Asian Journal of Information Technology 16 (7): 599-604, 2017
ISSN: 1682-3915

Psuedo Relevance Feeback for Literary Documents

[1]Yasir Hadi Farhan and [2]Shahrul Azman Mohd. Noah
[1]Department of Computer Science and Information Technology,
[2]Centre for Artificial Intelligence and Technology (CAIT),
University Kebangsaan Malaysia, Selangor, 43600 Bangi, Malayisa

Abstract: One of the biggest issues that affect the Information Retrieval (IR) systems performance is the difficulties facing users to define exactly what their information needs as that information might be a gap in their knowledge. Such an issue is more problematic for classical and literary documents such as the holy Qur'an. One of the approaches to overcome such an issue is pseudo-relevance feedback which assumes a small number of top-ranked documents as relevant in the initial retrieval results. It selects related terms from these documents to improve the query representation through query expansion. Among the issues in the Qur'anic text are ambiguities and complexit y of the text. Due to these issues, users need to reformulate and refine their queries to match their information needs. Pseudo-relevance feedback can help relieve these issues. The classic Rocchio algorithm has been widely used to support query reformulation in pseudo relevance feedbacks. In this research, a modified Rocchio algorithm was proposed by considering element of terms selection and query importance. In this case it combines the Term Frequency and Inverse Document Frequency (TF-IDF) weights and Rocchio's algorithm weights in order to generate a new query. It also uses the frequency of terms to choose suitable expansion words. Evaluation of the proposed algorithm were compared against the probabilistic IR Model implemented in Lucene toolkit and against the WordNet query expansion approach. The experiments only consider relevance feedbacks after two iterations. The evaluation used the Qur'anic dataset previously used by other researchers. Twelve queries were considered during the evaluation. The results of the experiments showed that the proposed method exhibit significant improvement in recall and precision. The average precision through pseudo relevance feedback for the first iteration was 8.3% and for the second iteration was 11.3% whereas, the average precision by Lucene was 3.3% and the average precision by WordNet query expansion was 2.7%. These results prove that the proposed method improves retrieval performance.

Key words: Information retrieval, pseudo relevance feedback, query expansion, Qur'anic text retrieval, expansion, algorithm

INTRODUCTION

Information Retrieval (IR) is a commonly utilized phrase today. IR is classified as the action of locating and retrieving relevant documents from a document collection in response to query related to information-need requests. IR techniques have been significantly developed to adapt to the niche needs of each field. One of the interesting fields is verse retrieval from Qur'anic texts (Sultan *et al*., 2011). Among the reasons for such interest in this field includes expansion of Islam, the openness of Muslims from migration and exchange of cultures between Muslims and Westerners and the large numbers of conversions to Islam (Hasan, 2012).

Users may have difficulties in presenting their information needs due to the aforementioned complexity of the nature of the Qur'anic text. Research has indicated that queries were usually short and between 2-3 words (Metzler and Croft, 2007; Rocchio, 1971). Due to these issues, most search applications include functions, like keyword suggestions, spelling correction and classification to help users in articulating their queries. Relevance Feedback (RF) will be one of the methods that will be discussed. However, limited efforts were focused on the application of RF for Qur'anic verse retrieval. Noordin and Othman (2006) have put forward a web-based system to retrieve Qur'anic texts and knowledge sourced from the understanding or citations of the Al-Qur'an from the web. Shoaib *et al*. (2009) have proposed a model that used WordNet relationships in a relational database model to improve retrieval of related verses to identify topics from the Al-Qur'an regardless of

Corresponding Author: Yasir Hadi Farhan, Department of Computer Science and Information Technology, University Kebangsaan Malaysia, Selangor, 43600 Bangi, Malayisa

the presence of a query-word. Also, Nasharuddin *et al.* (2010) have adopted the Stemming Semantic Query (SSQ) for Al-Qur'an documents. They concluded that the use of stemmers in the semantic approach improved retrieval results of relevant and related Al-Qur'an documents. Yu *et al.* (2003) proposed a framework for ontology representation for the Holy Qur'an. In spite of the studies mentioned before, the issue of the Qur'anic text retrieval is still arisen and it's not yet achieved the user satisfaction. In psuedo relevance feeback for literary documents order to address the aforementioned issues, pseudo relevance feedback are used in this study.

In order to develop and evaluate the Pseudo Relevance Feedback (PRF) performance for tackling the retrieval of the Qur'anic text an experimental study is conducted. A suitable PRF approach is proposed for improving the performance of Qur'anic retrieval by modifying the expansion technique.

MATERIALS AND METHODS

The conducted research method involves the following four phases: problem identification phase, pre-processing phase, requirement phase and evaluation phase.

Research phases

Problem identification phase: This phase aims to understand the Qur'anic text retrieval through defining its objectives and constraints and identifying the most related works. In particular, the most important challenges which encounter these works in developing their performance are presented in this phase. The effectiveness of relevance feedback which are presentedto improve the retrieval performance of the Qur'anic text is evaluated according to their results that have been obtained using pseudo RF. A test was conducted to determine the retrieval performance of Qur'anic texts. It involved twelve queries from Muslim scholars where all answers were provided by Muslim scholars. Each relevant documents for these queries has been assigned to represent as a correct answer for these queries.

Pre-processing phase: Designing the logical sequence for the suggested pseudo RF, represents the main objective of the current phase. In other words, the problem modelling including the data structure and the mathematical formula is done in this phase. Additionally, this phase focuses on designing the flowchart that is adopted to improve the retrieval performance of the Qur'anic text. The basic assumption of pseudo RF is that the top-ranked documents in the first retrieval result contain many useful terms that can help discriminate relevant documents from irrelevant ones. It automatically expands a query by assuming the top retrieved documents from the initial retrieval are relevant. The new terms from the initial documents are selected according to weighting functions. These functions are Rocchio's weights and TF-IDF where Rocchio's weights and IDF should be >1.5 and TF should be >2. These terms are then added to the original query terms for the next iteration. Pseudo RF have proved to be highly effective, especially for short query terms. However, the performance of pseudo RF highly depends on the quality of the initially returned documents. It works well for those good initial queries (Huang, 2006). The overall procedure of the proposed method can be summarized as follows:

- Calculate the weights of all the terms in the collection of results depends on Rocchio algorithm
- Calculate the inverse documents frequency FT-IDF, for all the terms

$$IDF_t = \log n\left(\frac{N}{N_t}\right)$$

Where, N is the total number of documents in |D|

- Determine which terms have Rocchio's weight and IDF >1.5
- Determine which terms have TF >2
- Pick all the terms which have all the conditions above

After applied the conditions on the terms then add all these terms to the original query to generate the new query that's will lead to get more documents which will be relevant with the new query.

Requirement phase: In this phase, the performance of information retrieval is improved. Particularly initial population generated through the previous phase is adopted as an input to this phase and its quality is improved via. the proposed pseudo RF. Pseudo RF which were designed in the preprocessing phase are coded in Java Net Beans IDE 8.0.2. Also, the equations which are used in this research are Rocchio's algorithm, TF-IDF and the modifying expansion equation where it applied to get the weights of the terms. Additionally, the corpus that contain the 6236 documents which refer to the Ayahs (Verse) or text of the Qur'an. Also, the standard queries that provide by the Islamic scholar and the correct answer of these queries. All of these requirements are applied to get the result where the standard queries are applied first

by using Lucene toolkit to get the initial results. Depends on the initial results, the standard query expanded by pseudo RF through the proposed method which depends on the Rocchio's weights and TF-IDF where the proposed method found a relation between these weights (Rocchio's weights and TF-IDF). Depends on this relation between the weights the new terms are picked to add to the original query in order to expand the original query that generally lead to improve the retrieval performance of the Qur'anic text. The expansion by using WordNet also used to compare with the proposed method.

Evaluation phase: This phase aims to evaluate the performance of the proposed modified pseudo RF. Evaluations were based on the effectiveness metrics of recall, precision and F-measure. The equations of each measures are as follows, respectively:

$$\text{Recall } R = \frac{TP}{TP+FN}$$

$$\text{Precision } P = \frac{TP}{TP+FP}$$

Where:
TP = The number of correct retrieved document
FP = The number of incorrect retrieved document
FN = Number of correct document but not retrieved

Hence, it is possible to calculate the overall accuracy which called F-measure based on the following equation:

$$\text{F-measure} = 2\times\frac{P\times R}{P+R}$$

Recall measure is the ability of the search application in finding all the relevant documents for a query whereas precision measures how well it is ranking relevant documents. F-measure is an effectiveness measure that summarize the effectiveness in terms of recall and measure in a single numerical value. In this study, the three effectiveness metrics were used as it has been widely implemented and became a standard in IR research field.

The proposed method; Modified pseudo relevance feedback: Pseudo-relevance feedback also known as local feedback or blind feedback. Pseudo RF is a technique commonly used to improve retrieval performance (Buckley *et al*., 1993; Efthimiadis, 1996). Its basic idea is to extract expansion terms from the top-ranked documents to formulate a new query for a second round retrieval. Through a query expansion, some relevant documents missed in the initial round can then be retrieved to improve the overall performance. Clearly, the effect of this method strongly relies on the quality of selected expansion terms. If the words added to the original query are unrelated to the topic, the quality of the retrieval is likely to be degraded (Yu *et al*., 2003). The technique of the Modified Pseudo RF (MPRF) method depends on three conditions and the new generated term should be picked depends on these three conditions. These conditions are: the Rocchio's weight and IDF of this term should be >0.15 whereas TF could be >2. These dependent weights (0.15 and 2) are comes from the experimental on the terms. All these conditions then the terms will pick to add to the initial query in order to expand the query. For example: the original query number six has five terms which are: (prohibition, appoint, infidel, leader and partner). After calculate the weights for the terms which are in the top results it found that there are three terms which are (surely, take and fight) has a number of Rocchio and IDF weights is >0.15 and TF is >2, so that, these new terms are picked to add to the original query to expand it. At the same time, there is another term which is (believe), also has Rocchio's weight >0.15 and TF is >2 but the IDF of this term is lower than (0.15), so that, this term didn't picked to add to the original query:

$$\overrightarrow{Q_{new}} = \overrightarrow{Q_{0_{xi}}} \cap \overrightarrow{Q_{1_{xi}}}$$

Where:
X_0 = A set of all terms from query Q_0
X_1 = A set of all terms extracted from the top 10 documents retrieved from the query Q_0, if $x_i \in X_1$

Then:

$$tf_{xi} > 2$$

And:

$$idf_{xi} > 0.15$$

And:

$$R_{xi} > 0.15$$

The proposed method presented here utilizes the collection distribution knowledge to refine the initial query. Due to the good generalization ability of the pseudo RF, the most relevant terms are selected automatically.

The effectiveness of the proposed pseudo-relevance feedback improvements is evaluated using the (English text of Qur'an), queries and relevance judgments are based on the research by Rahman. This dataset involves 6236 documents. Each of these has a title Doc XXX_YYY where "Xs" are represented as three bits and it's referred to the number of "Surah" in Qur'an and "Ys are also represented as three bits and it's referred to the number of

"Ayah" in Qur'an. For example, the Doc020_009 this document is (Surah 20 and Ayah 9) and also for Doc104-050 it means the document is (Surah 104 and Ayah 50).

RESULTS AND DISCUSSION

In this study, the proposed term weighting and selection method for modifying the original query of the Rocchio's was compared and evaluated against the baseline which is the probabilistic IR Model using the BM25 scoring function. The proposed method was also compared with the WordNet-based query expansion method. During the experiment only the first and second iterations were considered. The precision, recall and F-measure metrics were used to make the comparisons. During the evaluation, the precision were measured at four different cut-off points which are 10, 20, 30 and 40 for all the 12 queries used.

Precision at various cut-off points: The values of precision at 10 cut-off points are as shown in Table 1 where it shows that AP at top 10 results. AP by using Lucene was (0.036). Before using the RF and after RF, AP values became (0.058) and (0.083) for the first and second iterations, respectively whereas by using WordNet AP was only (0.042). As can be seen, the Average Precision (AP) of the proposed modified pseudo RF are generally higher than the Lucene BM25 weighting scheme (Lucene) and the WordNet (WN) query expansion. The best result was achieved through feedback after two interactions (RF2) for all cases with the exemption of 30 cut-off point where RF1 is better than RF2.

The correct choice of the terms by the proposed method which depends on the Rochio's weight and TF-IDF weighting scheme resulted in the high precision measures. For example: as noted in the standard query number three "God is Most Knowledgeable; God knows matter in the Sky and Earth". This query was got only one relevant document by using Lucene search toolkit. After the expansion by the proposed method, pseudo RF for the first iteration the new terms that was picked are: "Beneficent Allah Heavens" in which the new query became "God is Most Knowledgeable; God knows matter in the Sky and Earth, Beneficent Allah Heavens "for the first iteration the number of relevant documents became 10 relevant documents. For the second iteration of the proposed method pseudo RF, the new terms that was picked are "overtake night Originator Knower " in which the new query became "God is Most Knowledgeable; God knows matter in the Sky and Earth, Beneficent Allah Heavens, overtake night Originator Knower". In this iteration the number of relevant document became 20 relevant documents whereas by using WordNet for expanding the query, the number of relevant documents was only one relevant document. The number of relevant document by using WordNet decreased to less than that of the two iterations by pseudo RF. That means the new terms that added to the original query by using WordNet were incorrect choice because they retrieved new documents but these new documents are non-relevant with this query. Whereas the new terms that added to the original query by using the proposed method pseudo RF was correct choice because these new terms retrieved new relevant documents where these terms are matched with these relevant documents in the corpus.

As for the WordNet, the reason that make the precision low was due to the expansion by WordNet because WordNet is a general domain and not specific domain, so that, the incorrect choice of the Qur'anic terms by WordNet was not successful where it added new terms to the original terms. These new terms are non-relevant with the original query but it was retrieved due to the similarity measure and not on the meaning of the query. For example: query number one by using Lucene toolkit, the number of relevant documents was 3 relevant documents whereas after WordNet expansion the number of relevant document became only one relevant document. This can be attributed to the new terms that added through the expansion by WordNet in which the original query was: "Allah the One and Only, None like unto Him". But when the expanding by using WordNet has been done, the new terms that added to this query are: "Lord People" where the new query became: "Allah the one and only, none like unto Him, Lord people" after WordNet expansion. These new terms when they are added to the original query they retrieved new documents because of the similarity measure but these documents are non-relevant with this query (Table 1).

Recall at various cut-off points: As shown in Table 2, recall at top 10 results before using RF was (0.027) by using Lucene. After using RF, recall was found to be (0.030) for the first iteration and (0.051) for the second iteration. For WordNet, the value of recall was (0.025). Recall increased in the first and the second iteration more than the standard queries and more than the query expansion by WordNet. This could be attributed to the new terms that has been added to the new query through the expansion. These new terms help to retrieve more documents.

F-measure at various cut-off points: As shown in Table 3, F-measure at top 10 results by using Lucene

Table 1: Precision at top (10)

Query number	Initial query	RF (1st)	RF (2nd)	WordNet
Q1	0.300	0.200	0.100	0.200
Q2	0.038	0.100	0.100	0.200
Q3	0.100	0.000	0.200	0.100
Q4	0.000	0.000	0.000	0.000
Q5	0.000	0.000	0.000	0.000
Q6	0.000	0.000	0.200	0.000
Q7	0.000	0.000	0.000	0.000
Q8	0.000	0.000	0.000	0.000
Q9	0.000	0.100	0.000	0.000
Q10	0.000	0.000	0.100	0.000
Q11	0.000	0.300	0.300	0.000
Q12	0.000	0.000	0.000	0.000
AP	0.036	0.058	0.083	0.042

Table 2: Recall at top (10)

Query number	Initial query	RF (1st)	RF (2nd)	WordNet
Q1	0.066	0.044	0.022	0.044
Q2	0.250	0.125	0.125	0.250
Q3	0.011	0.000	0.022	0.011
Q4	0.000	0.000	0.000	0.000
Q5	0.000	0.000	0.000	0.000
Q6	0.000	0.000	0.181	0.000
Q7	0.000	0.000	0.000	0.000
Q8	0.000	0.000	0.000	0.000
Q9	0.000	0.000	0.000	0.000
Q10	0.000	0.000	0.083	0.000
Q11	0.000	0.176	0.176	0.000
Q12	0.000	0.000	0.000	0.000
AP	0.027	0.030	0.051	0.025

Table 3: F-measure at top (10)

Query number	Initial query	RF (1st)	RF (2nd)	WordNet
Q1	0.088	0.266	0.266	0.266
Q2	0.375	0.375	0.375	1.125
Q3	0.011	0.102	0.136	0.034
Q4	0.000	0.000	0.000	0.000
Q5	0.000	0.000	0.000	0.000
Q6	0.181	0.818	1.636	0.000
Q7	0.000	0.000	0.000	0.000
Q8	0.000	0.000	0.000	0.000
Q9	0.000	0.028	0.086	0.000
Q10	0.000	0.000	0.750	0.000
Q11	0.000	1.050	1.050	0.000
Q12	0.000	0.000	0.000	0.000
AP	0.055	0.220	0.358	0.119

toolkit is (0.082) by using RF for the first iteration F-measure became (0.066) and for the second iteration RF, F-measure was (0.153) and by using (WordNet) F-measure was (0.076). F-measure in the second iteration became greater than the standard queries and greater than the query expansion by WordNet. But the first iteration is decreased more than the standard queries this can be attributed to the incorrect choice of the terms that retrieved but it's non-relevant where it's sorted at the top results and the relevant documents is sorted as far as the top result it's caused the decrease of the precision which leads to decrease the F-measure.

Mean average precision: Table 4 shows the Mean Average Precision (MAP) for all the queries used. In this

Table 4: Mean Average Precision (MAP)

Query number	Initial query	RF (1st)	RF (2nd)	WordNet
Q1	0.150	0.275	0.300	0.125
Q2	0.075	0.075	0.075	0.075
Q3	0.025	0.175	0.300	0.025
Q4	0.025	0.050	0.050	0.025
Q5	0.000	0.000	0.000	0.000
Q6	0.075	0.100	0.175	0.050
Q7	0.000	0.000	0.000	0.000
Q8	0.000	0.000	0.000	0.000
Q9	0.000	0.100	0.125	0.000
Q10	0.025	0.050	0.075	0.000
Q11	0.025	0.175	0.250	0.025
Q12	0.000	0.000	0.000	0.000
AP	0.033	0.083	0.113	0.027

table MAP for 12 queries before using RF was (0.033) it is too weak. In order to increase this value, the proposed method applied to expand the original queries to get a higher number of MAP through the first iteration of RF. MAP at the first iteration RF became (0.083). Also, at the second iteration RF MAP became (0.113). Whereas, when the open-domain WordNet used to expand the queries MAP decreased to (0.027) it became less than the standard queries by using Lucene toolkit and less than RF with the two iterations. This can be attributed to the right choice of the new terms by applying the proposed method through RF. And the reason that make the expansion by using WordNet weak is the replacement of the terms of the original query with a new terms. These terms replaced depending on the synonyms dictionary of the open source domain where this domain is general and not specific domain where these a new terms did not have the same compatible with Qur'anic text.

CONCLUSION

The main objective of this study is to implement and examine a retrieval model and its behaviour when relevance feedback is used. Pseudo RF Model was implemented to be suitable for improving the retrieval performance of the Qur'anic text. MP RF is a novel method used in this study. The novelty of this method include expanding the queries using modified equations to pick the new terms. The proposed method has proved its success through the two iterations of the retrieve that has been done. This method was compared with Lucene search engine and WordNet. The reason that make the proposed method successful is the correct choice of the added terms in which these new terms led to retrieve relevant documents through the matching with these documents in the corpus.

Based on the results and evaluation performed by using the evaluation measures: precision, recall and F-measure on this model it can be concluded that MPRF works really well all by itself in extracting most of the

relevant documents. The selection of suitable terms through the expansion of queries led to retrieve more relevant documents that means the recall measure will increase. Also, the position of these documents near from the top results increases the precision measure. For example: by using Lucene search, the average precision AP was only 3.3 and for WordNet it was only 2.7. Whereas after the expansion by MPRF for the first iteration, AP became 8.3 and after the second iteration by MPRF it became 11.3. Also for recall measure it was only 8.2 by using Lucene search and 6.8 for WordNet expansion whereas it increased after using MPRF to be 15.4 and 18.3 for the first and second iteration, respectively. Depends on these numbers of AP and recall measure it can be concluded that the proposed method works well and it is able to improve the retrieval performance of the Qur'anic text.

ACKNOWLEDGEMENT

I would like to thank the anonymous referees for the comments given to this study and to the Universiti Kebangsaan Malaysia for making this research possible.

REFERENCES

Buckley, C., J. Allan and G. Salton, 1993. Automatic Retrieval with Locality Information using SMART. In: The First Text Retrieval Conference TREC-1, Harman, D.K. (Ed.). Diane Publishing, Collingdale, Pennsylvania, pp: 59-72.

Efthimiadis, E., 1996. Query expansion. Annu. Rev. Inform. Syst. Technol., 31: 121-187.

Hasan, S., 2012. People in the Muslim Majority Countries: History, Composition and Issues. In: The Muslim World in the 21st Century, Hasan, S. (Ed.). Springer, Berlin, Germany, ISBN:978-94-007-2633-8, pp: 115-130.

Huang, Y.R., 2006. Using Contextual Information and Machine Learning Technique to Improve Retrieval Performance. ProQuest, iAnn Arbor, Michigan, USA.,.

Metzler, D. and W.B. Croft, 2007. Latent concept expansion using markov random fields. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 23-27, 2007, ACM, New York, USA., ISBN:978-1-59593-597-7, pp: 311-318.

Nasharuddin, N.A., M.T. Abdullah, R.A. Kadir and A. Azman, 2010. A review on the cross-lingual information retrieval. Proceedings of the 2010 International Conference on Information Retrieval & Knowledge Managsement, (CAMP), March 17-18, 2010, IEEE, New York, USA., ISBN:978-1-4244-5650-5, pp: 353-357.

Noordin, M.F. and R. Othman, 2006. An information retrieval system for Quranic texts: A proposed system design. Proceedings of the Conference on Information and Communication Technologies, ICTTA'06, Vol. 1, April 24-28, 2006, IEEE, New York, USA., ISBN:0-7803-9521-2, pp: 1704-1709.

Rocchio, Jr. J.J., 1971. Relevance Feedback in Information Retrieval. In: The Smart Project Experiments in Automatic Document Processing, Salton, G., (Edi.), Prentice-Hall, Englewood Cliffs, New Jersey.

Shoaib, M., M.N. Yasin, U.K. Hikmat, M.I. Saeed and M.S.H. Khiyal, 2009. Relational wordnet model for semantic search in Holy Quran. Proceedings of the International Conference on Emerging Technologies, October 19-20, 2009, IEEE, Islamabad, Pakistan, ISBN: 978-1-4244-5631-4, pp: 29-34.

Sultan, A.M., A. Azman, R.A. Kadir and M.T. Abdullah, 2011. Evaluation of Quranic text retrieval system based on manually indexed topics. Proceedings of the 2011 International Conference on Semantic Technology and Information Retrieval (STAIR), June 28-29, 2011, IEEE, New York, USA., ISBN:978-1-61284-353-7, pp: 156-161.

Yu, S., D. Cai, J.R. Wen and W.Y. Ma, 2003. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. Proceedings of the 12th International Conference on World Wide Web, May 20-24, 2003, ACM, New York, USA., ISBN:1-58113-680-3, pp: 11-18.