EVALUATION OF INFORMATION RETRIEVAL SYSTEMS USING VARIOUS MEASURES

Princy Sisodiya* and Vandana S Sardar**

*Dept. of CSE, M.S. Ramaiah Institute of Technology, Bangalore, India
**Dept. of CSE, M.S. Ramaiah Institute of Technology, Bangalore, India

ABSTRACT: Our world revolves around technology and information. From a computer system present on desk to smart phone carried everywhere, the use of technology to aid human life has increased enormously. This leads to the production of massive amount of data; be it files belonging to an organization or a person's heartbeat rate. All data is stored. The main challenge is to retrieve information out of it. Additionally, a user specific information retrieval is also needed. Information Retrieval Systems is one of the most used applications in today's life, ranging from search engine searching for a given query to intelligently analyzing and retrieving accurate details of a particular disease. Along with predefined retrieval items, a user can give a new query to the system and relevant information will be retrieved. Since, the usage is wide; the need for evaluating such systems becomes a priority. Federated search is an information retrieval technology that allows the simultaneous search of multiple searchable resources and aggregates the results that are received from the search engines for presentation to the user. It has data for numerous queries and search engines. In this paper, various applications of Information Retrieval Systems are discussed, followed by different approaches used for the evaluation. The dataset used is Federated Web Search track TREC 2014 of FedWeb Greatest Hits collection which allows combining results of multiple search engines. The methods used for evaluation along with the results are provided.

KEYWORDS: Fedearted Web Search, Precision, Recall, F-measure.

INTRODUCTION

The collaboration of internet with computer systems, mobiles, smart devices has lead to enormous amount of data. The collected data is beneficial if information could be retrieved from it. The information derived from data provides insight to a system or may also help to conclude or arrive at a conclusion for a specific application. The first description of information searched by computer was mentioned in 1948 by Holmstrom, who mentioned Univac computer. The information need not only be retrieved but management of information is also essential. These features when combined leads to Information Retrieval System. The main feature of these systems is to provide information or output based on the query provided to it. The most commonly used information retrieval system is web search engine, in which a query is given and corresponding to it various outputs are shown. The vast spread usage of web search engine in today's knowledge society has encouraged further the need of very large scale retrieval systems.

With the advent of digital information, the need for availability of text databases online has also increased. The access to desired information has encouraged better techniques which also promoted research in the field of Information Retrieval. The Information System is broadly classified as traditional or classical system and modern systems. The main focus of former is classifying the problem while in the latter user acquires large number of relevant entities for the query.

Evaluation of Information Retrieval Systems requires test collection: corpora of documents, sets of topics, and relevance judgments pointing which documents are relevant to which topics. Test collection comprises of three attributes: a collection of documents, queries expressing the information need and set of relevance judgments for each query pair. By tuning parameter of a system and reporting results on test collection so that performance can be maximized on that the test collection, gives wrong results. This is because expected performance of the system is tuned. It can be improved by using more than one development test collections and tuning parameters on these development test collections.

The TREC (Text REtrieval Conference) effort founded in 1992 and continuing to the present, took as its prime motivation the need to create realistically-sized collections, and to consolidate and extend research in retrieval technology through the use of shared, high-quality experimental data and standard evaluation techniques. In addition to the test collections themselves, TREC has made large amounts of a novel kind of data available: namely, the document rankings or runsets submitted by participating groups. These runsets have provided the material for research on evaluation itself, and have inspired a large volume of such work over the past decade.

RELATED WORK

It is apparent that there is availability of one more than one system to do any specific task. Hence, there is need to select one among many to optimize desired result. This is where, evaluation comes into picture. The need to choose one among many available options, which would yield the best result, evaluating various systems is necessary. Evaluating an Information System has its base as relevance. The objects retrieved by an Information System have various degrees of relevancy, in other words some retrieved entities are more relevant than others. To aid this, rank is given to each retrieved entity. The entity having more relevance is given a better rank than less relevant entity. This is the concept of binary classification. For non-binary relevance two notions have been interpreted. One approach treats relevance as comparative notion leading to user preference documents and the other treating relevance as quantitative notion, which provides multi grade relevance.

An Information System may contain various formats of files such as PDF, PPT or DOC. An user may need to retrieve any specific document from the system, which should be capable of retrieving valuable information at a fast pace and accurately. Content Based Document Information Retrieval denoted as CBDIR system was developed which retrieved information from the actual content of a document. The major keywords were extracted using Latent Dirichlet Allocation (LDA) and the system was capable of communicating with existing web service. For indexing B-tree approach was used. This proposed system increased flexibility, effectiveness and lead to fast retrieval of information. It was concluded that CBDIR improves performance up to 20% compared to the baseline, average recall was increased by 30% and F-measure calculation showed that improvements were seen in all categories over the baseline.

Another feature to be added to information retrieval system is to make it more interactive is by introducing the concept of personalized information retrieval systems. The precision and quality of personal information retrieval is dependent on right degree of the user's interest. Another work done presents a personalized information retrieval system based on multi-agent, to accomplish information retrieval according to user interest knowledge using multi-agent collaboration to provide personal service to user. The user intent model was dependent on browsing history record and registration data and the system could update user interest model when user's interest changes. The system was developed using combination of apparent feedback method and connotative feedback method to discover user intent. In apparent feedback, user is capable of giving personal interest as input or evaluates the current work, while in connotative feedback method; system obtains information about user's interest via tracking user behavior and operation. The algorithm used could discover user interest in time, control safely the scale of user interest model and increase effectively document filtration efficiency. Also, precision was improved by 15 %-35%.

Web-oriented architecture of personalized intelligent information system was presented and the local domain ontology repository of biomedical disease system was constructed as a hierarchal tree, which used OWL (Web Ontology Language) language to detail annotate syntax and semantics of concepts, their attributes and relations among the concepts. The experiment results showed that Web-oriented personalized intelligent information retrieval system can obtain speedy search information mostly that are relative to preferences or interests from local domain ontological repository. The information semantics-based intelligent clustering algorithm and rules of inference in the systemic architecture were used to implement automatic acquirement of remote information resource on the Internet and pretreatment of local domain ontological repository. The system is capable of meeting the requirements of personalized intelligent information search diseases diagnosis, and leads to improvement of recall and precision on information retrieval on biomedical diseases system as knowledge maps leads to improvement of timely reflection of information on Internet and navigation of retrieving information.

The 2014 Federated Web Search (FedWeb) track promotes research on federated search with realistic web data. Federated Web search is the approach where multiple search engines are queried simultaneously, and their results are combined into one consistent search engine result page. The goal of the Federated Web Search track is evaluation of approaches to federated search which is done at very large scale in a realistic setting, by combining the search results of existing web search engines. In the following paragraphs, work done using Federated Web Search track is discussed.

The Information Management System Research Group of the University of Padua contributed in two tasks: vertical selection and resource selection. Their aim was to measure effectiveness of TWF_IRF in Federated Web search. The measures used in resource selection were nDCG, nDCG@10, nDCG@20, nP@5, np@1. Whereas Precision, Recall, F1-measure was used for vertical selection. Their experimental stated that TWF_IRF was not affected by stemming when

used in Web Search setting. On the other hand, TWF_IRF was improved by removing stop-words. Also, the usage of IRF differs when used for vertical representation or search engine ranking.
Chinese Academy of Sciences (ICTNET) made contribution to the task of Vertical Selection and Resource Selection. The task of vertical selection started with LSI model which used Google Custom Search API to represent vertical and query. Next, LSI model was used to calculate the similarity between them. Random Forest was used for text classification and Frequent Term Rank (FTR) was used for representation of verticals. Among various combinations of method used borda fuse combination of three methods gave the best result evaluated using Precision, Recall and F1-measure.The work in resource selection used LSI model which had top 20 resources for each query. The various methods used were Text classification strategy, LSI, resource's pagerank which were evaluated using nDCG@10, nDCG@20, nP@1, nP@5.
The work carried out as learning to rank approach was used for resource selection task and binary judgments were used to estimate collection relevance scores. For resource selection, Collection-centric (CC) and Document-centric (DC) were combined. The result showed that learning-to-rank outperforms the DC model by 13%. Also, it suggested that improvements in resource selection lead to improvement to vertical selection.
One of the most popular methods to describe resources is central sample index. This approach has disadvantages such as substantial storage space and administrative overhead. University of Twente, suggested the use of vocabulary-based resource description which was based on statistics of term related features in each shard used in ranking functions. The measures used were nDCG@20, np@1, and nP@5.
A query independent method to measure a search engine's impact called as Search Engine Impact Factor was used by East China Normal University for the task of vertical selection. Two methods were proposed: one based on economic exploration report regarding to the distribution of market shares of search engines and second to use TREC 2013 datasets. The different methods were used: first was to match keywords in query with the label of verticals and second to build a supervised machine learning model. The measures used were Precision, Recall and F-measure.
University of Delaware used various rules to rank a vertical. The rules included three cases: presence of interrogative word, absence of interrogative word and presence of vertical word. Based on these rules, verticals were given a score and top five verticals were selected for each query. It was observed that precision and F1 were improved.

APPROACH AND METHODOLOGY

Federated search is the approach of simultaneous querying of multiple search engines and combining their results into one coherent search engine result page. The motivation of the work is from TREC federated Web Search (FedWeb) track whose goal was to evaluate approaches to federated search at very large scale in realistic setting, by combining the search results of existing web search engines. The dataset used is FedWeb Greatest Hits collection. The details are provided for FedWeb 2014, which contains the following

- Search results and corresponding web pages crawled in April and May 2014 from 149 existing web searches.
- 4000 random single word queries issued to all search engines and set of topics consisting of 275 information needs and corresponding queries
- 231 information needs and keyword queries which were not used in the official FedWeb track.

Various tasks which can be performed for Federated Web Search are resource selection which selects the right resources (search engines) from a large number of independent search engines given a query, results merging which combines the results of several search engines into a single ranked list and vertical selection.
The work focuses on one of the tasks of FedWeb 2014 track namely Vertical Selection. Vertical Selection classifies each query into a fixed set of 24 verticals. Metrics to be used for vertical selection are precision, recall and F-measure.
Among various documents present in the datasets, topic pages and sample pages generated by various search engines were used for evaluation. Apache Lucene is used to index documents for every search engine. As shown in Fig. 1, the process starts with parsing of collection and indexing of search engines. The query is then provided to the search engine and the score of retrieved documents is obtained. If there are more queries or search engine, the loop is continued, unless there are no search engines or queries. Once, the loop terminates engines are categorized in verticals and score for each vertical is computed called as vertical score. The example run file consists of query matched to one or more verticals. The example run file is used to compute precision, recall and F-measure.

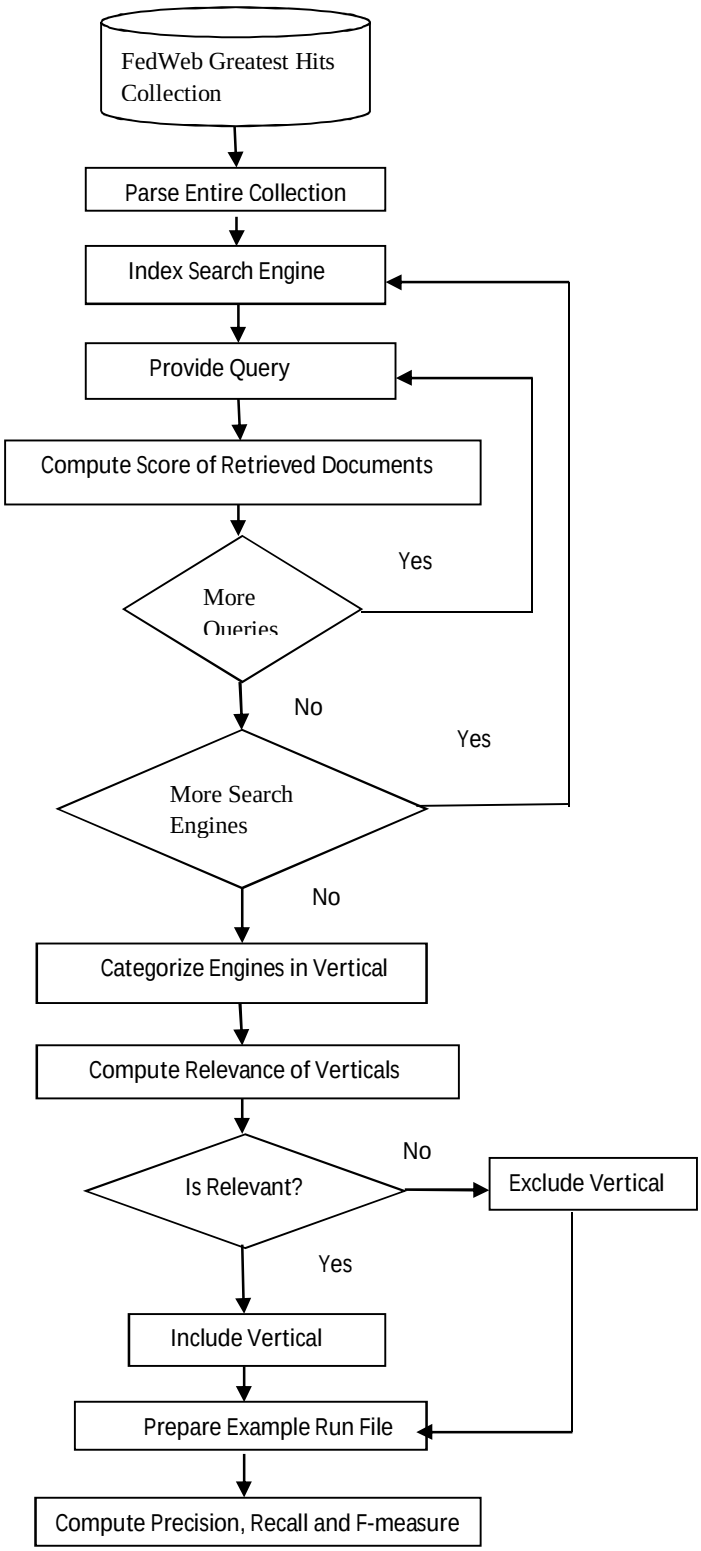

Fig.1. Flowchart for vertical selection and computation of measures

EXPERIMENTAL RESULTS

The following results were calculated using FedWeb Greatest Hits (FedWeb 2014) collected from TREC which includes topic docs and sample docs. In the vertical selection task, the scores obtained are provided for two queries as shown in Fig. 2 for the query 7015 (the raven) and in Fig. 3 for the query 7044 (song of ice and fire). Both Fig. 2 and Fig. 3 show

scores for 60 search engines. Search engines are numbered randomly as not all the search engines from the collection are being used for these results so far. The scores are calculated for 100 search engines out of 149 search engines for topic docs and 50 search engines for sample docs. For example, out of 8 search engines in Encyclopedia vertical, scores are being calculated for 7 search engines while scores are being calculated for all the four search engines under the vertical Blogs. The evaluation is carried for two scenarios i.e., top five verticals and average as threshold. For the scenario of top five verticals example run file includes scores for top five verticals and for average as threshold example run file contains verticals which are higher than average of all vertical score.

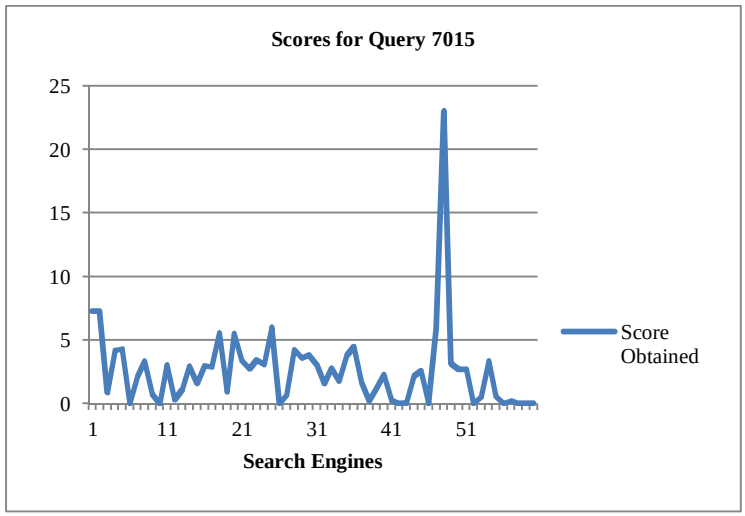

Fig.2. Score obtained for the query 7015 “the raven”

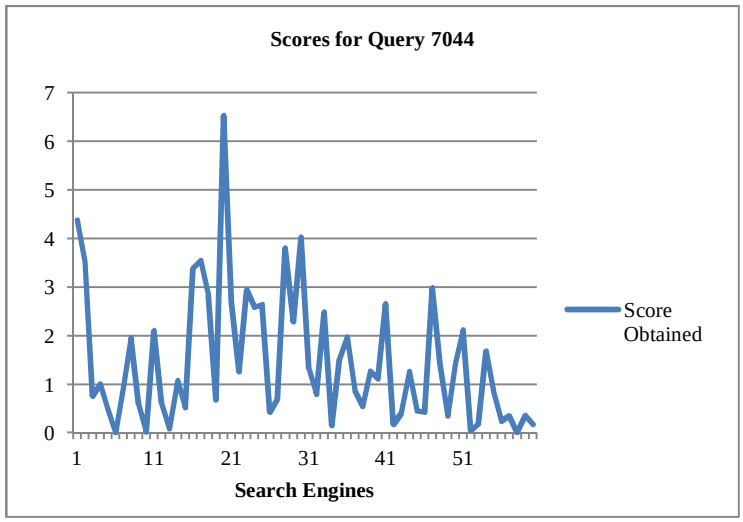

Fig.3. Score obtained for the query 7044 “song of ice and fire”

The example run file is a text file consisting of 231 entries. Each entry of example run file consists of three fields, i.e., topic number, the engine identifier of selected vertical and the run tag. The example run file is given as input to the evaluation script which calculates precision, recall, and F- measure. The various evaluation measures for a vertical selection task are shown in Table I. As the documents used are topic pages, recall is found to be 1 for most of the queries.

Table II shows outcome for topic docs and sample docs. It is observed that better results are obtained for top five verticals than average as threshold scenario. The results are shown in terms of mean precision, mean recall and mean F-measure.

Table 1. Evaluation measures for Vertical Selection

Query	Precision	Recall	F-measure
7015	0.2	1	0.33
7044	0.4	0.4	0.40
7045	0.2	1	0.33
7072	0.2	1	0.33
7092	0.6	1	0.75
7111	0.4	1	0.57
7123	0.4	1	0.57
7137	0.2	0.5	0.29
7146	0.2	1	0.33
7161	0.2	1	0.33

Table 2. Outcome for Vertical Selection

	Topic Docs Top Five Verticals	Topic Docs Average as Threshold	Sample Docs Top Five Verticals	Sample Docs Average as Threshold
Mean Precision	0.3000	0.2161	0.1000	0.1046
Mean Recall	0.8900	0.9600	0.2033	0.4367
Mean F-measure	0.4245	0.3395	0.137	0.1594

CONCLUSION

This paper has summarized the work done in the field of evaluation of Information Retrieval Systems. The techniques used by various researches with the obtained results have been discussed. The description about vertical selection task of Federated Web Track 2014 and FedWeb Greatest Hits collection has been discussed. The methodology used to proceed from score of a single search engine to the task of vertical selection has been clearly mentioned. The evaluation results of vertical selection task in terms of precision, recall and F-measure are provided for the work done so far. Better results are obtained for topic docs than sample docs as only 50 search engines have been considered for sample docs whereas 100 search engines have been considered for topic docs.

REFERENCES

[1] Bah A., Sabhnani K., Zengin M., and Carterette B.. “University of delaware at TREC 2014”. In The 23rd Text Retrieval Conference (TREC), 2014. National Institute of Standards and Technology (NIST), 2014.

[2] Balog K., “NTNUiS at the TREC 2014 federated web search track”. *In The 23rd Text Retrieval Conference (TREC)*, 2014. National Institute of Standards and Technology (NIST).

[3] Buccio E. D., Masiero I., and Melucci M., “University of Padua at TREC 2013: federated web search track”. *In Proceedings of the 22nd Text REtrieval Conference Proceedings (TREC),* 2013. National Institute of Standards and Technology (NIST), 2013.

[4] Cha Moon Soo, Kim So Yeon, Ha Jae Hee, Lee Min-June, Choi Young-June and Sohn Kyung-Ah, "CBDIR: Fast and effective content based document Information Retrieval system," Computer and Information Science (ICIS), 2015 IEEE/ACIS 14th International Conference on, Las Vegas, NV, 2015, 203-208.

[5] Clough Paul, Sanderson Mark, “Evaluating the performance of information retrieval systems using test collections”, *Information Research, Vol. 18No.2*, June 2013.

[6] Demeester T., Trieschnigg D., Zhou K., Nguyen D., and Hiemstra D. FedWeb Greatest Hits: Presenting the New Test Collection for Federated Web Search. In 24th International World Wide Web Conference (WWW 2015),2015.

[7] Giachanou A., Markov I., and Crestani F.,"Opinions in federated search: University of lugano at TREC 2014 federated web search track". *In The 23rd Text Retrieval Conference (TREC),* 2014. National Institute of Standards and Technology (NIST), 2014.

[8] Guan F., Zhang S., Liu C., Yu X., Liu Y., and Cheng X., "ICTNET at federated web search track 2014". *In The 23rd Text Retrieval Conference (TREC),* 2014. National Institute of Standards and Technology (NIST), 2014.

[9] Hiemstra Djored and Aly Robin, "Two selfless contributions to web search evaluation". *The 23rd Text Retrieval Conference (TREC),* 2014. National Institute of Standards and Technology (NIST), 2014.

[10] Holmstrom JE (1948). "'Section III. Opening Plenary Session". The Royal Society Scientific Information Conference, 21 June-2 July 1948: report and papers submitted: 85.W. Bruce Croft, "What do people want from Information Retrieval ", D-lib Magazine.

[11] Jin S. and Lan M.. "Simple may be best - a simple and effective method for federated web search via search engine impact factor estimation". In The 23rd Text Retrieval Conference (TREC), 2014. National Institute of Standards and Technology (NIST), 2014.

[12] Li W., Zhang X. and Wei X., "Semantic Web-Oriented Intelligent Information Retrieval System," BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on, Sanya, 2008, 357-361.

[13] Voorhees E. M., "The philosophy of information retrieval evaluation". *In CLEF '01: Revised Papers from the Second Workshop of CLEF*, pages 355-370, London, UK, 2002. Springer-Verlag.

[14] Voorhees Ellen M., Paul, Soboroff Ian, "Building Better Search Engines by Measuring search Qualities", Available: http://www.infoq.com/articles/building-better-search-engines-by-measuring-search-quality.

[15] Zhou Bing, Yao Yiju (2010) "Evaluating Information Retrieval System Performance Based on User Preference", *Journal of Intelligent Information Systems*, **34**, Issue 3, 227-248.

[16] Zhu Z. and Wang J. Y., "Research of personalized information retrieval system based on multi-agent and user interest model," Machine Learning and Cybernetics, 2009 International Conference on, Baoding, 2009, 2148-2152.