A proposed semantic keywords search engine for Indonesian Qur’an translation based on word embedding

Liza Trisnawati1,2, Noor Azah Binti Samsudin2, Shamsul Kamal Bin Ahmad Khalid2, Ezak Fadzrin Bin Ahmad Shaubari2, Sukri1,2, Zul Indra3
1Department of Informatics Engineering, Universitas Abdurrab, Pekanbaru, Indonesia
2Department of Software Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Malaysia
3Department of Computer Science, Universitas Riau, Pekanbaru, Indonesia

ABSTRACT
Obtaining relevant information from the Holy Qur’an can be really challenging for people who cannot speak Arabic, such as the Indonesian people. One technology implementation which is commonly used to tackle this problem is to develop a search engine application for Al-Qur’an verses. This paper proposes a search engine based on semantic representation keywords for the Indonesian translation of the Al-Qur’an which consists of 3 phases i.e., data preparation, document representation, and search engine development. In the first stage, the Al-Qur’an dataset was built using the official translation of the Al-Qur’an from the Ministry of Religion and then enriched with the Wikipedia corpus. The second phase is document representation which produces feature vectors by utilizing the Word2Vec algorithm. Finally, the development of a search engine that can find the most relevant verses by calculating the cosine similarity between the document and the keywords. It was found that the proposed search engine succeeded in exceeding the performance of ordinary search engines by finding wider information due to the use of semantic keywords. Apart from that, the proposed search engine succeeded in maintaining the relevance of search results by achieving precision and recall levels of 98.7% and 97.3% respectively.

Keywords: Al-Qur’an, Search engine, Semantic keywords, Word embedding, Word2vec

This is an open access article under the CC BY-SA license.

Corresponding Author:
Noor Azah Samsudin
Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM)
Parit Raja, Batu Pahat, Malaysia
Email: azah@uthm.edu.my

1. INTRODUCTION
The Holy Qur’an is one of the most important things for Muslims in carrying out their lives as good Muslims. In the view of Islam, if someone is willing to be a good Muslim, they must be able to understand the contents of the Qur’an and underlie all of their activity to the guidance that is in the Qur’an. This is because the Qur’an is one of the two main sources of law which is a guide for Muslims in living their lives. Qur’an contains things that are related to faith, science, law, muamalah (rules governing behavior and procedures for human life), stories of people before, worship and, tajwid. As a document, the Holy Qur’an is a scripture that is very long and rich in content where it consists of 30 juz (part), 114 surahs, and 6236 verses and is written in Arabic. This can be really challenging for most Muslims who are not proficient in Arabic, for example, Indonesian Muslims who speak in Indonesian language, to understand it. It will require a lot of time for them to find the appropriate information in the Qur’an if they have to refer to the Qur’an for certain problems. Hence, developing technology to make it easier for Muslims to discover knowledge contained in
the Qur’an is one of the efforts that is worth to be done [1]. One of the technology implementations that is commonly used to assist Qur’an knowledge discovery is developing a search engine application for Al-Qur’an verses [2].

Thus far, a lot of research has been carried out to help search for information in the Al-Qur’an by developing search engine applications. In the last 5 years, it was found that several related studies had worked on this topic. Even though a lot of research has been carried out, there is a big gap in existing research regarding the inability of the developed search application to understand and search for implicit (contextual) information in the Al-Qur’an. The majority of existing search engines only apply the concept of labeling [3], indexing and ranking as well as question answering textually without trying to understand verses contextually [2], [4]-[6]. This contextual understanding is very important in understanding the information contained in it since the Holy Al-Qur’an contains many implied meanings which is the biggest challenge in finding relevant information [7], [8].

One solution that can be applied to tackle this issue is to represent the Qur’anic text by paying attention to the semantic relationships that exist in each verse. Based on an in-depth study of the literature, it was found that only a few studies had tried this approach [9]-[11]. Moreover, the application of text representation based on semantic relationships in case studies of translations of the Indonesian Al-Qur’an can be said to be very rarely discussed. Existing research only applies a regular keyword-based approach [12] or keywords enriched with a Glossary [13]. Only a few studies have applied semantic-based document representation as carried out by Purnama et al. [14]. However, this research has not implemented dataset enrichment during document representation phase in order to obtain semantic keywords. To be concluded, the summary of the research gap and contribution offered by this research is illustrated in Figure 1.

![Figure 1. Summary of research gap and research contribution](image-url)

Based on Figure 1, it is hoped that this research can bridge those research gaps by developing a semantic search engine application. This proposed application is expected to be able to understand semantic relationships to understand Qur’anic verses textually and contextually by generating semantic keywords. In particular, this search engine application is intended for Indonesian translations of Qur’anic verses because Indonesians have limitations in understanding Arabic documents which is an absolute requirement to obtain information from the Al-Qur’an.

2. METHOD

As stated in the introduction section, this study aims to facilitate the search for knowledge in the Al-Qur’an which contains a lot of implied information. However, this study discovered that there are three main challenges to achieve this goal (1) how to prepare the data, (2) how to represent data in semantic relationship, and (3) how to develop proposed semantic search engine for Al-Qur’an. Therefore, this study is divided into three stages to overcome these challenges i.e., data preparation, creating semantic relationships, and developing semantic search engine applications. The overall architecture of the research method to conduct this study is illustrated in Figure 2.
2.1. Data preparation stage

As in general research related to data processing and information retrieval, the first stage in this research is data collection and preparation. As mentioned previously, the first problem in this research is how to prepare quality data for the Indonesian translation of the Al-Qur’an. To answer this problem, this research uses a translation of the Al-Qur’an sourced from the Indonesian Ministry of Religion. This translation can be considered the official and most valid translation of the Indonesian Al-Qur’an. Once the process of collecting the Indonesian Al-Qur’an dataset is conducted completely, the next task in the data preparation stage is augmenting the dataset. Enriching this dataset is very crucial for the next stage due to creating semantic relationships between texts requires a very large number of datasets.

2.2. Semantic text representation stage

The next stage of this research is the representation of Al-Qur’an verses based on their semantic relationship for the Al-Qur’an dataset that has been enriched in the first stage. Semantic relationships in a text refer to how words or phrases in a text interact meaningfully [15], [16]. Furthermore, these semantic relationships include the way words are related to each other and how the meaning of one part of the text can influence or be related to the meaning of other parts. In the field of natural language processing studies, especially in efforts to create a search engine for Al-Qur’an verses which contain many implied meanings, understanding semantic relationships is very important. Through this semantic relationship, a deeper understanding of the documents which are examined can be obtained [17]-[19], such as understanding implied meaning (contextual understanding), analogical reasoning, and increasing similarity calculations.

One method that is widely recommended for finding semantic relationships between words is to apply a word embedding algorithm [20], [21]. Semantic relationships in word embeddings refer to the way words are related or connected in meaning in a given vector space. This embedding captures semantic relationships by placing words with similar meanings close to each other in this space. Many algorithm choices can be used to represent documents based on their semantic relationships, such as Word2Vec, GloVe, or FastText [22]-[25]. In this research, the Word2Vec algorithm was chosen because of its ability to carry out analogous operations and understand contextual meaning better [26]-[28]. Apart from that, Word2Vec has advantages in its data processing capabilities where this algorithm is very appropriate for handling large data. The Word2Vec algorithm generally consists of 7 main stages i.e., choose Word2Vec model, prepare text corpus, tokenization and vocabulary formation, initialize word vectors, train the Word2Vec model, iterate through corpus and, update word vectors.

2.3. Search engine application development stage

The final stage of this research is the development of proposed search engine application for Indonesian translations of the Qur’an. Through this proposed application, it is hoped that users will be easier to find the information they need by simply entering the appropriate keywords. Furthermore, this proposed Qur’an search engine is a website-based application by employing the Django framework. Django is a Python-based web development framework designed to make it easy to create fast and efficient web applications [29]-[31]. Django is a popular choice for Python-based web development due to its ability to build small to large applications while maintaining good performance and high scalability [32]. The overall design for developing a search engine for the Al-Qur’an is shown in Figure 3.

Figure 2. Research framework
Based on the design of Figure 3, it can be seen that this research uses the cosine similarity algorithm to measure the level of similarity between the semantic keyword and the Al-Qur’an word vector. Semantic keywords are obtained from expanding queries typed by users based on their semantic relationship with the enriched Qur’an dataset cosine similarity is a metric used to measure the extent to which two direction vectors are close or the extent to which two documents or data sets are similar in a high-dimensional vector space. Cosine similarity is measured as the cosine of the angle between two vectors [33]-[35]. For example, there are two vectors, namely vector X and vector Y, which are two vectors in n-dimensional vector space, hence the cosine similarity value (cos θ) can be calculated using in (1):

\[
\text{Cosine Similarity} (X, Y) = \frac{X \cdot Y}{|X| \cdot |Y|}
\]

As for:
- \(X \cdot Y\) is the result of the dot product between vectors X and Y
- \(|X|\) and \(|Y|\) is the Euclidean norm of each vector

Once the results of the cosine similarity calculation have been successfully obtained, the search engine will rank it based on the level of similarity and then display it to the user through the search engine application interface.

3. RESULTS AND DISCUSSION

The performance of the proposed semantic Qur’an search engine application is determined based on precision and recall measurements. By using several keywords, experiments were carried out on the semantic Qur’an search engine application which implemented semantic word embedding and the search engine application without word embedding (ordinary search engine). After that, a performance comparison will be carried out to find out whether the proposed semantic Qur’an search engine application can exceed the performance of ordinary search engine applications.

### 3.1. Evaluation metrics

As previously mentioned, the performance of the Al-Qur’an verse search engine is calculated in terms of precision and recall. It is hoped that the proposed search engine can get high precision and recall values because this means that the search engine has good information search coverage as a sign that the search engine has succeeded in its expected performance. The measure of precision, also known as specificity, represents the correctness of information search results based on the ratio of relevant verses found (retrieved) to the total number of relevant and irrelevant verses found as shown in (2).
**Precision** = \( \frac{\text{relevant verses} \cap \text{retrieved verses}}{\text{retrieved verses}} \)  

The recall measure, also known as sensitivity, shows performance based on the ratio of the number of relevant verses found to the number of all possible related verses as shown in (3).

\[ \text{Relevant} = \frac{\text{relevant verses} \cap \text{retrieved verses}}{\text{relevant verses}} \]

### 3.2. Discussion

This research chose 7 topics i.e., prayer, zakat, fasting, pilgrimage, prophets, holy books, angels, apocalypse, trade, and clothing which are very important for Muslims as queries that will be tested. Augmented Al-Qur’an queries and datasets are converted into vectors based on semantic relationships using the word embedding Word2Vec algorithm. This semantic relationship-based vector conversion process will produce several keywords that will be used to search for information on Qur’anic verses according to relevant topics as described in Table 1 and Figure 4.

<table>
<thead>
<tr>
<th>Query</th>
<th>Top 7 generated semantic keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>prayer</td>
<td>shalat/ 0.915, salat/ 0.885, berjamaah/ 0.799, solat/ 0.77, subuh/ 0.769, ashar/ 0.763, tarawih/ 0.755</td>
</tr>
<tr>
<td>zakat</td>
<td>infaq/ 0.768, infal/ 0.702, wakaf/ 0.692, shadaqah/ 0.669, amil/ 0.628, fitrah/ 0.603, pembiayaan/ 0.585</td>
</tr>
<tr>
<td>fasting</td>
<td>berpuasa/ 0.794, ramadan/ 0.722, wakaf/ 0.692, sholat/ 0.669, shalat/ 0.663, adh/ 0.668, maulid/ 0.646</td>
</tr>
<tr>
<td>pilgrimage</td>
<td>umrah/ 0.678, khatib/ 0.594, syaikhon/ 0.59, manasik/ 0.563, syech/ 0.552, arsyad/ 0.551, husin/ 0.549</td>
</tr>
<tr>
<td>prophets</td>
<td>nabi/ 0.657, infil/ 0.631, paulus/ 0.63, timotius/ 0.627, matius/ 0.616, yohanes/ 0.615, bupa/ 0.592</td>
</tr>
<tr>
<td>angels</td>
<td>jibril/ 0.667, izrafil/ 0.598, iblis/ 0.59, tuhan/ 0.584, rasul/ 0.582, mikail/ 0.574, rob/ 0.572</td>
</tr>
<tr>
<td>apocalypse</td>
<td>berbangkit/ 0.555, penghakiman/ 0.543, qiymat/ 0.52, petang/ 0.517, ajal/ 0.51, dagal/ 0.508, harmagedon/ 0.505</td>
</tr>
</tbody>
</table>

Figure 4. Search engine interface for Zakat topics

*Table 1. Generated semantic keywords for topic*
As seen in Table 1, each topic has several keywords based on their semantic relationships. It can be concluded that broader information can be obtained due to the large number of keywords that are formed instead of just relying on limited keywords as is done by ordinary search engines. Once the vector conversion process and keyword evocation are complete, the similarity calculation between the topic and the verses of the Al-Qur'an is carried out using the cosine similarity algorithm. Furthermore, sorting is carried out based on the results of this calculation to obtain the most relevant verses (keywords) for each topic as shown in Table 2.

Table 2. Comparison results for proposed semantic and ordinary search engine

<table>
<thead>
<tr>
<th>Topic</th>
<th>Ordinary search engine</th>
<th>Proposed semantic search engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 15 Qur'an verses</td>
<td>Relevant</td>
<td>Top 15 Qur'an verses</td>
</tr>
</tbody>
</table>

Based on the results displayed in Table 2, it can be concluded that the proposed semantic search engine has many advantages over ordinary search engines in terms of its ability to search for broader information and the relevance of search results. It can be seen that the proposed search engine can search for more information due to the expansion of queries based on semantic keywords obtained from the word embedding representation. By utilizing generated semantic keywords, this proposed search engine can obtain additional information such as adding synonyms of words and other implied meanings. Furthermore, even though the proposed search engine retrieves more information, it also manages to maintain the relevance of search results in terms of precision and recall levels. The proposed search engine succeeded in getting very good performance with precision and recall values of 98.7% and 97.3% respectively. Therefore, this proposed search engine is very reliable in helping search for information from Indonesian translations of the Al-Qur’an.

Even though it has been proven to be successful in exceeding the performance of ordinary search engines, the proposed search engine still has several weaknesses in terms of the keywords it produces. Sometimes the resulting semantic keywords are not related to the topic the user wants to search for, such as the topic of prophets. Several keywords from this topic are more suitable for Christian religious documents, such as the keywords Paul, Timothy. This is due to the use of a wiki corpus which contains general words and is not specific to Islamic religious documents only. One solution that can be applied to overcome this problem is to add the concept of ontology to the semantic representation. Ontologies are expected to play a key role in searching for more specific information for certain domains. This is because ontology-based retrieval can enable users to search, organize, and present information in a more contextual and meaningful way.

4. CONCLUSION

The research proposes a semantic representation-based search tool for Indonesian translations of the Al-Qur’an. This Al-Qur’an verse is then augmented with the Wiki corpus to get a richer dataset to produce better semantic relationships. Then by using the word embedding concept based on the Word2Vec algorithm.
this augmented dataset will be converted into a vector based on semantic relationships. Finally using the cosine similarity algorithm, the cosine similarity is calculated between the augmented dataset and the selected topic. The performance of the proposed search engine is measured by comparing its ability to a regular search engine (without word embedding and augmented dataset) in finding relevant verses based on precision and recall metrics. In addition, the relevance of the search results for Al-Qur’an verses was evaluated by Al-Qur’an experts as additional validation. The performance obtained was 98.7% and 97.3% for precision and recall. The next research that is worth doing is the application of ontology so that users can search, organize, and present information in a more contextual and meaningful way. Apart from that, implementing other words embedding algorithms such as GloVe and FastText is also interesting to work on.

REFERENCES


Liza Trisnawati, S.T., M.Kom. is a lecturer at Abdurrab University, Indonesia. She obtained her Master in Information Technology from Putera Indonesia University (YPTK, Indonesia). He is currently pursuing his Ph.D. in Doctor of Philosophy of Information Technology at Tun Hussein University Onn Malaysia (UTHM). Her research interests include data mining and machine learning. She can be contacted at email: liza.trisnawati@univrab.ac.id.

Assoc. Prof. Dr. Noor Azah Binti Samsudin is Associate Professor at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). He holds a Ph.D. degree from the University of Queensland. Bachelor Of Computer Science, Universiti Kebangsaan Malaysia, Bachelor Of Computer Science, University Of Missouri, Columbia, US. She is Deputy Director of the Center for Academic Excellence and Development and Principal Investigator of the Faculty of Computer Science and Information Technology, Soft Computing, and Data Mining Center (SMC). Her areas of expertise are information, computer and communications technology (ICT), artificial intelligence, and machine learning. Her research area interests were then classification and feature selection. She can be contacted at email: azah@uthm.edu.my.

Dr. Shamsul Kamal Bin Ahmad Khalid is lecturer at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). The B.Sc. degree in Bachelor of Computer Science from New York University, US, the M.Sc. degree in Bachelor of Science (Computer), Universiti Kebangsaan Malaysia, and the Doctor of Philosophy in Information Technology from Tun Hussein University Onn Malaysia (UTHM), Malaysia. He is the Principal Investigator, Faculty of Computer Science and Information Technology, Center of Information Security Research (CISR), Universiti Tun Hussein Onn Malaysia (UTHM). His areas of expertise are information, computer and communications technology (ICT), security systems, and security services (including digital forensics, steganography, network security, public key infrastructure, and biometrics). His research area of interest was watermarking, and his innovation fields of interest are GPS-based applications and the internet of things (IoT). He can be contacted at email: shamsulk@uthm.edu.my.
Dr. Ezak Fadzrin Bin Ahmad Shaubari 他的更高研究在 MS 食品工程学由课程工作在在食品科学的学院，利兹大学从 1981-1982。他被附属于生物技术和生物分子科学学院。他的研究领域是喷雾干燥的食品。通过一个小研究资助由 UPM，他开发了生产喷雾干燥的椰奶过程这使全国的头条。他的丰富经验和专长在生物技术和生物分子科学已经使他成为该领域的一个全国参考点。他曾经于科学顾问办公室，总理的部门在国家生物质利用项目和是亚洲生物质协会的全国代表在东京，日本。他可以通过电邮: ezak@uthm.edu.my 联络。

Sukri, S.T., M.Kom. 是目前著于阿布杜拉巴大学，印尼。她获得信息技术的硕士学位在印度尼西亚伊斯兰大学，印度尼西亚。他正在 pursue 他的 Ph.D. 在信息技术的博士学位在 Tun Hussein Universiti Onn Malaysia (UTHM)。他的研究兴趣包括数据挖掘和机器学习。他可以通过电邮：sukri@univrab.ac.id 联络。

Zul Indra, M.Sc. 是一名讲师在里亚大学，印尼。他获得计算机和信息科学的硕士学位从 Universiti Teknologi PETRONAS。他正在 pursue 他的 Ph.D. 在计算机科学。他的研究兴趣包括软件发展，数据挖掘，和机器学习。他可以通过电邮：zulindra@lecturer.unri.ac.id 联络。