Journal of Engineering Science and Technology
Vol. 15, No. 1 (2020) 675 - 692

QUERY EXPANSION METHOD FOR QURAN SEARCH USING SEMANTIC SEARCH AND LUCENE RANKING

NUHU YUSUF[1,2], MOHD AMIN MOHD YUNUS[1,*], NORFARADILLA WAHID[1], NAZRI MOHD NAWI[1], NOOR AZAH SAMSUDIN[1], NUREIZE ARBAIY[1]

[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), 86400 Parit Raja, Batu Pahat Johor, Malaysia
[2]Management and Information Technology Department, Abubakar Tafawa Balewa University Bauchi, Bauchi State, Nigeria
*Corresponding Author: aminy@uthm.edu.my

Abstract

Search engines are becoming an instrument for users to search for needed information. The web search engine is one of the most popular search engines that are successfully implemented in many application areas. A major challenge to a web search engine is vocabulary mismatch, specifically term selection and weighting. With the advent of query expansion techniques, the performance of web search engines has been improved in terms of retrieving users needed information. These techniques add additional terms to a query for better search results. The results of these techniques still lack higher precision values. This paper proposed a new hybrid query expansion approach to improve Quran search results using semantic search and Lucene ranking. More specifically, a novel semantic search for Quran search is first presented, in which, Quran search queries are expanded with word synonyms and combined with Quran ontology to get the relationships between concepts within the expanded query. Based on the proposed semantic search, the Lucene ranking algorithm is adopted and modified with stemming, stop words and derivatives to improve the results of query expansion for Quran search. To assess the performance of the proposed query expansion, semantic search and Lucene ranking experiments were conducted using 8 Quran datasets from the Tanzil web site. Overall, the results indicate that the proposed LQA and SSLR on Arberry dataset is superior to other query expansion techniques for Quran search in MAP with 55% and 48% respectively. Future research work should focus on improving Quran ontology and utilizes within a distributed semantic representation

Keywords: Information retrieval, Lucene ranking, Query expansion, Quran search, Semantic search.

1. Introduction

Query expansion is among the primary technique used in search engine performance. Particularly, it is important to many search engine developers for improving performance. Recently, relevant feedbacks and word synonyms are widely used by Singh and Sharan [1] to enhance the search performance of information retrieval applications. However, the word synonyms can only match two words that are the same. This overlooked some word stem and root, which are necessarily the segments of any word. Consequently, this allows researchers to provide an improvement in query expansion methods. It is becoming important to ensure that an improved query expansion method retain reasonable search performance.

Query expansion methods are used to improve search engine performance by expanding the original user query to obtain better search documents results [2]. The major assumption is that the document can be easily retrieved if the user query contains sufficient words that match the documents. Other methods are mostly supplement of the query expansion methods [3, 4], specifically the query expansion based on relevance feedback and the query expansion based on Wikipedia [5]. These signify how the query expansion methods continually attract scholars' interest. The term selections for query expansion are very important aspect prompted researchers to modify existing methods for better search performance.

For instance, the frequency of co-occurrence method of Diao et al. [6] expands the query with words that have a high frequency. In line with Zingla et al. [7], an extended Boolean method with the similarity between the user query and a document can achieve search performance when compared with the results obtained using the likelihood model. It has also been described by Rui-Navas and Miyazaki [8] that better search performance improvement using the query expansion method obtained based on appropriate term selection methods.

Generally, providing a relevance judgment on a given document is quite easy for an expert, but difficult to get the right queries for users. Both relevant judgement and queries are considered in improving the query expansion results. Yusuf et al. [9] compare different queries, which can be used for query expansion. Rasheed and Banka [2] examine a query expansion for Urdu. Singh and Sharan [10] propose a new query expansion approach using fuzzy rule and relevant judgement. Their approach utilizes pseudo-relevance feedback where top documents are assumed to be relevant. Subsequently, Singh and Sharan et al. [11] present a new query expansion method based on pseudo-relevance feedback and word2vec to obtain word similarity and expand the query. The user sends a query and the top documents retrieved will be regarded as a relevant document. Hybrid query expansion was suggested for term selection to improve search performance [12, 13]. Furthermore, Khennak and Drias [14] relates query expansion and accelerated particle swarm optimization. Each of the query expansion methods, aims to provide better search performance. Puspitaningrum et al. [15] proposed a query expansion method based on the work of Puspitaningrum et al. [16].

In query expansion-related work, approaches based on relevant judgements focus on improving the search results, which are relevant to users [1]. Rasheed and Banka [17] examine how the precision results would increase to improve

search engine using specific queries. The paper used relevance feedback given on 52 queries for Urdu dataset. Liu et al. [3] present an approach to query expansion that utilizes word embedding and fuzzy rules. The paper weighted both the original and the expanded query for testing the performance of the search. Furthermore, Ma et al. [4] investigate how to query expansion can work in two different languages. The paper suggested a model map Chines to Mongolian queries. However, Wu et al. [5] investigate information retrieval in health sectors. The paper used the term weighting and filter to expand the original user queries. Ma et al. [4] present term re-weighting approach for improving the performance of search results using query expansion. In addition to that, Zingla et al. [7] propose a query expansion method to address both term generation and selection. The paper used semantic and similarity between terms to measure the search performance. Gupta and Saini [18] described the importance of using support vector machine within search algorithms.

Currently, the term selection assumptions are mainly concentrated on the probability of word occurrences. For instance, Singh and Sharan [11] approach mainly assumed the chance of word occurrence. The term selection depends closely on how words relate to each other. Zhou et al. [19] proposed a personalized query expansion. Also, recently Yusuf et al. [20] proposed query expansion for English Quran based word synonymous to address term selection problems. Still, the query expansion approaches lack precision results. Consequently, our work improves the query expansion method using semantic search and Lucene ranking to improve Quran search engine results.

In this study, we provide the following contributions:

- To obtain the relevant results from the Quran search, this paper proposes a semantic search, which introduces lexical resources and Quran ontology to Quran search query. According to the term's synonyms and their relationship, the query Quran with search is then expanded. Then, the expanded semantic search is combined into Lucene ranking, which synchronizes with derivatives to additionally expand the Query. The Lucene ranking not only improves the results but also expands the query within the index search
- To expand the query of Quran search, four semantic search methods, lexical Quran ontology (LQO), semantic search Lucene ranking (SSLR) and three search ranking algorithms are presented. The LQO expand the query by lexical and Quran ontology expansion. By the mean average precision (MAP) on 8 datasets, the LQO revealed its efficiency and can be applied to the huge dataset. Based on the LQO method, a query expansion ranking algorithm SSLR is proposed to expand the Quran search query and improve the ranking results. SSLR is a search engine ranking algorithm. Similar to LQO, better search results can be achieved in all the evaluation metrics listed.
- To verify the effectiveness of SSLR, and LQO expansion, the MAP is compared with other states of the art methods and algorithms using 8 Quran datasets collected from Tanzil. The results of the experiment show that the proposed query expansion can achieve better search performance.

The rest of the paper is organized as follows. We present materials and methods in Section 2. Experimental results and discussions are also presented in Section 3 and Section 4 contains conclusions and future work.

2. Related Work

2.1. Quran search

A Quran search method can represent Quran text information retrieval using different search criteria. However, the quality of Quran search results was attributed due to vocabulary mismatch. Improving the quality of Quran search method has remained a challenge. For years, many contributions towards Quran search method have been presented. Among them is then presented by Chefrour and Amirat [21], which proposed an auto-completion method for android Quran based on search speed using the Boyer Moore algorithm. However, the paper concentrated mainly on keyword search rather than query search based.

A commonly used search method, WordNet has also been introduced as a search improvement technique in Quran search. Similarly, Quran search using ontology [22] provides better results. Quran ontology helps to clearly understand Quran knowledge. Text categorization was proposed by Rostam and Malim [23], which improve the Quran search by determining the relation between Quran verse and hadith. Support vector machine can successfully improve search results by categorizing the text. The knowledge retrieved from the Quran was evaluated by an expert. The expert can determine the level of search accuracy [24]. Those experts are the best in domain-specific search.

Recently, Beirade et al. [22] developed Quran ontology that is capable of presenting word relationships. This clearly shows that Quran query terms can be expanded based on ontology representation. The present paper focuses on Quran search using a query. Although the keyword-based search is important and regards to the foundation of a query-based search, this paper was excluded. This limitation was due to important of getting relevant information needs within a few retrieved results.

2.2. Query expansion methods

Many query expansion methods are now available for testing the performance of search engines. According to Azad and Deepak [25], the query expansion methods can be categorized into six (6) approaches. Out of these approaches, linguistic-based, concept-based, relevance feedbacks-based and ontology-based are the most commonly used in query expansion. Linguistic-based approaches based on WordNet and Word-sense disambiguation proved effective in query expansion methods. The WordNet has successfully provided the synonyms of words in different languages. Multilingual WordNet [26] allows researchers to compare the word syntax from different languages.

The Bilingual WordNet [27] is also capable of comparing two language word syntax. Also, Alkhatib et al. [28] utilize Arabic WordNet to properly understand some important lessons about Hadith. However, WordNet lacks acceptable translation, especially for Arabic words. Most of the Arabic WordNet collections have been translated by non-Arabic speakers. This hurts the quality of results. Moreover, word-sense disambiguation [29] ascertains the meaning of a word from a particular sentence, especially, a word with multiple meanings.

Lopez-Arevalo et al. [30] described how word-sense disambiguation within the specific domain for particular collections. In addition to that, Pal and Saha [31] present how the exact meaning of the word in Bengal uses Bengal WordNet.

However, different test collection available in a different language makes it difficult to judge the quality of word-sense disambiguation. Word2vec [32] has been used to expand the users' query by removing irrelevant document terms.

To determine the context of word2vec *context(w_i) = w_{i+1}*, Liu et al. [33] provide two ways of word2vec representing: Continuous bag-of-words model (CBOW) and a Skip-gram model. The CBOW can be computed based on Eq. (1) while skip-gram on Eq. (2):

$$L(D) = \frac{1}{T}\sum_{i=1}^{T} log\left(\frac{exp(x_i \cdot x_c)}{\sum w \in W\ exp(x \cdot x_c)}\right) \quad (1)$$

where $\frac{exp(x_i \cdot x_c)}{\sum w \in W\ exp(x_i \cdot x_c)} = pr(w_i / w_{i+k}, \dots,\ w_{i-1}, w_{i+1}, \dots, w_{i+k})$ and $w_{i+k}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+k}$ represent the sequence of the word in document (*D*).

For skip-gram, the following computation should be considered:

$$L(D) = \frac{1}{T}\sum_{i=1}^{T} \sum_{-k \le c \le k, c \ne 0} log\left(\frac{exp(x_{i+c} \cdot x_i)}{\sum w \in W\ exp(x \cdot x_i)}\right) \quad (2)$$

where $= \frac{exp(x_{i+c} \cdot x_i)}{\sum w \in W\ exp(x \cdot x_i)} = pr(w_{i+c} / w_i)$, T; is the training size of the text collection, k; is the context size of the target word, w; is the vocabulary, x_i; is the vector representation of the ith word, x_c is the average vector of all the contextual words, x_{1+c}; is the vector of context word. The differences between the continuous bag-of-word (CBOW) and the skip-gram is that the CBOW represents a word in a document as a generalized concept (bag) without considering the word arrangement or syntaxes in a text. In skip-gram, syntaxes take into consideration to reflect the structure of a word in a sentence. Further explanation of their different computation can be found by Shi et al. [34] their research paper.

The concept-based query expansion approach requires adding additional terms similar to the concept terms used in the query. Afzal and Muktar [35] present a concept search approach for English Quran. The approach improves concept search using WordNet semantic similarities and achieved better recall performance. However, Kolhe and Sawarkar [36] present a semantic concept search using WordNet but utilizing clustering from the concept proposed. Bhargavi and Reddy [37] also used WordNet into concept-based unstructured data representation. Moreover, Tran et al. [38] proposed a concept-based method that utilizes semantic to expand the query.

In terms of relevant feedback approaches, the explicit feedback technique has shown effective in retrieving the quality results from the search engine. In explicit feedback, a query will be sent to the search engine. The expert will be asked to judge the relevance of the results obtained from that query. Also, an expert must have prior knowledge of the domain considered. The results that expert judged most relevant will be given to a system to compute the performance search engine. Liu et al. [39] proposed a method that combines both explicit and implicit feedbacks into consideration. The paper emphasizes on addressing the challenges that may come from the top-*N* recommendation. Jiang et al. [40] suggested the use of multiple explicit feedbacks for improving the relevant judgment in subsequent information retrieval evaluations. Mach et al. [41] evaluate whether to include expert judgment can yield positive results in climate change assessment. The paper invited experts from different domains for their inputs on

the issues. Their result shows that using expert judgment can transparently and consistently improve the quality of the needed results.

The ontology-based approach can be either domain-independent such as vocabulary relations or domain-dependent, which is based domain areas such as Quran, agriculture or law. For domain-dependent ontology, Jiang et al. [42] present how semantic search based on government open data on transport sector can be improved in term of ontology. Bartussek et al. [43] developed risk analysis, ontology to predict any risk associated with medical and health-related devices. Furthermore, Maran et al. [44] use computing ontology to query a relational database. Their work considered context-awareness issues of information system. Wang et al. [45] present how chemical toxicity can effectively assess if domain-specific ontology in chemical has been considered. Beirade et al. [46] propose Quran search engine based on semantic can provide better results when using Quran ontology. Utomo et al. [47] consider using machine learning to develop Quran ontology. In addition to that, Alqahtani and Atwell [48] develop Quran ontology that contains both English and Arabic semantic. Gong et al. [49] proposed query expansion utilizing domain-independent ontology. Their approach utilizes semantic and relevant feedbacks within the query expansion. Sharma et al. [50] use ontology for the semantic web to improve search results.

3. Semantic Search for Quran Search

Precisely, Quran search for information relevant to user need was the primary concern of the query expansion. This section presented a new search method in Quran search, which introduces lexical-semantic and Quran ontology to retrieve Quran search. A good search method should contain many similar terms of the Quran search when the user specifies his/her need through a query. Quran search contains both similar and related words, therefore, lexical-semantic and Quran ontology can fulfil all these needs, and these are introduced to the Quran search. The process of Quran search modification into semantic search is presented.

Semantic search

Based on the previous studies, lexical words serve as the basis for expanding query-based semantic similarity, which addresses term selection challenges. Hence, the lexical words have some restrictions in term of words-contexts as appears in the sentence, which can affect the precision results values. This work proposes a hybrid approach with lexical words and Quran ontology. The proposed approach is called LQO, which can address not only term selection challenges, but also improve precision results. This paper uses lexical words to expand the query and then search the Quran ontology to capture the word's relationship.

Firstly, the paper defines a query (X) and lexical document (Y). The query (X) is assumed as:

$X = (x_1, x_2, ..x_n)$

where $x_1, x_2,..x_n$ represent the query terms. The lexical document (Y) is defined as:

$Y = (y_1, y_2,..y_n)$

where y_1, y_2,..y_n represent lexical terms available in WordNet. According to Eqs. (1) and (2), synonyms of terms in equation 1 will be obtained from Eq. (2) to expand query X. Hence, this query expansion does not take into consideration the context of words. Based on this reason, this paper proposes the utilization of Quran ontology to get the term relationship so that the context of words will be addressed. Instead of taking the relationship between the original queries (X), the expanded query with lexical terms has been used so that words with similar meaning and relationship from Quran ontology can properly provide relevant results.

To identify query word sense during searching, this study uses the WordNet lexical structure of Princeton University [51] for English while other languages used Bond lexical [52]. These provide a necessary relationship between words. In addition, this paper utilizes the ontology of Quranic concepts provided by the language research group, University of Leeds [53].

4. Lucene Ranking

In light of the proposed semantic search method, the new Quran search ranking is proposed in this sub-section. While the result obtained from the semantic search, ranking the expanded query has transformed into a new Quran search Lucene ranking. This sub-section proposes a query expansion method, which considers Lucene structure. The proposed method is semantic search and based ranking (SSLR) with word derivatives, which tokenize, remove stop words, match derivatives words and rank the scores.

Proposed SSLR Lucene ranking

The basic idea behind Lucene ranking in Quran search is to use the semantics of words and address term weighting challenge to retrieve relevant information needed for users based on their query. However, the longer document can easily be ranked above other lower document length. This is because the terms will frequently occur more in longer document lengths. Also, more noise may likely contain in longer documents, which may result in ranking such documents high while are irrelevant. Thus, it is necessary to normalize documents based on their length specifically, as we rank them. The more balanced rating of a term appears in the document. Hence, the matching result is only obtained based on the description of attributes in a document. To appropriately describe document attributes, we introduce expanded query information using word derivatives in Lucene ranking based on semantic search before SSLR ranking. The challenge with calculating SSLR ranking scores is that it requires appropriate semantic similarity measures. This necessitates the need for semantic similarity that would compute the distance between two concepts. For this reason, the cosine similarity used by Baeza-Yatez and Ribeiro-Neto [54] has been adopted, in which, the angle between two vectors is the measure.

To compute cosine similarity, we perform some pre-processing activities to clean and process the dataset. The first pre-processing step is the data cleaning where all the comments on the characteristics of the datasets such as names, translator, language, ID, last update and source are all removed from the dataset. In addition, in this step, all the documents are split, and each line is saved as a document to make a total of 6236 documents. Moreover, stop words are then

removed from these documents and various stemming algorithms have been applied to reduce words in these documents to their stem.

We used Arabic light stemmer [55] to stem both Arabic and Urdu documents. Hausa stemmer [56] to stem Hausa document and sastrawi [57] to stem Malay Bahasa document. Porter Algorithm for English document, Secondly, to process the data we, first of all, generate derivative of our query terms from the WORDAPI [58]. Finally, the cosine similarity is calculated using terms in modified query A and document B is interpreted as in Eq. (3)

$$\sigma(A,B) = \frac{\sum a_i b_i}{\sqrt{a_i^2}\sqrt{b_i^2}} \tag{3}$$

where *a* and *b* represent word vector components of weights assigned to query *A* and weight assigned to document *B*.

The cosine scores obtain can be used to rank the scores using TFIDF, but the irrelevant document can still be generated, and the results will not improve the precision value. Therefore, instead of directly using TFIDF, we incorporate normalization on a very long document in IDF part to avoid negative value specifically if the length of the words in document D $|D|$ is larger than the average length of the document *avgdl* of text collection where document D drew. The TFIDF is defined by Yusuf et al. [59] as in Eq. (4)

$$xy(a,b,C) = tf(a,b) \cdot idf(a,C) \tag{4}$$

where *xy*(*a*,*b*,*C*) represents the term frequency in the inverse document. The number of times term *a* occurs in document *b* is represented as *tf* (*a*,*b*) while the number of the document in corpus *C* where term *a* appears is also represented as *idf*(*a*,*C*). The inverse document frequency can simply compute as in Eq. (5)

$$IDF = \log_2 \frac{N}{DF} \tag{5}$$

where *N* represents the total number of the document in corpus *C* and *DF* represents the number of documents where term *a* appears. If the term did not appear in the document, the value will be zero. Therefore, we normalize the *D* in SSLR algorithm. Integrating normalization reduces the effect of multiple occurrences of the term.

The SSLR algorithm consists of two parts; the first is the term frequency (TF), which computes the number of time term *a* occurs in the document *D*. The second is the inverse document frequency with document length normalization *D*, which measures how important a term *t* appears in a text document collection. Hence, the SSLR score of document *D* and query *Q* is defined as in Eq. (6):

$$SSLR\ Score(D,Q) = \sum_{t \in Q} tf(t,D) * idf(t) * normDLength\ (D) \tag{6}$$

where *tf*(*t*,*D*) is the term *t* frequency in document *D*. the inverse number of documents, in which, *t* term appears is represented as an *idf*(*t*) while *normDLength(D)* is the length normalization factor for document *D*.

Specific Lucene analysers are used to analyse and understand user queries. These analysers include standard analyser and language analysers. A standard analyser is a commonly used analyser, which converts lower cases, removes stop words and remove apostrophes while the language analyser handles many language

texts. EnglishAnalyser, MalayAnalyser, ArabicAnalyser, UrduAnalyser and HausaAnalyser were used in SSLR.

The proposed SSLR is a ranking algorithm, which can be applied in many different search engine or information retrieval application optimizations. Additionally, the proposed SSLR can easily accommodate query expansion of Quran search with different term selection strategies.

Lucene index is responsible for stem documents in a search index. It is referred to the inverted index because it mapped all the query terms to documents and reverses such mapping. The procedure for scoring the search result was also provided by the inverted index. Any document that maps the term is the return as the relevant document.

5. Experimental Results and Discussion

5.1. Datasets

In this experiment, the 6326 verses of the Holy Quran each from 8 different datasets have been used. This paper used Tanzil [60] free datasets for our experiments. The 8 datasets are Classical Arabic, Hausa, Urdu, Malay, Yusuf Ali, Arberry, Sarwar and Modern Arabic. The Classical Arabic clean is the clean copy of the Holy Quran verses without marks such as fat-ha, dhamma, kasra, sukoon, pause marks, sajdah signs as well as rub-el-hizb signs.

Hausa is the Quran translated by Sheikh Abubakar Mohmoud Gumi. He is an Islamic scholar from the northern part of Nigeria. He is Sunni denomination and among the founder of Izala. Urdu is the translated Quran verses from Arabic to the Urdu language. It was translated by Abul Aala Maududi, a British India and Sunni from Pakistan who produces several Islamic books for the benefits of Muslims world.

Lastly, Malay is the Quran translation by Abdullah Muhammad Basmeih who hailed from Yemen and settled in Malaysia. He contributed to writing many Islamic books that now widely used in Malaysia and its neighbours. Yusuf Ali is an English Quran translation dataset by Islamic scholar Abdullahi Yusuf Ali. The scholar was born in India and wrote many Islamic books during his lifetime.

Yusuf Ali dataset gains acceptance by many English-speaking countries. Arberry is also another English version of a Quran translation dataset by Arthur John Arberry. Arthur John Arberry is a non-Muslim scholar but helps to translate the Quran into English.

Sarwar was the third Quran translated dataset by Shaykh Muhammad Sarwar. He is an Islamic scholar from Pakistan who publishes some Islamic books, including the Quran English translation.

Modern Standard Arabic is another dataset of the tafsir Al-muyassar dataset, which is the copy of the Quran translated and printed by the King Fahad Quran Complex, Kingdom of Saudi Arabia.

Table 1 presents the sample of the Arberry dataset where each line represents a verse. The first numerical number of the verse represents a chapter while the second number in between the bars represents verses. All the remaining datasets also had the same structures as Arberry.

Table 1. Sample of Arberry dataset.

Line number	Verses
1	1|1|In the name of God, the Merciful, the Compassionate
2	1|2|Praise belongs to God, the Lord of all being
3	1|3|the All-merciful, the All-compassionate
4	1|4|the master of the day of doom
5	1|5|Thee only we serve; to Thee alone we pray for succour
6	1|6|Guide us in the straight path
7	1|7|the path of those whom Thou hast blessed, not of those against whom Thou art wrathful, nor of those who are astray
8	2|1|Alif Lam Mim
9	2|2|That is the Book, wherein is no doubt, a guidance to the Godfearing
10	2|3|who believe in the unseen, and perform the prayer, and expend of that we have provided them

5.2. Evaluation metrics

Precision and recall [10, 12], and mean average precision [12, 13] are frequently used performance evaluation metrics. Our proposed methods utilize mean average precision (MAP). Based on Moawad et al. [12], Eq. (7) present how precision can be computed:

$$MAP = \sum_{i=1}^{NQ} Avg(q_i) \tag{7}$$

Mean average precision uses rankings from different users' queries and then average them to obtain average precision. MAP will be a good measure to our proposed approach because large datasets and many queries are presented. This meant issues that might arise if some documents are relevant to some queries and few relevant to others.

5.3. Evaluation metrics

This section presents experimental results obtained with 8 datasets using mean average precision (MAP). To understand whether the results obtained are relevant or irrelevant to our queries, the paper used 70 queries out of them, 36 were adopted by Yusuf et al. [59] Quran relevant judgement as our benchmark. The proposed approach is set to retrieve the results within the first 100 documents.

The proposed SSLS will be compared against different traditional searching algorithm results for retrieving Quran search.

5.3.1. Semantic search results and discussion

This part show and discuss the results of semantic search experiments. The semantic search results are compared with other states of the art semantic methods to evaluate its performance. These methods include lexical resource, word morphology, domain-based independent ontology and Quran domain-based ontology. In these methods, the Quran with semantic search is used to expand the query.

Table 2 compares the MAP results of each expanded query. Generally, different methods can perform better on different datasets. The proposed LQO achieve best results on 5 datasets with 55% as the highest results on Arberry dataset. The results

on the remaining 3 datasets show that the proposed LQO still does not perform worse because the results are slightly below them. In terms of datasets languages categorization, the proposed LQO performs best on all the three datasets namely; Yusuf Ali, Arberry and Sarwar. For other languages, it achieves the best results in Malay and modern standard Arabic. Therefore, the proposed semantic search can significantly improve the performance of Quran search as compared with other semantic methods in terms of MAP.

Figure 1 shows the MAP on 8 datasets by making use of morphology, lexical resources, domain-independent ontology, Quran domain ontology and proposed LQO. The proposed LQO is superior in retrieving relevance results on Yusuf Ali. It achieves 51% improvements as compared to morphology with 33%. Though, it performs better when compared lexical resources with 45%, domain-independent ontology with 31% and Quran domain ontology with 31%. For Sarwar dataset, the results show that the proposed LQO achieve 49%, morphology 31%, lexical resources 42%, domain-independent ontology 39% and Quran domain ontology with the least by 29%. The results indicate that the proposed LQO and lexical resources can perform better on the Sarwar dataset as compared to other methods.

Table 2. MAP performance comparison for semantic search.

Datasets	Morphology (%)	Lexical resource (%)	Domain-based independent ontology (%)	Quran domain-based ontology (%)	Proposed LQO (%)
Yusuf Ali	33	45	42	31	51
Sarwar	31	42	39	29	49
Arberry	32	42	39	26	55
Malay	35	25	19	28	47
Urdu	15	18	17	19	32
Hausa	17	16	19	22	36
Classical Arabic	40	17	22	57	37
Modern Arabic	35	38	33	25	48

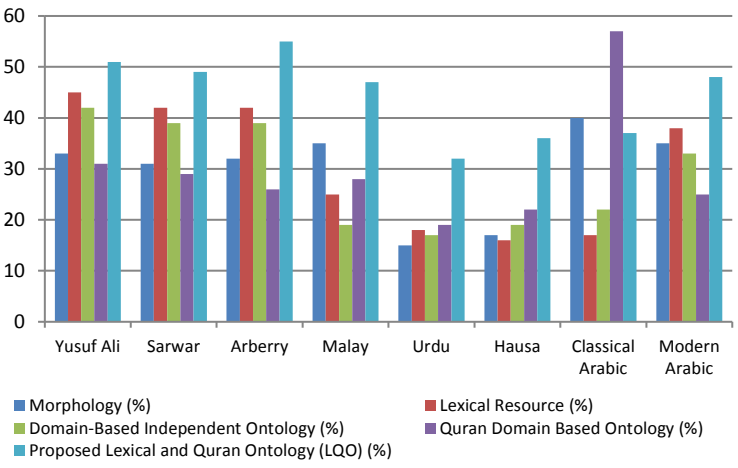

Fig. 1. MAP performance results for semantic search.

The results obtained for Arberry dataset have some similarities with the Yusuf Ali dataset on the MAP. However, the results slightly have little differences and it improves search performance by achieving 55% as compared to lexical resources with 42% and domain-independent ontology with 39%. The least result was achieved on Quran domain ontology. The Malay dataset shows that the proposed LQO is significant on MAP as compared to morphology, lexical resources, domain-independent ontology and Quran domain ontology. Such a result shows that Quran domain ontology alone cannot provide significant on Malay dataset. For Urdu dataset, the results achieved with morphology, lexical resources, domain-independent ontology, Quran domain ontology and proposed LQO. When comparing the results, the search engine has 32% performances when applying proposed LQO, which is greater than the morphology, lexical resources, domain-independent ontology and Quran domain ontology with 15%, 18%, 17% and 19% respectively.

The MAP results on Hausa dataset show that the proposed LQO is also superior in retrieving relevant results. It achieves 36% improvements as compared to morphology with 17%. Though, it performs better when compared to lexical resources. The classical Arabic dataset shows the proposed LQO is inferior in retrieving relevant documents as compared with Quran domain ontology with 57% achieve results. Such results can be applied in a practical situation as classical Arabic remains the actual Quran language. The proposed LQO achieved 37%, while lexical resources and domain-independent ontology achieve 17% and 22% respectively. Lastly, the MAP results obtained using modern standard Arabic dataset show that the proposed LQO method achieved significant improvement over other semantic search methods. It achieves 48% results, performance compared with lexical resources with 38%, morphology with 35% and Quran domain ontology with the least results of 25%.

5.2.2. Lucene ranking results and discussion

This section presents the experiment conducted on Lucene ranking based on the semantic search. The proposed Lucene ranking was compared with the states of the art ranking algorithms. These algorithms are BM25, TFIDF and Lucene. Table 3 compares the results of four ranking algorithms. Better performance was still achieved on Arberry dataset with 48% performance improvement on the MAP. In addition, the proposed SSLR achieve better performance almost on all the datasets except classical Quran.

Table 3. MAP performance comparison for Lucene ranking.

Datasets	BM25 (%)	TFIDF (%)	Lucene (%)	Proposed Semantic Search with Lucene Ranking (SSLR) (%)
Yusuf Ali	27	23	11	33
Sarwar	22	14	19	29
Arberry	20	11	27	48
Malay	10	12	22	30
Urdu	12	11	21	31
Hausa	4	4	11	13
Classical Arabic	21	31	31	6
Modern Arabic	25	35	18	30

Figure 2 shows the MAP on 8 datasets by making use of BM25, *tf-idf*, Lucene and proposed SSLR. The proposed SSLR is superior in retrieving relevance results on Yusuf Ali. It achieves 33% improvements as compared to BM25 with 27%. Though, it performs better when compared TFIDF with 23% and Lucene with 11%. In Sarwar dataset, 19% has been achieved on Lucene as compared to the proposed SSLR with 29% results. The Arberry dataset results show the proposed SSLR is significant on both MAP as compared to BM25, TFIDF and Lucene traditional methods. The proposed achieves 48% results as compared to BM25 with 20%. The Malay dataset results also show that the proposed SSLR is significant and obtained 30% results as compared to Lucene with 22%, TFIDF with 12% and BM25 with 10%.

Moreover, the MAP on Urdu dataset by making use of BM25, TFIDF, Lucene and proposed SSLR show that the proposed SSLR is performing best in retrieving relevance results. It achieves 31% improvements as compared to BM25 with 12%. Though, it performs better when compared TFIDF with 11% and Lucene with 21%. The proposed SSLR is superior in retrieving relevance MAP results on Hausa dataset. It achieves 13% improvements as compared to BM25 with 4%. Though, it performs better when compared TFIDF with 4% and Lucene with 11%. In addition, better results have been shown for the proposed SSLR is inferior as compared to the other three ranking methods on classical Arabic, which achieve 6% results. Lastly, the MAP result on Modern Arabic dataset show better results for both the proposed SSLR and TFIDF rankings. The MAP results obtained show the percentage that our proposed approach can retrieve documents within 100 ranked lists.

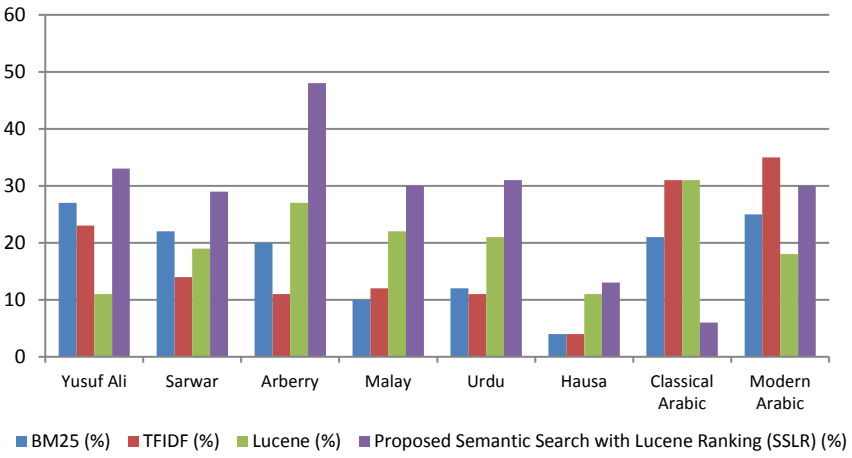

Fig. 2. MAP performance results for ranking algorithms.

6. Conclusions

This paper proposed a new query expansion method that utilizes semantic search and Lucene ranking to improve Quran search results.

The main contribution of the proposed approach is to obtain the relevant results from the Quran search. This paper proposes a semantic search, which introduces lexical resources and Quran ontology to Quran search query. According to the

term's synonyms and their relationship, the query Quran with search is then expanded. Then, the expanded semantic search is combined into Lucene ranking, which synchronizes with derivatives to additionally expand the Query. The Lucene ranking not only improves the results but also expands the query within the index search. The second contribution is to expand the query of Quran search, four semantic search methods, lexical Quran ontology (LQO), semantic search Lucene ranking (SSLR) and three search ranking algorithms are presented. The LQO expand the query by lexical and Quran ontology expansion. By the mean average precision (MAP) on 8 datasets, the LQO revealed its efficiency and can be applied to the huge dataset. Based on the LQO method, a query expansion ranking algorithm SSLR is proposed to expand the Quran search query and improve the ranking results. SSLR is a search engine ranking algorithm. Similar to LQO, better search results can be achieved in all the evaluation metrics listed.

The third contribution is to verify the effectiveness of SSLR, and LQO expansion, the MAP is compared with other states of the art methods and algorithms using 8 Quran datasets collected from Tanzil. The results of the experiment show that the proposed query expansion can achieve better search performance. From the research experiments, it was shown how words that are exactly or nearly the same use to improve search results. It was also observed that relevant documents have used different terms. In contrast, the search engine must have a distributed representation of term with semantic metadata so that the meaning of a word can be better processed using machine learning algorithms such as neural network and therefore, a beneficial to improve the performance of query expansion for better results in future.

Acknowledgement

This research project has been sponsored by research fund UTHM, E15501 for financially supporting this research and research fund of RMC Vot E15501, Universiti Tun Hussein Onn Malaysia (UTHM) and grant Tier 1 vote no. U898, Enhancing Quran Translation in Multilanguage using Indexed References with Fuzzy Logic.

References

1. Singh, J.; and Sharan, A. (2018). Rank fusion and semantic genetic notion based automatic query expansion model. *Swarm and Evolutionary Computation*, 38, 295-308.
2. Rasheed, I.; and Banka, H. (2018). Query expansion in information retrieval for Urdu language. *Proceedings of the Fourth International Conference on Information Retrieval and Knowledge Management*. Kota Kinabalu, Malaysia, 171-176.
3. Liu, Q.; Huang, H.; Lut, J.; Gao, Y.; and Zhang, G. (2017). Enhanced word embedding similarity measures using fuzzy rules for query expansion. *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Naples, Italy, 1-6.
4. Ma, L.; Lai, W.; Guo, L.; and Zhao, X. (2018). Mongolian-Chinese cross-language query expansion based on cross-language word vectors. *Proceedings of the International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*. Changsha, China, 196-200.

5. Wu, H.; Li, J.; Kang, Y.; and Zhong, T. (2018). Exploring noise control strategies for UMLS-based query expansion in health and biomedical information retrieval. *Journal of Ambient Intelligence and Humanized Computing*, 1-12.
6. Diao, L.; Yan, H.; Li, F.; Song, S.; Lei, G.; and Wang, F. (2018). The research of query expansion based on medical terms reweighting in medical information retrieval. *EURASIP Journal on Wireless Communications and Networking*, 105, 7 pages.
7. Zingla, M.A.; Latiri, C.; Mulhem, P.; Berrut, C.; and Slimani, Y. (2018). Hybrid query expansion model for text and microblog information retrieval. *Information Retrieval Journal*, 21(4), 337-367.
8. Ruiz-Navas, S.; and Miyazaki, K. (2018). A complement to lexical query's search-term selection for emerging technologies: The case of big data. *Scientometrics*, 117(1), 141-162.
9. Yusuf, N.; Yunus, M.A.M.; and Wahid, N. (2019). A comparative analysis of web search query: informational vs. navigational queries. *International Journal on Advanced Science, Engineering and Information Technology*, 9(1), 136-141.
10. Singh, J.; and Sharan, A. (2017). A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. *Neural Computing and Applications*, 28(9), 2557-2580.
11. Singh, J.; and Sharan, A. (2015). Co-occurrence and semantic similarity based hybrid approach for improving automatic query expansion in information retrieval. *International Conference on Distributed Computing and Internet Technology*, 415-418.
12. Moawad, I.; Alromima, W.; and Elgohary, R. (2018). Bi-gram term collocations-based query expansion approach for improving Arabic information retrieval. *Arabian Journal of Science and Engineering*, 43(12), 7705-7718.
13. Khatoon, T.; and Govardhan, A. (2018). Query expansion with enhanced-bm25 approach for improving the search query performance on clustered biomedical literature retrieval. *Journal of Digital Information Management*, 16(2), 85-98.
14. Khennak, I.; and Drias, H. (2017). An accelerated PSO for query expansion in web information retrieval: application to medical dataset. *Applied Intelligence*, 47(3), 793-808.
15. Puspitaningrum, D.; Yulianti, G.; and Prasetya, I.S.W.B. (2017). Wiki-MetaSemantik: A wikipedia-derived query expansion approach based on network properties. *Proceedings of the 5th International Conference on Cyber and IT Service Management (CITSM)*. Denpasar, Indonesia, 1-6.
16. Puspitaningrum, D. (2015). An MDL-based frequent itemset hierarchical clustering technique to improve query search results of an individual search engine. *Proceedings of the 11th Asia Information Retrieval Societies Conference*. Bribane, Australia, 279-291.
17. Rasheed, I.; and Banka, H. (2018). Query expansion in information retrieval for Urdu language. *Proceedings of the Fourth International Conference on*

Information Retrieval and Knowledge Management. Kota Kinabalu, Malaysia, 1-6.

18. Gupta, Y.; and Saini, A. (2017). A novel fuzzy-PSO term weighting automatic query expansion approach using combined semantic filtering. *Knowledge-Based Systems*, 136, 97-120.
19. Zhou, D.; Wu, X.; Zhao, W.; Lawless, S.; and Liu, J. (2017). Query expansion with enriched user profiles for personalized search utilizing folksonomy data. *IEEE Transactions on Knowledge and Data Engineering*, 29(7), 1536-1548.
20. Yusuf, N.; Yunus, M.A.M.; and Wahid, N. (2019). Query expansion based on explicit-relevant feedback and synonyms for English Quran translation information retrieval. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(5), 227-234.
21. Chefrour, D.; and Amirat, A. (2019). Using fast string search for Quran text auto-completion on Android. *Advances in Computing Systems and Applications*, 92-101.
22. Beirade, F.; Azzoune, H.; and Zegour, D.E. (2019). Semantic query for Quranic ontology. *Journal of King Saud University-Computer and Information Sciences*. (in press)
23. Rostam, N.A.P.; and Malim, N.H.A.H. (2019). Text categorisation in Quran and Hadith: Overcoming the interrelation challenges using machine learning and term weighting. *Journal of King Saud University-Computer and Information Sciences*. (in press)
24. Safee, M.A.M.; Saudi, M.M.; Pitchay, S.A.; Ridzuan, F.; Basir, N.; Saadan, K.; and Nabila, F. (2018). Hybrid search approach for retrieving medical and health science knowledge from Quran. *International Journal of Engineering and Technology*, 7(4.15), 69-74.
25. Azad, H.K.; and Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing and Management*. 56(5), 1698-1735.
26. Franco-Salvador, M.; and Leiva, L.A. (2018). Multilingual phrase sampling for text entry evaluations. *International Journal of Human-Computer Studies*, 113, 15-31.
27. Goikoetxea, J.; and Agirre, E. (2018). Bilingual embeddings with random walks over multilingual wordnets. *Knowledge-Based Systems*, 150, 218-230.
28. Alkhatib, M.; Monem, A.A.; and Shaalan, K. (2017). A rich Arabic wordnet resource for Al-Hadith Al-Shareef. *Procedia Computer Science*, 117, 101-110.
29. Saif, A.; Omar, N.; Zainodin, U.Z.; and Ab Aziz, M.J. (2018). Building sense tagged corpus using wikipedia for supervised word sense disambiguation. *Procedia Computer Science*, 123, 403-412.
30. Lopez-Arevalo, I.; Sosa-Sosa, V.J.; Rojas-Lopez, F.; and Tello-Leal, E. (2017). Improving selection of synsets from WordNet for domain-specific word sense disambiguation. *Computer Speech and Language*, 41, 128-145.
31. Pal, A.R.; and Saha, D. (2017). Word sense disambiguation in Bengali: An unsupervised approach. *Proceedings of the 2nd IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. Coimbatore, India, 1-5.

32. Damiano, E.; Minutolo, A.; Silvestri, S.; and Esposito, M. (2017). Query expansion based on wordnet and word2vec for Italian question answering systems. *Proceedings of the International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. Barcelona, Spain, 301-313.
33. Liu, Q.; Huang, H.; Lut, J.; Gao, Y.; and Zhang, G. (2017). Enhanced word embedding similarity measures using fuzzy rules for query expansion. *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Naples, Italy, 1-6.
34. Shi, Y.; Zheng, Y.; Guo, K.; and Li, W. (2018). Word similarity fails in multiple sense word embedding. *Proceedings of the International Conference on Computational Science*. Wuxi, China, 489-498.
35. Afzal H.; and Mukhtar, T. (2019). Semantically enhanced concept search of the Holy Quran: Qur'anic English wordnet. *Arabian Journal for Science and Engineering*, 44(4), 3953-3966.
36. Kolhe, S.R.; and Sawarkar, S.D. (2017). A concept driven document clustering using WordNet. *Proceedings of the International Conference on Nascent Technologies in Engineering*. Navi, Mumbai, India, 1-5.
37. Bhargavi, C.; and Reddy, A.B. (2018). Transforming unstructured data into conceptual representation using WORDNET. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3(1), 389-395.
38. Tran, M.-T.; Dinh-Duy, T.; Truong, T.-D.; Vo-Ho, V.-K.; Luong, Q.-A.; and Nguyen, V.-T. (2018). Lifelog moment retrieval with visual concept fusion and text-based query expansion. *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF)*. Avignon, France, 1-12.
39. Liu, M.; Pan, W.; Liu, M.; Chen, Y.; Peng, X.; and Ming, Z. (2017). Mixed similarity learning for recommendation with implicit feedback. *Knowledge-Based Systems*, 119, 178-185.
40. Jiang, J.; He, D.; and Allan, J. (2017). Comparing in situ and multidimensional relevance judgements. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Shinkuju, Tokyo, Japan, 405-414.
41. Mach, K.J.; Mastrandrea, M.D.; Freeman, P.T.; and Field, C.B. (2017). Unleashing expert judgment in assessment. *Global Environmental Change*, 44, 1-14.
42. Jiang, S.; Hagelien, T.F.; Natvig, M.; and Li, J. (2019). Ontology-based semantic search for open government data. *Proceedings of the IEEE 13th International Conference on Semantic Computing (ICSC)*. Newport Beach, California, United States of America, 7-15.
43. Bartussek, W.; Weiland, T.; Meese, S.; Schurr, M.O.; Leenen, M.; Uciteli, A.; Kropt, S.; Heere, H.; Goller, C.; Blohm, P.; Lauer, W.; and Seidel, R. (2018). Ontology-based search for risk-relevant PMS data. *Proceedings of the 3rd Biennial South African Biomedical Engineering Conference (SAIBMEC)*. Stellenbosch, South Africa, 1-4.
44. Maran, V.; Machado, A.; Machado, G.M.; Augustin, I.; and de Oliveira, J.P.M. (2018). Domain content querying using ontology-based context-awareness in information systems. *Data and Knowledge Engineering*, 115, 152-173.

45. Wang, R.-L.; Edwards, S.; and Ives, C. (2019). Ontology-based semantic mapping of chemical toxicities. *Toxicology*, 412, 89-100.
46. Beirade, F.; Azzoune, H.; and Zegour, D.E. (2019). Semantic query for Quranic ontology. *Journal of King Saud University-Computer and Information Sciences.* (in press)
47. Utomo, F.S.; Suryana, N.; and Azmi, M.S. (2019). New instances classification framework on Quran ontology applied to question answering system. *TELKOMNIKA*, 17(1), 139-146.
48. Alqahtani, M.M.; and Atwell, E. (2018). Developing bilingual Arabic-English ontologies of Al-Quran. *Proceedings of the 2nd IEEE International Workshop on Arabic and Derived Script Analysis and Recognition*. London, United Kingdom, 96-101.
49. Gong, H.; Du, J.; and Wang, W. (2018). Microblog query expansion based on ontology expansion and borda count rank. *Proceedings of the Chinese Intelligent Systems Conference*. Wenzhou, China, 271-280.
50. Sharma, S.; Kumar, A.; and Rana, V. (2018). Ontology based informational retrieval system on the semantic web: Semantic web mining. *Proceedings of International Conference on Next Generation Computing and Information Systems (ICNGCIS)*. Jammu, India, 3 pages.
51. Miller, G.A. (1995). WordNet: A lexical database for English. *Communication of the ACM*, 38(11), 39-41.
52. Bond, F.; and Foster, R. (2013). Linking and extending an open multilingual wordnet. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.* Sofia, Bulgaria, 1352-1362.
53. Dukes, K. (2017). Ontology of Quranic concept. Retrieved June 30, 2019, from http://corpus.quran.com/ontology.jsp.
54. Baeza-Yatez, R.; and Ribeiro-Neto, B. (1999). *Modern information retrieval* (1st ed.). New York, United States of America: ACM Press.
55. Al-Omari, A.; and Abuata, B. (2014). Arabic light stemmer (ARS). *Journal of Engineering Science and Technology (JESTEC)*, 9(6), 702-716.
56. Bimba, A.T.; Idris, N.; Khamis, N.; and Noor, N.F. (2016). Stemming Hausa text: Using affix-stripping rules and reference look-up. *Language Resources and Evaluation*, 50(3), 687-703.
57. Robbani, H.A. (2019). Sastrawi. Retrieved June 15, 2019, from https://pypi.org/project/Sastrawi/.
58. WORDAPI. (2019). An API for the English language. Retrieved June 17, 2019 from http://www.wordsapi.com.
59. Yusuf, N.; Yunus, M.A.M.; and Wahid, N. (2019). Arabic text stemming using query expansion method. *Proceedings of the 4th International Conference of Reliable Information and Communication Technology.* Johor, Malaysia, 3-11.
60. Tanzil. (2019). Quran translations. Retrieved May 5, 2019, from http://tanzil.net/trans/.