20	0.13	0.16	0.38	0.50	5.56	0.31
Dog Pile	0.21	0.14	0.17	0.22	0.20	4.40	0.24

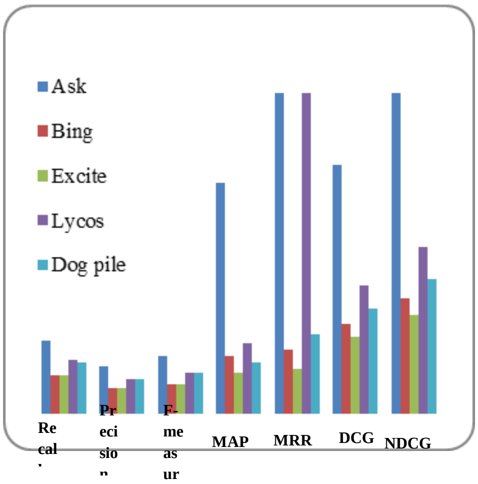

Fig 3: Evaluation metrics results based on the 1st query

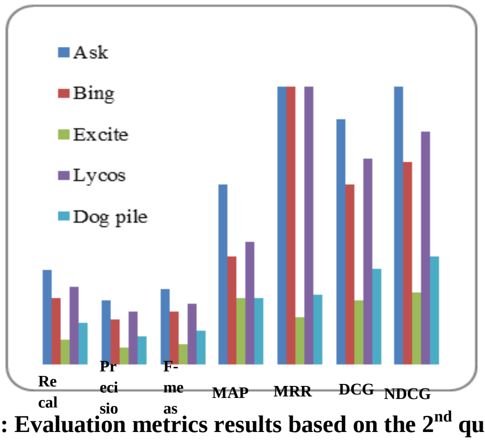

Fig 4: Evaluation metrics results based on the 2nd query

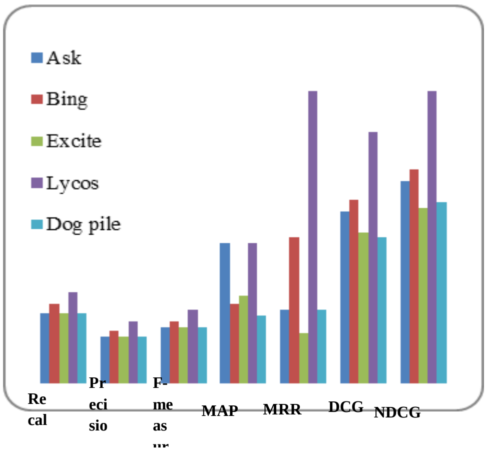

Fig 5: Evaluation metrics results based on the 3rd query

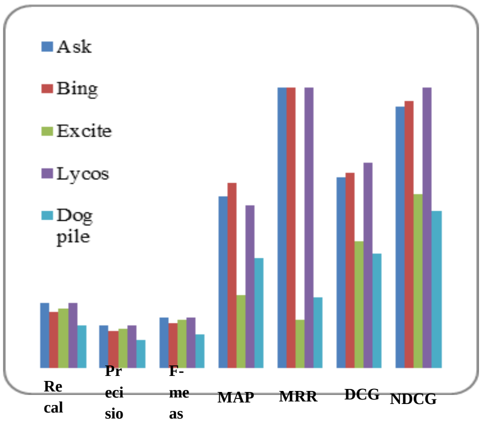

Fig 7: Evaluation metrics results based on the 5th query

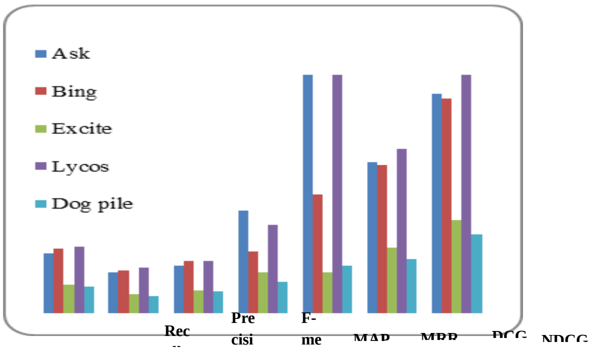

Fig 8: Evaluation metrics results based on the 6th query

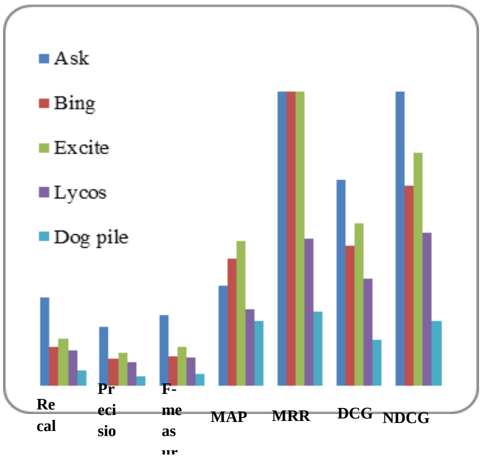

Fig 6: Evaluation metrics results based on the 4th query

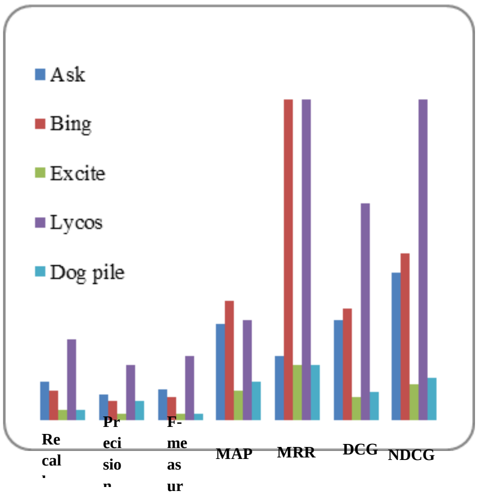

Fig 9: Evaluation metrics results based on the 7th query

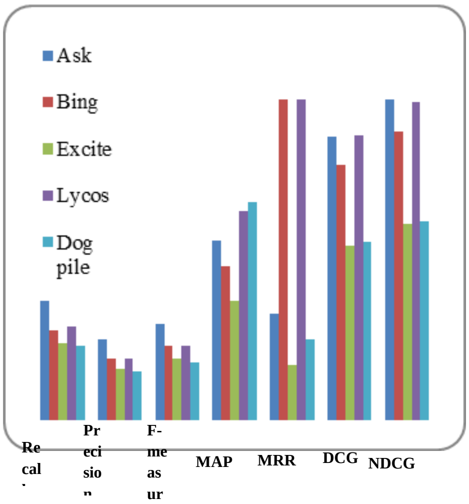

Fig 10: Evaluation metrics results based on the 8th query

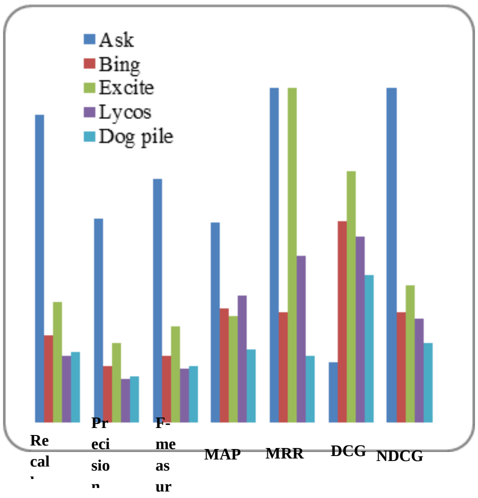

Fig 11: Evaluation metrics results based on the 9th query

From figures and tables 3 to 11 it could be concluded that *ASK* search engine had the best performance among other search engines for Q_1, Q_2, Q_4, and Q_8 and Q_9 queries, based on DCG and NDCG criteria. Also, *Lycos* search engine had the best performance in Q_3, Q_5, Q_6, Q_7 queries based on DCG and NDCG metrics. The results show that MRR metric won't explicitly express the difference between the overall performance of search engines because, the performance of *Ask*, *Bing* and *Lycos* search engines were the same for Q_2 query, the same for *ASK* and *Excite* in Q_9, the same for *Bing* and *Lycos* in Q_7 and Q_8. Other observations of the non-differentiation nature of MRR metric also exist in the tables above. In terms of MAP criteria, *Ask* search engine had the best operation in replying to Q_1, Q_2, Q_3, Q_6 and Q_9 queries, while based on MAP, the performance of *Bing* search engine is the best for replying Q_5 and Q_7 queries among other search engines. In terms of f-measure metric, *Ask* search engine had the best operation comparing others about Q_1, Q_2, Q_4, Q_5, Q_8 and Q_9 queries and about Q_3, Q_6 and Q_7 queries, *Lycos* search engine had the best performance.

Among all metrics to study search engines' performance in order to reach a metric in combination approach, regarding the same view of results obtained from DCG and NDCG metrics, the NDCG metric is chosen because of its normalized final result. Among precision, recall and f_1-measure criteria, f_1-measure is chosen, because of yielding a result that carries both precision and recall concepts. Considering the inefficiency of MRR metric in differentiation among search engines, it has no place in the integrated solution. Instead, MAP metric is selected for the combined result. In all sent queries to five search engines of this study, both NDCG and f_1-measure metrics have confirmed the priority of one search engine. The search engine, based on the stated numbers for 1st, 2nd, 4th, 8th and 9th queries, is *Ask* and for 3rd, 5th, 6th and 7th is *Lycos*. In this paper, the integrated solution is called *SIEM* stands for *Search engine Integrated Evaluation Metric* and it is calculated as shown in equation (6) below;

$$SIEM = [\alpha * NDCG + \beta * F_1\text{-}measure + \delta * MAP] / 3 \quad (6)$$

In equation (6), α, β and δ are set to one. The greater value of SIEM indicates the better performance of a search engine in compare to others. Table 12 shows the calculated SIEM values for nine queries in search engines of this research. Figure 12 depicts the performance of the five search engines of this experiment regarding the SIEM metric based on the average results for different queries. As figure 12 shows the performance of *ASK* is the best followed by *Lycos*, *Bing*, *Excite* and *Dogpile* respectively.

Table 12- SIEM values for monitored search engines

	ASK	*Bing*	*Excite*	*Lycos*	*DogPile*
Q1	0.63	0.21	0.53	0.29	0.24
Q2	0.64	0.44	0.19	0.50	0.25
Q3	0.45	0.40	0.36	0.58	0.35
Q4	0.53	0.40	0.47	0.29	0.16
Q5	0.57	0.59	0.35	0.59	0.36
Q6	0.52	0.46	0.22	0.53	0.18
Q7	0.28	0.32	0.07	0.50	0.09
Q8	0.62	0.54	0.39	0.62	0.49
Q9	0.78	0.29	0.34	0.28	0.21

For adapting equation (6) for future growth in considering other metrics, the formula is generalized as shown in equation (7):

$$SIEM = \frac{\sum_{i=1}^{n} \alpha_i * M_i}{n} \qquad (7)$$

In equation (7), n is the number of involved metrics in the integrated plan, M_i is any of the considered metrics and α_i is the considered coefficient for each metrics.

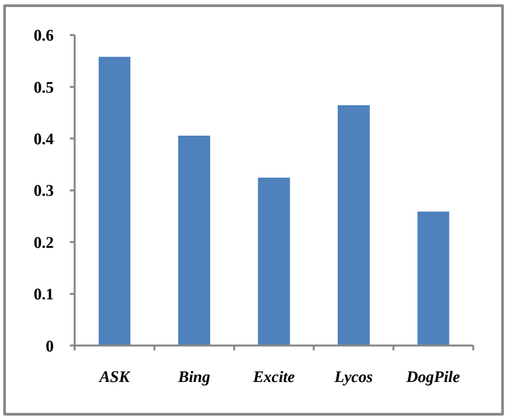

Fig 12: Evaluating the five search engines based on SIEM metric

5. CONCLUSION

As Lord Kelvin said, if we can’t measure something, we won’t be able to improve that. Therefore, evaluation of search engines reveal the weakness level of these Web-based applications regarding their accuracy in obtaining relevant Web documents and performing the ranking process of obtained links. Performance measurement of search engine could result in better designation of the specific modules which results in better overall efficiency of the application.

In this paper after a review on the most famous evaluation metrics, an experiment is conducted through which several queries have been issued to different search engines. Keeping the *Google* result set for each of the queries as the base line, the performance of each search engine is measured against the discussed metrics. The integrated solution is based on analyzing and combining the obtained results from various metrics. More generalization of the above-said integrated solution depends on extending the test through sending more queries to more search engines and also involving other evaluation metrics in the provided analytical-combined approach. The future work of this research aims at developing a stand-alone metric for measuring the performance of search engines’ result set independent of the discussed metrics.

6. REFERENCES

[1] Ahmadi-Abkenari, F., Selamat, A. 2012. “An Architecture for a Focused Trend Parallel Web Crawler with the Application of Clickstream Analysis”, International Journal of Information Sciences, Elsevier, Vol. 184, pp. 266-281.

[2] Ahmadi-Abkenari, F., and Selamat, A. 2013. “Advantages of Employing LogRank Web Page Importance Metric in Domain Specific Web Search Engines”. JDCTA: International Journal of Digital Content Technology and its Applications. Vol. 7, No. 9. pp. 425-432.

[3] Ahmadi-Abkenari, F., and Selamat, A. 2012. “LogRank: A Clickstream-based Web Page Importance Metric for Web Crawlers”. JDCTA: International Journal of Digital Content Technology and its Applications. Vol. 6, No.1. pp. 200-207.

[4] Clarke, C. L. A, Craswell, N. and Soboroff, I . 2009. “Overview of the TREC” 2009 Web Track.

[5] Cleverdon, C. W. and Keen, M. 1966. “Factors Determining the Performance of Indexing Systems”. Cranfield, England, Aslib Cranfield Research Project.

[6] Cooper, W.S. 1973. “On Selecting a Measure of Retrieval Effectiveness”. pp. 87-100.

[7] De Kunder, M. 2012. “The Size of the World Wide Web”. Retrieved from: Retrieved from http://www.worldwidewebsize.com/.

[8] Fawcelt, T. 2006."An Introduction to ROC Analysis". Pattern Recognition Letters 27 (8). pp. 861 – 874.

[9] Heinrich, H. 2012. “On Search Engine Evaluation Metrics”. Dusseldorf. pp. 3-192.

[10] Järvelin, K. and J. Kekäläinen . 2000. “IR Evaluation Methods for Retrieving Highly Relevant Documents”. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval .pp.41-48.

[11] Malekian, Ehsan. 1979. “Principles of Internet Engineering”. NAS Publications. pp. 481-486.

[12] Perruchet, P. and Peereman, R. 2004. "The Exploitation of Distributional Information in Syllable Processing". Journal of Neurolinguistics 17. pp. 97−119.

[13] Powers, D.M.W, 2011. "Evaluation: From Precision, Recall and F-measure to ROC Informedness, Markedness and Correlation". Machine Learning Technologies 2 (1). pp. 37–63.

[14] Rakesh Kumar, P.K.Suri & Chauhan, R.K. 2005. “Search Engines Evaluation”. Desidoc Bulletin of Information Technology.Vol.25, No.2. pp.3-10.

[15] Song, M., Wu, Y. 2009. “Handbook of Research on Text and Web Mining Technologies”. Information Science Reference. Vol. 2.

[16] Van Rijsbergen, C.J. 1979. “Retrieval effectiveness. In: Progress in communication science”. Vol.1. pp. 91-118.

[17] Vaughan. L. 2004. “New Measurements for Search Engine Evaluation Proposed and Tested”. 40(4). Information Processing and Management. pp. 677-691.

[18] Voorhees, E. M. 1999. “The TREC 8 Question Answering Track Report”. In Proceedings of the 8th Text Retrieval Conference (TREC 8) Gaithersburg. pp. 77-82.

[19] Wesley, A. 2008. “Evaluating Search Engines”. Chapter 8. pp. 1-40.