A Review on Web Search Engines' Automatic Evaluation Methods and How to Select the Evaluation Method

Masomeh Azimzadeh, Reza Badie, Mohammad Mehdi Esnaashari
Information Technology Research Group
Information and Telecommunication Research Center (ITRC)
Tehran, Iran
{azim_ma, rezabadie, esnaashari}@itrc.ac.ir

***Abstract*— Nowadays search engines are recognized as the pathway for accessing the tremendous amount of information in the internet. They provide aids and services for solving users' different information needs. Thus, being able to evaluate their effectiveness and performance is constantly gaining importance because these evaluations are useful for both developers and users of search engines. Developers can use the evaluation results for improving their strategies and paradigms in the development of search engines. Users, on the other hand, can identify the best performing search engines and in a better, quicker and more accurate way, gratify their information needs. Evaluation of search engines can be done in two different ways; either manually using human arbitrators or automatically using automatic machinery approaches which do not use human arbitrators and their judgments. In the case of manual evaluation methods, by now numerous and standard activities had been carried out by organizers and participants of conferences like TREC or CLEF. In the case of automatic evaluation methods, unlike variety of efforts which had been done by different researchers, no categorization and organization of such methods exists so far. As a result, anyone that wants to use one of the automatic evaluation methods must read all the relevant literature of these methods which is very time consuming and confusing activity. In this paper, we have reviewed almost all the important reported automatic methods for evaluation of search engines. Analyzing the results of this review, we have stated the requirements and prerequisites of using any of these methods. At the end, a framework for selecting the best pertinent method for each evaluation scenario has been suggested.**

Instead, automatic methods of evaluation are cheaper and faster to be performed.

***Index Terms*—Web search engine, information retrieval, automatic evaluation methods.**

I. Introduction

Web search engines already have indexed millions to billions of web pages and must be able to answer to the huge number of users request on a daily basis. They should be able to provide high quality results to the users' information needs in the least amount of time possible. These facts clearly bold the search engines' effectiveness evaluation problem. Evaluation is an important aspect in creation of search engines with high quality and measurement of their progress by the passage of time. As a result evaluations can help creators and developers of search engines to enhance and improve their product and the users to pick the best search engine which more effectively help them in solving their information needs.

There are numerous different type of evaluation methods that based on the requirements and limitations of implementation and goals of evaluation, can be carried out in black box or white box style. In the white box evaluation style there is a need of access to the components' of the search engine under evaluation and also some details of how this search engine is implemented. On the other side, in the black box style of evaluation, the search engine is viewed as a single undividable component which it is assumed that does not have any sub-components. In this black box style the search engine is evaluated by sending queries and receiving the results it provide and evaluating these results in a systematic approach. The developers of search engines often for self-evaluation of their product use these two styles of evaluation. But the outside evaluators because of the limitations that they have in the access to the components of the search engine normally only use the black box style of evaluation.

In another view, the search engine evaluation methods are grouped in two different categories; the manual evaluation approach and the automatic evaluation approach. In the manual evaluation approach, the effectiveness of the search engine is verified by some human arbitrators while in the automatic approach, the goal is to use the least possible intervention of humans' intellectual abilities in the evaluation of the search engines. In other words, the most prominent difference between the manual and automatic evaluation methods is their source of relevance judgment acquisition. The manual approaches of evaluations are more accurate than the automatic approaches although they are slower and are considerably more resource demanding and more expensive than the automatic approaches.

In this paper the main concentration is on providing a review of automatic approaches of evaluations of a group of search engines in the black box style of carrying out such evaluations. Different research activities have been carried out in the context of automatic evaluation of search engines and their comparisons. These different approaches can be categorized in four different classes as below:

- Based on user feedbacks [13] [14] [15] [16] [18].
- Based on voting and consensus among the search engines'' results [1] [8].
- Based on rank aggregation [3] [5] [12] [19] [20] [21].
- Based on known items search [4] [10] [11].

II. A REVIEW ON AUTOMATIC METHODS OF EVALUATION

A. *Methods based on users' feedbacks*

In [13] a solution for construction of training sets that can be used for learning of retrieval functions from the observed behavior of users is proposed. In this work, implicit feedbacks achieved from users' clicks on the web pages among results search engine has shown to them, are compared with explicit feedbacks that have gathered manually in order to determine that how much these implicit feedbacks can show a document in the result list can be considered as relevant. In this paper the users' clicks are considered as preferences signs that relatively show how much users prefer a document in comparison to other ones because of bearing more quality and relevance than those others which is different than other works based on users' clicks which consider users' clicks as absolute relevance signs.

The approach in [14] considers the users' behaviors and their clicks while they are searching and using search engines for finding the navigational queries and their target answers which is a specific web page and then by using these queries try to evaluate and measure the effectiveness of different search engines. This scheme is used because for every navigational query there exists only one right answer which we call it as the target. So this approach won't encounter with the complexities the approaches base on informational queries encounter with. For informational queries there are numerous relevant answers which recognizing them automatically and by an algorithm is hard while for navigational queries there exists only one possible right answer. In the [14] target for each navigational query is considered the web page among the results which were shown for a query that is clicked the most by the users whom have sent this query to the search engine under evaluation. After identification of targets, the evaluation is accomplished by computing the MRR measure for all the queries used in the evaluation. Although this approach was able to achieve high level of correlation with the manual evaluations and their ranking of the search engines base on their performance, but its most prominent weak point is that it only can be used for navigational queries. Moreover, in this approach there is a need of having access to different search engines' log files which are not publicly available. In addition to that in this approach there is a need to be able to recognize the type of query (i.e. it is a navigational or informational query). In the [14] for the type recognition of queries the approach introduced in [15] is used. In this paper for type recognition or classification of queries into one of the three classes of navigational, informational and transactional, the distribution of users' clicks on web pages present in the results' list shown for each of the queries is used. In [15] it is assumed that if most of the users' clicks are concentrated on only one specific web page then the query is of navigational type.

In the approach of [16] is a semi-automatic method that uses man-power for gathering the data about users' clicks on results of different search engines in a fair and just manner. But the humans used for gathering the clicks data are not required for providing any relevance judgements. In this approach it is just asked from user to the normal behavior and interaction that they have with search engines when they are search for finding information and solving their information needs, i.e. they should send queries to search engines and then click on results that they find interesting among web pages that the search engine has shown to them. The fair approach for gathering the clicks data acts in this way that the user first sends a query to an interface then this query is sent to both search engines A and B under comparison by this interface. Then the results' lists of these two search engines are mixed in a way that in the L upper links of the final mixed ranking there are k_a links from search engine A and k_b links from search engine B in a way that $|k_a - k_b| \leq 1$. At the end the mixed ranking is shown to the user and the links that the user is clicked on will be saved.

In the [18] it is assumed that the users' behaviors can be interpreted as signs of relevance or non-relevance of different web pages among the results. In this paper the relevance model used is beyond binary model of relevance and considers different levels of relevance that for detection of the appropriate level of relevance for a specific web page, different actions of a user with that web page will be considered. These actions are things like: copying a web page, adding to favorites, bookmarking a webpage, printing, saving and scrolling. Doing each one of these actions means different level of relevance of a web page to the user query. So these actions can be used as implicit feedbacks and if a user doesn't do any one of these actions with a web page that page can be considered as a non-relevant page. As a result by adopting this definition for the relevance of web pages, one can evaluate search engines in a real-time manner without the use of pre-specified queries and their relevant set of web pages and measure how much effective they are. This approach inconspicuously gathers users' opinion about the results a search engine is providing to them and so it is a powerful approach. The source of this strength originates from this fact that all the other approaches only consider a very limited set of parameters in their evaluation of search engines (that all these parameters and lots of other ones had used by the search engines' their selves for construction of the relevant list of documents in the first place) but this approach instead extract the users' views about the performance of a search engine. Users' views that in this methods are inferred by the actions they had done during their search sessions are completely homogenous and are of the same nature with the judgments the human arbitrators made during their judgements, so this method can have the highest correlation with human judgements. The weak point of this approach is the complexity of its implementation.

B. Methods based on consensus among engines

In the Reference Count approach introduced in [8] which is based on overlapping structure among the results lists of different engines, firstly a number of queries are picked. Then in each step to each of the engines under evaluation one of the queries is sent. Then in each step and for each engine the top ten results are considered and it is determined that how much these top results of the current engine are duplicated and present among top results of other engines. The total number of presence of top results of an engine among results of other engines is considered as its score. Finally the engines are ranked based on the scores they have achieved. This approach is called the basic scheme of reference counting. This approach will have the best performance when all the engines under investigation have similar indexes and as a result there is high probability for existence of overlapping among their results' lists. So when evaluation web search engines, because of great level of discrepancy among their indexes the amount of overlaps among their results' list will be lower than the case when the engines are indexing a limited closed set of documents which results for them having similar indexes. In other words, using Reference Count approach in web environment, will results in lower correlation scores when it is used for example for TREC conference data sets and retrieval systems that have participate in it. Also if all the engines are good performers but their results lacks adequate amount of overlaps, then usage of this approach won't results in high quality rankings on the engines. Moreover this approach considers no importance for the web pages content and another weak result of this approach is its low correlation scores with the official TREC conference rankings on the participated retrieval systems.

C. Methods based on rank aggregation

In [5] a random way for designating the relevant documents is proposed. The authors put forward the idea of random selection of relevant web pages for each query and used this scheme for evaluation of queries in the TREC conference. In this work by using the results introduced in [7] it is mentioned that the unanimousness among different judgment is rather low and even there exist discrepancies among a single human arbitrator in its judgement of relevance of different documents. As a result it is proposed to pick the relevant documents for a query among all the documents returned for it randomly. But the main weak point of this approach is its low correlation scores with official TREC ranking on the participating retrieval engines. Although there is discrepancy among users' judgments but these different judgements don't have any conceivable effect on the final ranking of the engines. It is because on the judgment of high quality documents you can't see much difference among arbitrators and discrepancies normally happen for documents that normally making a judgement for them is difficult. In other words discrepancies happen for documents that neither are having very low level of relevance and quality to be completely considered as non-relevant nor they are bearing enough levels of quality and relevance to be considered relevant surely. So picking relevant documents randomly cannot achieve the same rankings on engines that human judgments on relevance can produce. Also using this approach on web can have more unpredictable results because the distribution of relevant documents on the web is different than this same distribution for the TREC conference while this method constructs its statistical model of relevant documents based on the TREC conference data.

The AWSEEM approach is introduced in [12]. In this method firstly some information needs and queries related to them are picked (number of information needs used is 25). Then these queries are sent to AlltheWeb, AltaVista, HotBot, InfoSeek, Lycos, MSN, Netscape and Yahoo search engines. Then the first 200 results of each of these search engines are gathered into a single pool and re-ranked based on their relevance to the currently sent query. Then some top of the documents in this pool after re-ranking are considered as the relevant documents (although they are really pseudo-relevant documents). Then the vector space approach is used for computation of similarity of the top t documents of each search engines with the top s documents of the pseudo-relevant documents present in the pool constructed earlier (normally s is considered to be 50 or 100). In the last phase the search engines are ranked based on the similarity of their top results with the top documents of the pseudo-relevant documents set. Also the amount of correlation achieved by this automatic approach and manual evaluation is reported too which had high level of correlations. One of the weak points of the AWSEEM approach is that the search engines consider so many parameters in building their result list but this method in re-ranking the documents for constructing the pseudo-relevant set considers only the textual contents of documents. This approach was the basis for the work in [3] and for improving it other parameters like PageRank and AlexaRank are added to it.

The work in [19] is a method based on rank aggregation and machine learning and in it a scheme for ranking of retrieval systems is proposed. Rank aggregation is an approach that in it different rankings on documents in response to queries are merged in a way to construct a final mixed ranking that can more accurately provide answer to user query. In this paper the most important phase is the learning step that in it ranking rules by the usage of for different methods which are PageRank algorithm, Binary similarity approach (Binary Retrieval), Vector Space method and users implicit feedback are extracted. For this rules to be extracted the algorithm follows the following steps. It first finds the results for a number of queries by each of the four basic retrieval methods mentioned. Then by cross-checking of the results of these approaches with each other, the ranking rules pertinent for documents ranking will be learned (these learned rules have the characteristics of the four basic retrieval approaches in their selves and so by using these rules, automatically the rankings constructed by these for basic approaches will be mixed as a one final aggregated ranking). Like almost all the automatic approaches, in this method too, only a few number of parameters are considered in construction of the aggregated ranking while the search engines their selves consider so numerous other parameters for their retrieval in the first place and so have higher precisions in their construction of the results set. Also this approach is a machine learning one and so in the learning phase requires a big set of textual tagged documents which such a set is not available yet.

In the [20] three different approaches of data fusion which are rank position approach and approaches based on voting like Borda Count and Condorcet for ranking of retrieval systems are used. Data fusion approaches are used for construction of pseudo-relevant set. In these methods the results of different search engines are mixed by different methods and then some of top documents are considered as pseudo-relevant ones. Then these pseudo-relevant document sets are used for evaluating the effectiveness of retrieval systems. In methods consider the overlap structure among the results of search engines but does not take into account the content of web pages. Lack of overlap among the results of engines will make serious problems for the precision of the achieved ranking on engines when this approach is used. Also for this method to achieve its best results the search engines under evaluation must all have same levels of performance and effectiveness.

D. Methods based on known items search

The method in [10] uses documents in the online directories. In this method a page is considered the answer of a query, if the query terms are equal with the title of the page entry on the ODP directory completely. This approach is just applicable for navigational queries. In this approach selecting all the queries from the log of an engine can produce bias to that engine. But the harsher problem is that all search engines won't let their logs to be available publicly which this fact make it hard for carrying out this evaluation method. Also the ODP directory does not support all the languages equally well and for example for Persian language one must use another online directory. So the need for parsing the ODP directory, having access to logs of multiple search engines, and detection of navigational queries all are requirements of this approach. The similar research activity is the paper in [4] that is tailored for Persian documents and in it Iran.ir portal which contains 17000 Iranian sites which are categorized in different topics, are used.

III. Evaluation method selection criteria

If the goal is to select an evaluation method, the managers of these evaluations must have a set of rules or a framework so that they are able to pick up the most pertinent approach for evaluation which considers the evaluation project conditions. For achieving this framework, there still points about automatic evaluation methods that we should mention in what follows:

- Methods based on user feedbacks: if the search engine under evaluation is being used by numerous people on a daily basis, using this category of methods is most effective. This is because by tracking users' behavior we can infer the effectiveness of search engine and users' satisfaction approximately without becoming worry about that our approach is biased toward a specific engine or not. But when the search engine is not used by adequate number of users on a daily basis these approaches are not as effective. Also in this case we have to tell the users that they are using the engine for the goal of its evaluation which will have destructive impact on the normal user behavior i.e. the pattern of behavior they normally show when they use search engine is not achieved in these circumstances because users' behavior changes subtly.
- Methods based on consensus: when there is an engine that is performing very better than all the other engines then methods based on consensus are not quite effective. This is because the better performing engine will have distinct results which do not have enough overlaps with other engine results which results in distorted ranking on the engines. These class of approaches are good for when there is high level of overlap among search engine indexes and these methods have good accuracy for closed set of documents too because in this case retrieval systems will have similar indexes with each other. Also for using these approaches there is a need for procurement of adequate number of queries with good quality that cover a good volume of different topics.
- Methods based on rank aggregation: these approaches are good for when the engines under evaluation don't have high levels of overlaps among their results and

the cover of different topics and parts of web in their indexes and also there is no access available to implicit users' feedback. But main drawback of these approaches is that they only use very limited set of parameters in their construction of re-rankings of documents in response to different queries while the engines their selves had used a lot more parameters for the providing results in the first place.

- Methods based on known items search: evaluation using these approaches is only possible when queries and their target answers are available. So these approaches can be used only by leveraging navigational queries. As a result another important factor in determining the approach which is going to be used is the type of queries which are going to be used in that approach. Methods which are using navigational queries are more accurate and simpler to implement than other automatic approaches of evaluation of search engines.

IV. Suggested framework for method selection

Based what has been mentioned about the characteristics of different methods and the different criteria for method selection, a framework for method selection is proposed here which is shown in Fig. 1. In accordance to this framework, the first criterion for method selection is query type. Other criteria are available data about search engines (e.g. their logs) and the maturity and so the user-base of the engines. If an engine possesses a considerable portion of the search market and so have a big user-base and so consequently its usage statistics are high, the methods based on user feedbacks can be used for its evaluation. Although the availability of access to the feedback data of users must be considered too. Otherwise because of the simplicity of the methods based on consensus in comparison to methods based on rank aggregation, if there is a good level of overlaps among engines results the methods based on consensus are recommended. But in the web environment because there is no adequate level of overlap among results of web search engines especially for the national search engines and international ones like Bing and Google these kinds of approaches are ineffective. In these circumstances the third group of approaches which are based on rank aggregation are recommended for leverage. When these approaches are going to be used you have to be cautious that the parameters considered for re-ranking or learning rules for construction of the final aggregated ranking do not create any bias toward any of the search engines.

At the end it should be noted that another important factor in the success of an evaluation method is the ability of creation of queries that are good representatives for actual users' information needs. It is obvious that these set of queries must have enough queries and also have enough topical divergence which is in accordance to users' information needs. In the best condition the access to engine log can be used for creating the test set queries. But in most cases there are limitations for example often there is no access to search engine log or the log given publicly does not represent users' real information needs perfectly. Also queries classification to two classes of navigational and informational can be challenging.

V. Conclusion

As introduced in this paper, there are numerous different approaches for evaluation of search engines. In this paper a classification on these approaches was introduced and a framework for choosing the most pertinent approach, regarding the evaluation scenario at hand, was proposed.

In the evaluation process the most important requirement is the construction of relevance judgment set which is the core of the evaluation approach. Constructing this set highly depends on the types of queries, used for evaluation. For instance, for navigational queries, the process of constructing the set can be easily automatized.

In addition, it is worth to be mentioned here that users' feedback is a precious source of information which is used by many commercial search engines for improving search results. If such valuable source of information is available, then it is strongly recommended for automatic evaluation methods to utilize it during the process of constructing the relevance judgement set.

In future research activities, we aim at conducting an automatic approach for comparison of national and international search engines. The approach is a hybrid method of the consensus based and rank aggregation methods. The reason of choosing this type of methods is the lack of access to the users' behaviors and their implicit feedback data and also because of the lack of adequate amount of overlap among results of search engines under evaluations.

References

[1] M. Azimzade, S. Samuri, A. Yari, "Verification and qualitative comparison of search engines in the Persian web field" (In Persian language), 18th annual national computer association conference, February 2013.

[2] S. Moosavi, M. Azimzadeh, M Mahmoudy, A. Yari, "Presenting an efficient and complete framework for Persian search engine evaluation" (In Persian language), 18th annual national computer association conference, February 2013.

[3] R. Badie, M. Azimzadeh, M.S. Zahedi, S. Samuri, "Automatic evaluation of search engines: Using webpages' content, web graph link structure and websites' popularity" Seventh International Symposium on Telecommunications (IST2014), September 09-11, 2014.

[4] M. Mahmoudy, M. Sadegh zahedi, M. Azimzadeh, "Evaluating the retrieval effectiveness of search engines using Persian navigational queries", Seventh International Symposium on Telecommunications (IST2014), September 09-11, 2014.

[5] I. Soboroff, C. Nicholas, and P. Cahan, "Ranking retrieval systems without relevance judgments," in Proceedings of the 24th annual

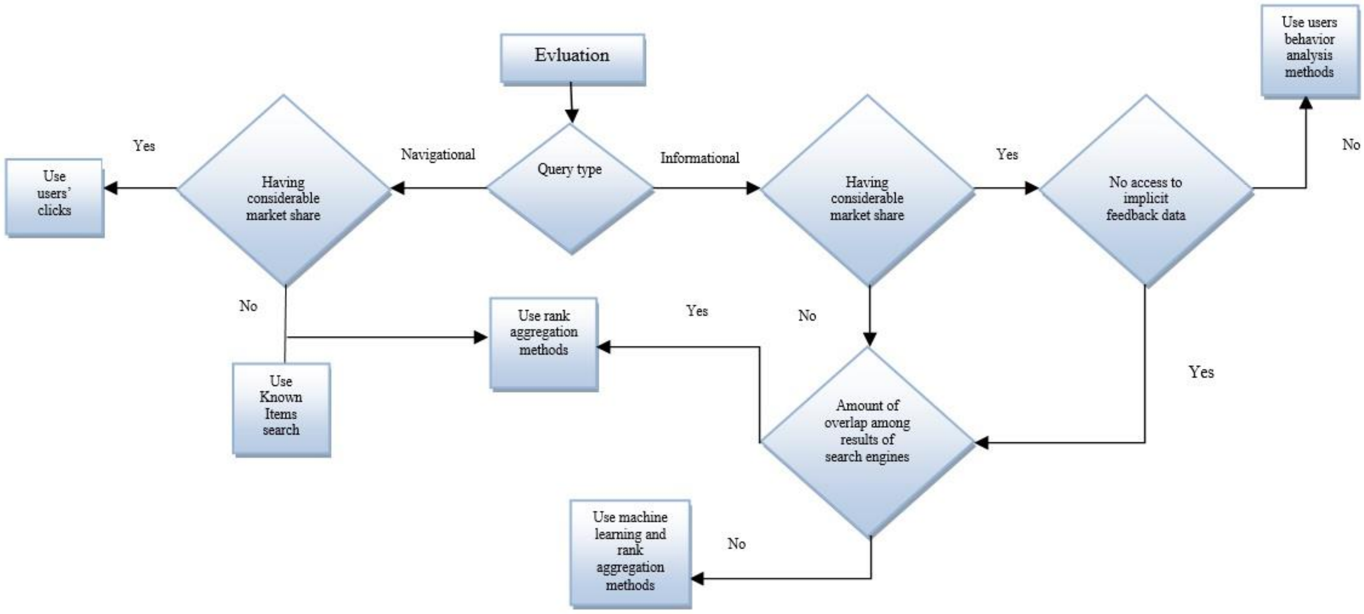
Evluation
Query type
Navigational
Informational
Having considerable market share
Yes
Use users' clicks
No
Use Known Items search
Use rank aggregation methods
Having considerable market share
Yes
No access to implicit feedback data
No
Use users behavior analysis methods
Yes
No
Amount of overlap among results of search engines
Yes
No
Use machine learning and rank aggregation methods

Fig. 1. Automatic evaluation method selection framework

international ACM SIGIR conference on Research and development in information retrieval, 2001, pp. 66-73.

[6] S. P. Harter, "Variations in relevance assessments and the measurement of retrieval effectiveness," JASIS, vol. 47, pp. 37-49, 1996.

[7] A. Spink and H .Greisdorf, "Regions and levels: measuring and mapping users' relevance judgments," Journal of the American Society for Information science and Technology, vol. 52, pp. 161-173, 2001.

[8] S. Wu and F. Crestani, "Methods for ranking information retrieval systems without relevance judgments," in Proceedings of the 2003 ACM symposium on Applied computing, 2003, pp. 811-816.

[9] J. Callan, M. Connell, and A. Du, "Automatic discovery of language models for text databases," in ACM SIGMOD Record, 1999, pp. 479-4.۹۰

[10] A. Chowdhury and I. Soboroff, "Automatic evaluation of world wide web search services," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 421-422.

[11] S. M. Beitzel, E. C. Jensen, A. Chowdhury, and D. Grossman, "Using titles and category names from editor-driven taxonomies for automatic evaluation," in Proceedings of the twelfth international conference on Information and knowledge management, 2003, pp. 17-23.

[12] F. Can, R. Nuray, and A. B. Sevdik, "Automatic performance evaluation of Web search engines," Information processing & management, vol. 40, pp. 495-514, 2004.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 154-161.

[14] Y. Liu, Y. Fu, M. Zhang, S. Ma, and L. Ru, "Automatic search engine performance evaluation with click-through data analysis," in Proceedings of the 16th international conference on World Wide Web, 2007, pp. 1133-1134.

[15] Y. Liu, M. Zhang, L. Ru, and S. Ma, "Automatic query type identification based on click through information," in Information Retrieval Technology, ed: Springer, 2006, pp. 593-600.

[16] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data," ed: Citeseer, 2003.

[17] G. Mood, "Boes, Introduction to the theory of statistics," McCraw-Hill Statistics Series, 1974.

[18] H. Sharma and B. J. Jansen, "Automated evaluation of search engine performance via implicit user feedback," in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval ,۲۰۰۵ ,pp. 649-650.

[19] R. Ali and M. S. Beg, "Automatic performance evaluation of web search systems using rough set based rank aggregation," in Proceedings of the First International Conference on Intelligent Human Computer Interaction, 2009, pp. 344-35۸

[20] R. Nuray and F. Can, "Automatic ranking of information retrieval systems using data fusion," Information processing & management, vol. 42, pp. 595-614, 2006.

[21] H. Sadeghi.,(2011). "Automatic Performance Evaluation of Web search Engines using judgements of Meta search Engines", Online Information Review,ISSN:1468-4527,Emerald Publishing Limited, pp.957-971.

[22] W. Tawileh, J. Griesbaum, T. Mandl ,Evaluation of five web search engines in Arabic language. Proceedings of LWA. (2010).

[23] D. Lewandowski, Evaluating the retrieval effectiveness of Web search engines using a representative query sample. Journal of the Association for Information Science and Technology (2015).

[24] J. Bar-Ilan, M.Levene, A method to assess search engine results. Online Information Review 35(6), 854-868. (2011).