MATEC Web of Conferences **135**, 00073 (2017) DOI: 10.1051/matecconf/201713500073
ICME'17

A Survey: Framework of an Information Retrieval for Malay Translated Hadith Document

Nurul Syeilla Syazhween Zulkefli[1,*], *Nurazzah* Abdul Rahman[1], and *Mazidah* Puteh[2]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2]Faculty of Computer and Mathematical Science, Universiti Teknologi MARA, Kampus Dungun Terengganu, Sura Hujung Dungun, Terengganu, Malaysia

Abstract. This paper reviews and analyses the limitation of the existing method used in the IR process in retrieving Malay Translated Hadith documents related to the search request. Traditional Malay Translated Hadith retrieval system has not focused on semantic extraction from text. The bag-of-words representation ignores the conceptual similarity of information in the query text and documents, which produce unsatisfactory retrieval results. Therefore, a more efficient IR framework is needed. This paper claims that the significant information extraction and subject-related information are actually important because the clues from this information can be used to search and find the relevance document to a query. Also, unimportant information can be discarded to represent the document content. So, semantic understanding of query and document is necessary to improve the effectiveness and accuracy of retrieval results for this domain study. Therefore, advance research is needed and it will be experimented in the future work. It is hoped that it will help users to search and find information regarding to the Malay Translated Hadith document.

1 Introduction

Hadith plays an important role as a significant source of reference throughout the world for the Muslims as well as the non-Muslims since it is a second fundamental source in Islam after the holy book, Al-Quran [3]. The actions, characteristics, stories and the sayings of the Prophet Muhammad are forms of Hadith [8,9]. Therefore, this source of reference is used by various parties for various reasons [1]. With the advancement of technology today which serves the platform in searching information more effectively, it is now easy for the users to search for Hadiths through online. Thus, there has been an increasing search activity for this information [12]. Most of the online searching activities on the web are based on the Information Retrieval (IR) framework. IR is the process of "representation, storage, organization of an access to information items" [2]. It aims to provide easy access to information based on the search request. Basically, user need to type in keyword (query text) for the information they want and the system will immediately respond by displaying the information sought in the user screen. In IR query, natural language is usually used and

*Corresponding author: syeilla.syazhween@gmail.com,

text from query and document collection is not always well structured. In retrieving Hadith information, users normally prefer to search for Hadith information in the language that they understand. Nevertheless, the collection of Hadiths that are published either in the form of books or digital resources are mostly in Arabic, English and Indonesian and for other languages are very limited. Therefore, the accurate retrieval of the Hadith information that they want would be difficult. To illustrate, the web offers a variety of application to retrieve relevant information with regards to Hadith in Arabic, English and Indonesian [1]. Even so, in retrieving Hadith information written in Malay language, the IR which was designed for Arabic, English and Indonesian may not work as effectively for the searching of Malay Translated Hadith document [12]. Despite this limitation, little amount of research has been made concerning Hadith searches in Malay language. Based on the past literature, Mutiara Hadis *(http://sigir.uitm.edu.my/webhadis/)* and JAKIM e-Hadith (*http://www, islam.gov.my/portal/e-hadith.php*) are two examples of on-line searches for Malay Translated Hadith documents based on Malay query. Both systems utilise keyword searches but results are significantly limited due to the word used in the query text. In addition, both retrieval systems do not support complex queries. Therefore, this limitation should be reviewed and extended in order to improve the IR in retrieving Hadith information in Malay languages (both query and document retrieved). The IR methods must not only be able to interpret the information, both in the query text and documents content, but also rank them accordingly to the degree of relevance based on user's information need (query). Therefore, this paper offers a review and proposed an improvement of IR framework for Malay Translated Hadith information retrieval to help people who are interested in searching and retrieving Malay Translated Hadith information for knowledge and references.

2 Proposed Framework

In this paper, searching information based on the text of Hadith written in Malay language is focussed. Therefore, this paper has been carried out to review and analyse the limitation of the simple and basic IR framework in retrieving Malay Translated Hadith documents. Based on the limitation, proposed framework is presented and described. Figure 1 below illustrates the basic IR framework and also the proposed framework for further investigation for Malay Translated Hadith document retrieval.

2.1 Dataset

In IR, it is very crucial to prepare a dataset. The dataset that is needed is a list of query text, document collection and a set of relevance judgement by the domain experts. The list of query text must be related to the information needed from the document collection. While, the document collection used must be the ones that have been evaluated by the domain experts (relevance judgement). A set of relevance judgement is referring to the number of relevant documents in a collection for each query. In this paper, a benchmark dataset related to the Shahih Bukhari collection written in Malay language is used for analysis and discussion purposes regarding to the effectiveness of IR in retrieving Malay Translated Hadith Document.

2.2 IR Process

Figure 1 is designed based on the basic IR framework, as presented in [2]. Basic IR framework consists of three steps; (1) Input query text by user; (2) query processor; (3)

retrieval and ranking process of relevant document related to a query. Meanwhile, document processing and indexing process has been done offline, before query is inputted. Due to the limitation of basic IR in retrieving Malay Translated Hadith information (discussion as in Section 3), the proposed framework is presented.In the proposed framework, sematic information extraction method from query and document will be explored. Also, conceptual similarity of information is focused. Therefore, semantic understanding of query, document and collections is necessary in order to improve the effectiveness of IR in retrieving accurate information from Malay Translated Hadith document.

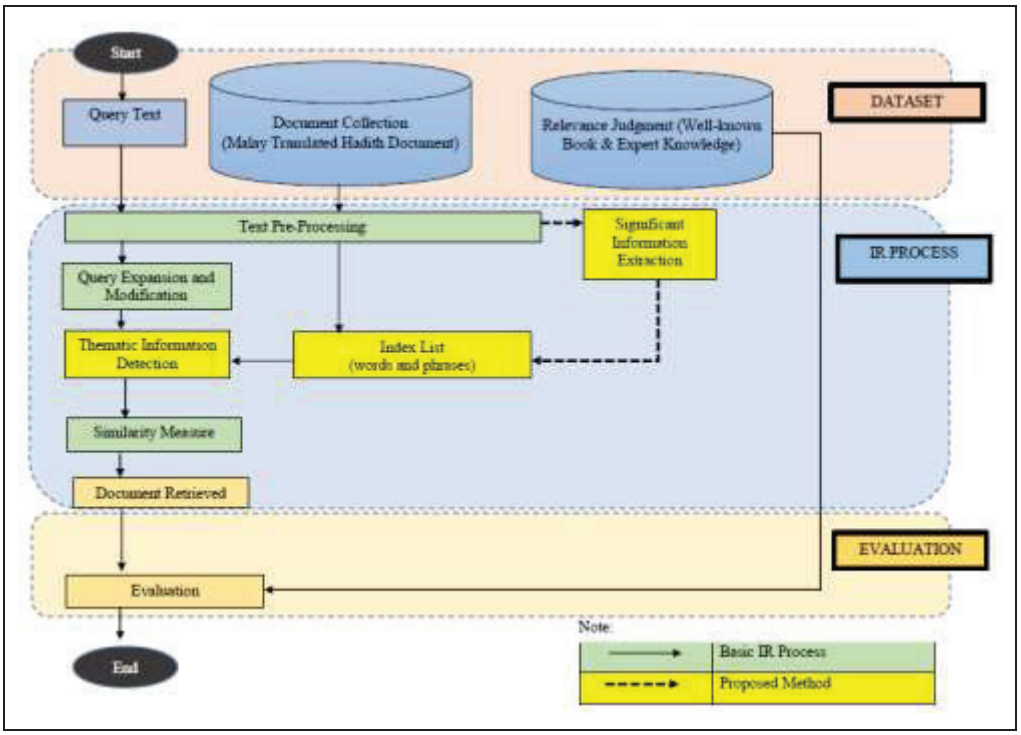

Fig. 1. Proposed framework for Malay Translated Hadith Information Retrieval

2.3 IR Evaluation

A set of documents are retrieved based on the input query text. The documents retrieved for each query is saving for the evaluation purpose. Most of the previous researches evaluate the effectiveness of IR methods using Precision and Recall scores based on the search retrieval results. Precision score measures the likeliness of a document to be relevant to the query while Recall score measures the retrieval of relevant documents. Others measures found in literature include E-Measure, F-Measure and MAP (Mean Average Precision scores for each query). However, this paper only looking for Precision and Recall scores for the review, analysis and discussion purposes.

3 Review and Discussion

This section presents the analysis and discussion of existing method applied for Malay Translated Hadith document retrieval. The analysis is based on the experimental results

from previous studies, which are using similar dataset (list of query to be testing, document collections, and set of relevance judgment).

3.1 Analysis of Retrieval Results

In investigating the effectiveness of IR in retrieving Malay Translated Hadith documents based on query search, the benchmark dataset from Nurazzah [12], consists of 36 numbers of queries, 2028 digital Hadith documents from Sahih Bukhari collection in which written in Malay language and a set of relevance judgement for 36 queries is used. This paper chooses to make a direct comparisons from previous experimental results [10-12]. Therefore five (5) similar queries used from previous studies are selected. There are three (3) methods were used in the experiment: VSM, LSI and Clustering Method. Table 1 illustrates the Precision (P) and Recall (R) scores using these three methods in retrieving Malay Translated Hadith document. A set of document are retrieved based on the input query text. The document retrieved for each query is saved for the evaluation purpose. This paper however only measures the efficiency of the retrieval process using P and R scores.

Table 1. Previous experimental results: Precision (P) and Recall (R) scores for three different methods in retrieving Malay Translated Hadith Document

Query No	Query Text	VSM		LSI		Clustering	
		P	R.	P	R.	P	R.
Q1	*Adab-adab berkaitan makan dan minum*	0.04	0.75	0.07	0.20	0.10	0.39
Q2	*Tuntutlah ilmu hingga ke liang lahad*	**0.50**	0.65	0.07	0.20	**0.50**	0.50
Q3	*Apakah hokum berhias bagi kaum wanita*	**0.01**	0.18	**0.01**	0.27	0.08	0.32
Q4	*Bagaimana cara solat jenazah*	0.07	**0.79**	**0.01**	0.46	0.13	0.33
Q5	*Pembahagian harta mengikut faraid*	0.04	0.68	0.02	0.67	0.17	**0.16**

The results show the low scores for P and R. The highest P is 0.50 which is from Q2 based on VSM and Clustering method. For the R scores, 0.79 is the highest score, which is from Q4 based on VSM. In the other hand, the lowest P score is 0.01, captured from LSI method for Q3 and Q4. The R score using clustering method for Q5 is the lowest score. Meaning that: (1) in the most cases, all the relevant hadith are not retrieved; (2) the sequence of retrieved Hadith does not meet the requirement from the user; and (3) some irrelevant Hadith documents are also retrieved. Therefore, the advance research on IR for this domain study should be extended since searching information from Hadith text is important for Muslims especially to receive their inspirations from the reliable source of their religious beliefs and daily practice from the sayings of their Prophet Muhammad, called Hadith. Due to the importance of Hadith, it is evident that more intelligent searching-based information for Hadith document is needed.

3.2 Limitation of Existing Method

In retrieving Malay Translated Hadith documents, the application enables its users to search through their website based on IR process [12]. The theoretical foundation of Malay Translated Hadith IR is the Vector Space Model (VSM) [10-14]. It is a model to represent the documents content [2]. Baeza-Yates and Ribeiro-Neto (2011 and 1999), Frakes and Baeza-Yates (1992), Jurafsky and Martin (2000), Manning and Schutze (1999) and Van Rijsbergen (1979) are some of the authors that described the IR process and vector space model in their publications well. In the VSM, each query text and Hadith document in the corpus is represented by a vector. Each query and document gets weights based on what words appear in them. These weights model is corresponding to the content of the query and document. They are dependent on how often the word appears in the query, document and in the entire collection. Documents whose vectors are close to each other in this space are considered as similar document content to the query.

Some modification of the vector space model has been proposed in retrieving Malay Translated Hadith documents. Very common changes used are based on Malay stopword list and Malay stemming rules in creating the document representation during the text pre-processing stage. Stopword list is a list of words which does not contribute to the content of the documents such as "*dan*" (and), "*atau"* (or), "*adalah"* (is), etc. The words in the stopword list are simply removed. Malay stemmer in the other hand can describe the Malay morphological in the Malay Translated Hadith documents.

Both stopword list and stemming rules can usefully improve the search retrieval results. Most keywords put in the query text are simple and unstructured. Therefore, it is hard to retrieve a number of relevant documents, where the documents do not contain the few query words. Dealing with this limitation, authors from [14] have expanded user's query using Malay Thesaurus in the process of retrieving Malay Translated Hadith document. The authors come to a conclusion that thesaurus does improve the retrieval effectiveness in this domain although the improvement is very low in percentage.

Latent Semantic Indexing (LSI) approach has also shown to give improvements in the retrieval results for Malay Translated Hadith documents [10]. In LSI process, a more compact representation is presented. Also, the unimportant data can be eliminated. The reduction in the dimension takes related words and documents close to each other.

There has also been proposed research on Malay Translated Hadith documents retrieval using clustering method during indexing process [11,12]. Clustering is an unsupervised learning method, where it is automatically group the documents with similar subject [5,6]. The results are based on the document representation, the similarity measure and the clustering algorithm [2]. In Malay Translated Hadith document clustering, VSM is used to represent the information in the document. Then, cosine similarity measure is reported as an effective method to compute the similarity among Hadith documents [12]. Previous researcher used Agglomerative Hierarchical Clustering Algorithm in the experiments of grouping similar Hadith documents [12]. Among five linkage method, the complete linkage is outperformed [11].

Observed that the existing method used only depends on the words information extraction and it ignores word sequences (phrases) information contained in the query text and document collections. Depending only on the words extraction of information from the query and documents will not fully describe the Hadith contents and this will result in the low quality of retrieval results. Even though, the previous researchers did use various method in retrieving Hadith documents however, there are still issues to be solved. Previous research have not focussed on semantic extraction from query and document. Existing method was unable to handle the issue of conceptual similarity of information from the

query, document and collections. For instance, there are two documents with similar subjects, but both documents contain two distinct contents of information. They are considered as two different documents. Based on the analysis, the reason for the limitation of IR in retrieving Malay Translated Hadith document is due to the complexity of information contained in Malay Translated Hadith document, which is use various of expression to deliver the similar content. Also, the increase of space dimensions while applying the word extraction method has made it hard for the data to be analysed. Moreover, significant information extraction from document is ignored since the method has not considered the word sequences information of documents. Therefore, it seems problematic for IR application to extract the significant information from the collection using the basic and simple method. In addition, the limitation is due to the limited digital resources that capture related information for Malay translated Hadith documents. Presently, Malay language resources are not provided with a large corpus. Furthermore, linguistic database for Malay language that is similar to WordNet is not yet available [16]. Also, the existing digital Malay Thesaurus used by [14] is not completed yet and can be extended.

In addition, it is reviewed that the re-evaluation for the document collections based on the query set is needed. This is because the existing relevance judgement is based on the well-known book [17] only. Existing relevance judgement for each query is based on the topic listed in the reference book [17]. For example, in query text of "*Bagaimana cara solat jenazah*", the researcher will select Hadith document titled as '*solat jenazah*' found from the table of contents in the book and regarded them as relevance judgement for the query. Even though there were other relevant documents contained in the book regarding to this query but, they were categorised under other topics in the book. Thus, the researcher assumed that they were all irrelevant to the query. By merely referring to books of Hadith collection, researchers will unintentionally overlook the relevant documents which are placed under other topics which then will be discarded. Therefore, this paper suggests the re-evaluation process for the relevance judgement. Re-evaluation process involves well-known Hadith book and domain expert knowledge. Both relevance judgements from the two distinct methods are compared. The missing relevant document in benchmark dataset is updated and the redundant relevant document is eliminated.

3.3 Proposed Method

In fact, the retrieval application should be able to understand the contents of the documents and query information need. It is apparent now that the retrieval of Malay Translated Hadith documents is needed to be supported by a more efficient method. The method must not only be able to interpret the content of information from a collection but also be able to identify the related information from the collection based on the search request. The real content of a document is totally different from the words in the vector space model. Thus, this paper designs the proposed framework for Malay Translated Hadith information retrieval. The sections below describe a detailed explanation regarding the methods that are to be utilised in the proposed framework. The proposed methods involved are Significance Information Extraction and Thematic Information Detection.

3.3.1 Significant Information Extraction

Significant information in this paper is referring to the words and phrases, also called as keyphrase [15] which are important and meaningful information for a specific subject in the entire collections. A keyphrase is described as important and significant information as a greatly shortened summary for a document [4]. However, it has received less attention in

the IR study, even though it is an important research topic. Keyphrase extraction is given a more focus from the previous research for journal or conference articles summarizations. Also, the researchers mostly applied for English language dataset [15] in NLP study. This paper however uses Keyphrase Extraction in IR methods. It is mainly used for Malay language dataset. Malay Translated Hadith Document is chooses since it is one of the most important sources of Islamic knowledge and it have no author-assigned keyphrases. Due to its advantage, this paper chooses to exploit Keyphrase Extraction method in the proposed IR framework for Malay Translated Hadith information retrieval. Therefore, query and document content analysis should be observed in future study, so that we can get a criteria and characteristics of significant information from Malay Translated Hadith document.

3.3.2 Thematic Information Detection

In the existing methods, the documents are independent of each other. The word extraction has no interactions for each document in the entire collections. However, some subject-related documents actually contain useful information that can be used as clues to extract significant information from each other. For example, two documents of about the same topic "*Zakat Fitrah*" would share a few common phrases, e.g. "*segantang gandum*", "*segantang kurma*", "*sedekah*", etc. They can provide additional knowledge for each other to better evaluate and extract relevant information from each other in the entire collection. Therefore, given a specific query or document, IR can retrieve a few similar documents in the entire collections through Thematic Information detection. In the literature study, thematic information can be captured based on external resources; corpus-based and knowledge-based resources [16]. However, there are no external resources available for Malay languages [16], and digital resources specifically for Hadith information in Malay language is also not yet available. Therefore, in the proposed framework, this paper chooses to focus on method that will automatically detect the Thematic Information based on Malay Translated Hadith document. Thematic information in this paper is referring to the group of the related information which is significant for a specific subject. The words and phrases are significant if they can deliver the content of query and document in collections [5,6].

4 Conclusion and Future Work

This paper has discussed the limitation of IR methods in retrieving the most relevant Malay Translated Hadith document regarding to the Malay query (information need). This paper then discusses method which aims to improve the effectiveness of IR in retrieving the most relevant information regarding to the Malay Translated Hadith document. The discussion shows that the advance research for IR methods is needed and can be extended. Existing method used in IR framework have not focussed on semantic extraction from query and document. Most of the methods are based on the words extraction from query text and documents. Depending on the words only, the significant information which is consisted of the sequence of words (phrases) is ignored. Also, it does not consider the semantic interpretation of the query text and documents and hence, the subject-related information is discarded. The bag-of-words representation is often unsatisfactory as it ignores the conceptual similarity of words with various expression of information. The limitation has produced low scores for Precision and Recall measures, as it is reported in the experimental results in previous studies which have used the similar dataset. The low scores mean that the current methods applied for IR framework in retrieving information from Malay Translated Hadith document is less effective and can should be reviewed, observed and improved. A more efficient IR framework is needed and the proposed method is presented and discussed to be experimented in future work. The methods involved are semantic

information extraction and thematic information detection. It is hoped that the proposed method is able to improve the effectiveness of IR in retrieving Hadith information from Malay Translated Hadith document in order to help people in searching and retrieving accurate Hadith information written in Malay language.

References

1. S. Abdul Rauf, J. S. Waqar, Information Mining from Muslim Scriptures. *In: International Joint Conference on Natural Language Processing (WSSANLP)* pp. 66-71 (2013)
2. R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval: The concepts and technology behind search* (*Second Edition*), Pearson Education Limited (England: Pearson) Chapter 1-3 pp. 1-130 (2011)
3. J. Dakir, F. A. Shah, A contextual approach in understanding the prophet's hadith. *Applied Sciences*, **8**(7): 3176-3184 (2012)
4. Hammouda KM, Kamel MS (2004) *Efficient phrase-based document indexing for web document clustering. In: IEEE Transactions on knowledge and data engineering*, **16**(10), pp. 1279-1296 (2004)
5. M. Ilic, P. Spalevic, M. Veinovic, Suffix Tree Clustering–Data mining algorithm. *In: Twenty-Third International Electrotechnical and Computer Science Conference (ERK 2014)*, pp. 15-18 (2014)
6. A. K. Jain, S. Maheshwari, Phrase based Clustering Scheme of Suffix Tree Document Clustering Model. *Computer Applications*, **63**(10): 30-37 (2013)
7. K. Jbara, Knowledge discovery in Al-Hadith using text classification algorithm. *American Science*, **6**(11): 409-19 (2010)
8. N. Moath, A. Abdelkarim, A. Musab, O. Abdelrahman, A Lexicon for Hadith Science Based on a Corpus, *Computer Science and Information Technologies*, **6**(2): 1336-1340 (2015)
9. A. S. Mohammad, I. Norisma, M. Rohana, J. Salinah, D. Thorleuchter, G. Abdullah, Hadith data mining and classification: a comparative analysis. *Artificial Intelligence*, 46(1): 113-128 (2016)
10. T. M. R. Amirah, *Latent Semantic Indexing (LSI) Using Parallel Programming Technique For Malay Hadith Translated Document Retrieval*. Unpublished master disertation thesis, Universiti Teknologi Mara (UiTM), Malaysia (2014)
11. N. A. Rahman, Z. A. Bakar, N. S. S. Zulkefli, Malay document clustering using complete linkage clustering technique with Cosine Coefficient. *In: Open Systems (ICOS 2015)*, pp. 103-107: IEEE (2015)
12. N. A. Rahman, *Evaluating The Effectiveness of Clustering Techniques In Retrieving Malay Translated hadith Text.* Unpublished doctoral thesis, Universiti Teknologi Mara (UiTM), Malaysia (2011)
13. N. A. Rahman, Z. A. Bakar, T. M. T. Sembok, N. K. Ismail, Cluster-Based Hadith Retrieval System. *In: Proceedings of the International Conference on ICT for the Muslim World (ICT4M*), Kuala Lumpur (2006)
14. N. A. Rahman, Z. A. Bakar, T. M. T. Sembok, Query Expansion using Thesaurus in Improving Malay Hadith Retrieval System, *In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* pp. 46-54: ACM (2010)
15. I. Sahmoudi, H. Froud, A. Lachkar, A new keyphrases extraction method based on suffix tree data structure for Arabic documents clustering. *Database Management Systems*, **5**(6):17-33 (2014)
16. S. A. Noh, N. Omar, A. Y. Amru, Evaluation of Lexical-Based Approaches to the

Semantic Similarity of Malay Sentences. *Qualitative Linguistics,* **22**(2):135-156 (2015)
17. H. Zainuddin, H. S. Fachruddin, T. Nasharuddin, A. Johar, M. A. A. R. Zainuddin, *Terjemahan Hadis Shahih Bukhari Jilid I,II,III,IV (Cetakan Keenam),* Kuala Lumpur Malaysia: Klang Book Centre, pp. 1-220 (2005)