KARYA ILHAM PUBLISHING
CONNECTING MINDS, BRIDGING IDEAS, INSPIRING READERS

Journal of Advanced Research in Social and Behavioural Sciences

Journal homepage:
https://karyailham.com.my/index.php/jarsbs/index
ISSN: 2462-1951

Ontology Score of Malay Translated Hadith as BM25 Ranking Indicator for Malay Text Retrieval

Shaiful Bakhtiar Rodzman[1,*], Zulhilmi Mohamed Nor[2], Zamri Abu Bakar[3], Mohammad Nasrul Hakim Roslan[4], Mohd Azim Zainal[5]

[1] College of Computing and Mathematics, Universiti Teknologi MARA Pahang Branch Raub, Pahang, Malaysia
[2] Fakulti Pengajian Quran dan Sunnah, Universiti Sains Islam Malaysia, Bandar Baru Nilai, Malaysia
[3] Centre of Foundation Studies, Universiti Teknologi MARA Selangor Branch Dengkil, Malaysia
[4] Academy of Contemporary Islamic Studies, Universiti Teknologi MARA, Raub Pahang, Malaysia
[5] College of Computing and Mathematics, Universiti Teknologi MARA Pahang Branch Raub, Pahang, Malaysia

ARTICLE INFO

Article history:
Received 3 March 2026
Received in revised form 24 May 2026
Accepted 2 June 2026
Available online 2 July 2026

Keywords:
Ontology score; Information Retrieval Malay Malay translated hadith; positive ranking indicator

ABSTRACT

Information Retrieval (IR) especially search engine is widely used worldwide. Unfortunately, Information retrieval is lacking in the searching performance. With the usage of ontology, its searching performance can be increased. Ontology is the base of knowledge that keep all the knowledge about the domain. Malay ontology, especially in an Islamic base like Hadith Concept Ontologies on the other hand, are very limited. This research aims to develop a scoring technique using the Malay Translated Hadith Ontology Model as a positive BM25 ranking indicator for Malay Text Retrieval. The results were compared with the original scores of the BM25 Model and the Vector Space Model. The proposed scoring technique outperformed the original BM25 score in 15 queries, while the Vector Space model showed better results in six queries. The ontology score has been found to yield more accurate and superior documents compared to non-ontology search engines, particularly in specific domains or concept areas.

1. Introduction

Information retrieval (IR) is the process of retrieving materials from unstructured texts within large collections, typically stored on computers, to meet an information need *Manning et al.*, [1] also stated if in the past, people prefer to get information from other people rather than using IR System, nowadays as the IR system has been rapidly improved for better performance, more people prefer to search using the IR system. Many methods have been implemented in the IR such as using synonyms, thesaurus, indexing and now most developers are trying to research on ontology. IR has been extensively researched in the education field, primarily focusing on its use in providing relevant

* *Corresponding author.*
E-mail address: shaifulbakhtiarrodzman@uitm.edu.my

https://doi.org/10.37934/arsbs.44.1.213220

information to students and instructors [2]. One example suggests using information retrieval techniques to recommend patient education materials for diabetic questions [3].

The researchers define an ontology as a formal description of concepts, their properties, and restrictions, which together with individual instances of classes form a knowledge base, with a fine line between the two [4]. As there are rapid increase in the amount of Internet users, the demand for Islamic resource documents is also increasing. According to Drew and David [5], there are 1.6 billion Muslim around the world. It is the second biggest religion in the world. Hence, Islamic documents are heavily searched by the Muslims around the world as their source of religious knowledge and education, and there is a dire need of a good system to ensure efficiency. Two-third or over 62% out of 1.6 billion Muslims are living in Asia-Pacific region. Malay is one of the biggest races in Asia-Pacific. There are many documents that are in Malay language. However, it is sometimes difficult to search the Malay words. The words may have the same spelling as the English words though they carry different meaning. The need for language-aware information retrieval in education has been highlighted in previous studies [6].

Nowadays, most search engines have high percentage of retrieving irrelevant documents. Even if the document retrieved is relevant, sometimes it is placed lower than the irrelevant document. Some of the search engines even need additional tools and features to increase the search precision which may lead to more space usage of the computer. To prove this, the study had been made by using Google as the most used search engines, the word used for the search is 'sembahyang'(prayer). The result showed that the first ranked document that appears on the page is not really giving the knowledge about prayer.

In this research, the Ontology BM25 Malay Information Retrieval System (OBMIR) aims to improve search engine efficiency by incorporating more related terms to the keyword searched, enabling the retrieval of related documents that do not contain the keyword entered. Take the topic of Prayer as an instance, Prayer has many terms that are related to it. The terms can be listed with the help of ontology. The ontology is the way to make a certain concept or domain specific [7]. Which means, prayer as the domain will be broken into more specific terms related to it such as the definition, steps of prayers, type of prayers and many more to get the knowledge about this domain.

Furthermore, despite numerous research, including Ahmad *et al*.'s Semantic-based Ontology for Malay Qur'an Reader in Furthermore, despite numerous studies, including Ahmad et al.'s *Semantic-based Ontology for Malay Qur'an Reader* [8], no study has produced Hadith Concept Ontologies in Malay to improve the retrieval of Malay Translated hadith. The study by Yahya *et al*., [9], suggests that a Malay ontology is needed to improve English document retrieval accuracy, but it has shown a decline in accuracy for Malay documents, emphasizing the need for further research.

2. Literature Review

2.1 Related Work

2.1.1 Research on ontology of Islamic domain

Research was done by Saad *et al*., [10] that producing the ontology of domain Solat, which is one of the five pillars of Islam. The ontology is produced from the awareness that there is less attention on the aspect of practical results of the application development method. The research is done by referring mainly to the Quran which is the center of all knowledge about Islam and from the hadiths. The research had two stages to complete. In the first stage, the best topic is identified, and all documents of the best text are comprehended for further study in the field. The second stage is where the development of the ontology occurred which then is involved in the design and implementation. Other than that, Salim [11] proposes an ontology extraction based on fuzzy-swarm

algorithm for Islamic knowledge (IK) that will contribute lastly to the performance of Information Retrieval.

2.1.2 Research on Information Retrieval Using Ontology

Much research has been done so far regarding the application of Ontology in Information Retrieval such as Ayalew [12] that presents the results of experimental evaluation of an ontology-based frequently asked question retrieval system in the domain of HIV and AIDS with retrieval models BM25 and PL2. For the measurement of performance, the author used the measures of mean average precision, mean reciprocal rank, and precision at 5. The results suggested that frequently asked question retrieval with ontology are effective compared to the original scoring technique. NAda *et al*., [13] presents the adaptive IR system based on fuzzy sets and keyword extraction from the Ontology. The most important modules in the system are building fuzzy user models. The input of these modules was the keyword (term) weight in the initial user interest generated from the system in ontological form. The generated fuzzy user-model based on adaptive information retrieval IR system was developed. The experiment results of each experiment showed that the proposed approach can improve the performance of adaptive IR systems in commerce systems. Deshmukh *et al*., [14] proposes the IR System that can yield Improvement of retrieval process by using similar measures derived from knowledge about relations between concepts in ontologies.

The researchers were conducted on effective medical resource retrieval using an ontology-driven information system for the H1N1 case [15]. The aim of the study was to determine whether the ontology can enhance user performance in information searching via IR System, the difference in user performance and operationalized in terms of recall and search time that was using hyperlink technology and the system that was using ontology are taken. The result showed that the system with ontology driven has better performance recall and search time compared to the hyperlink-based information retrieval system. Moreover, Comfort *et al*., [16] presents a fuzzy-ontology based information retrieval system that determines the semantic equivalence between terms in a query and terms in a document by relating the synonyms of query terms with those of document terms. The results showed that relevant document to user's query was retrieved successfully. Green *et al*., [17] creates ontologies that are relevant to phenotype and maize such as Gene Ontology, Plant ontology, Train ontology and phenotype and trait ontology. These ontologies are created due to the need of flexible and accurate search mechanisms to retrieve documents related to their completed queries. Lastly, Lin [18] proposes the development of an intelligent system for book recommendation and retrieval by using personal profile created by fuzzy ontology map that helps reduce the search time. The fuzzy decision-making is based on the interest level generated from user preferences, product feature selection and possible system recommendation. The simulation testing showed promising results.

2.2 Improvement of BM25 Model

BM25 model is an extension of the Probabilistic model. Probabilistic model is one of the Classic IR Model. Contrary to the original conceptions of the classic probabilistic model, the BM25 model can be computed without any relevant information provided by the user [19]. Nowadays, plenty of research has been done regarding the improvement of BM25 model. Khatoon *et al*., [20] proposes an enhanced BM25 mathematical approach to retrieve the most query relevant literature from the clustered Biomedical literature bank with query expansion, by analyzing the clustered biomedical topics with different pre-processing methods and term-weighting functions and found the best

values for the tuning parameters K1, b, K3 of BM25. The results showed the improvement of Mean Average Precision and R-precision. Other than that, Runjie [21] proposes the relative term frequency to be integrated into traditional probabilistic models. In other words, a set of three influence functions with the application of relative term frequency to model are introduced. This enhances the performance of the fundamental probabilistic weighting function, BM25.

3. Research Method

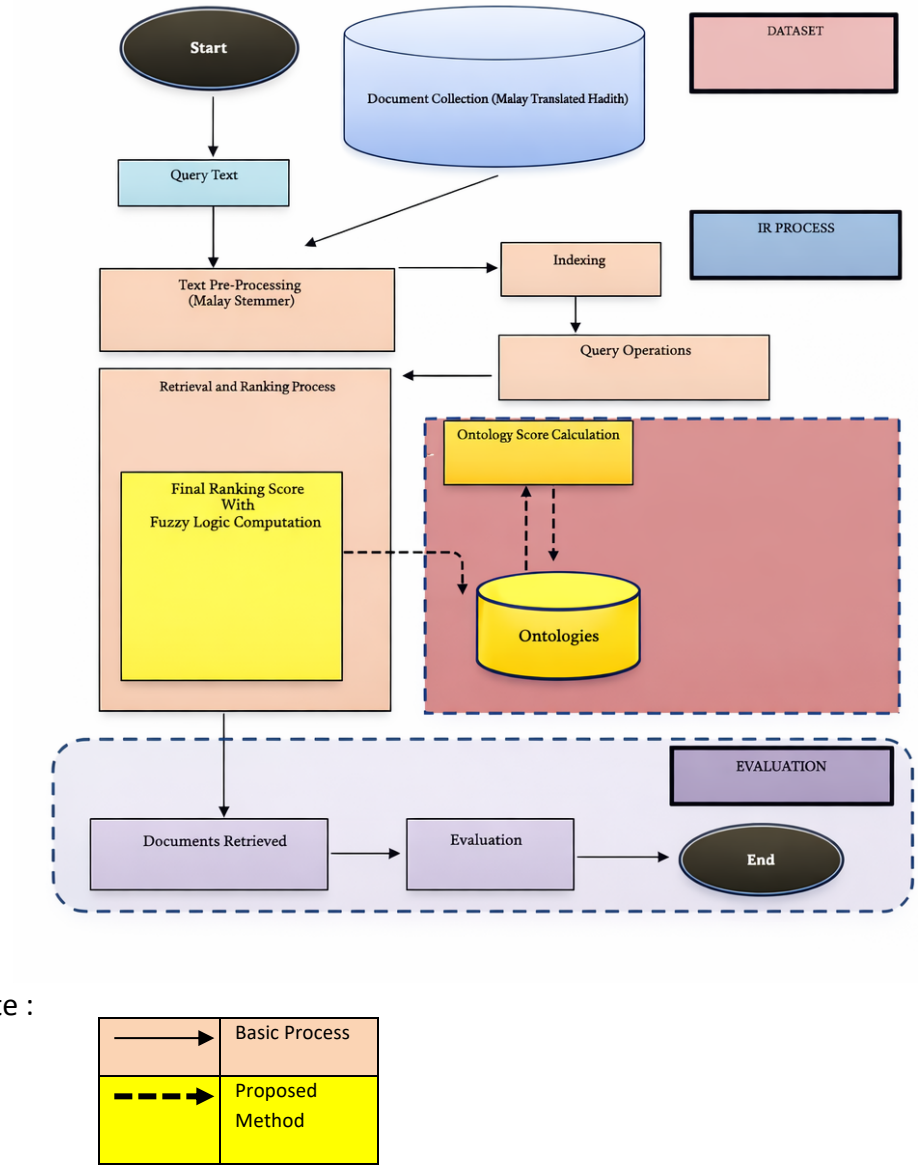

Fig. 1. Conceptual framework of OBMIR

Figure 1 illustrates the framework of The Ontology BM25 Malay Information Retrieval System (OBMIR) which is a search engine prototype designed to enhance the ranking function of the BM25 by using the Ontology Score of Document Query as a positive indicator. The framework includes the basic processes of Information Retrieval, including Pre-Processing, Query Operation, Retrieval and Ranking Process, which are new contributions to this research. The overall processes will be evaluated by comparing them with the original score and Vector Model Malay IR System (VM).

3.1 Information Retrieval System

Algorithms

1) Modification of Model: The BM25 model is used as the original score in this system due to its superior results compared to the Vector Space Model in general collections in Ricardo and Berthier [19]. It is widely used as a baseline for evaluating new ranking methods and is currently used alongside ontology to improve English Information Retrieval, as demonstrated by Ayalew *et al*., [12].

2) Ranking Function: The system will enhance its ranking function by incorporating an additional Ontology scoring using the BM25 model, with the BM25 score, D, based on the query containing keywords from q1 to qn :

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i).\left(\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1.\left(1-b=b.\frac{|D|}{avgdl}\right)}\right)$$

3) Algorithm: The search engine uses a system to clean queries from special characters or symbols using Malay's Fatimah Stemmer. After the stemming process, the ontology indexes are pre-processed and assigned to an array. The system intelligently selects one topic to represent the entire query. For each query term, it searches for the term's children, siblings, and parents if none exist. If the parent does not exist, the term is not in the ontology indexes. If the algorithm finds a specific term in the ontology, related terms are involved in Ontology scoring using the standard BM25 ranking function formula. The system continues to search for related terms if the term is not in the ontology indexes. The formula of Ontology Score as below:

$$\text{Ontology Score} = \sum_{i=1}^{n} IDF(q_i).\left(\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1.\left(1-b=b.\frac{|D|}{avgdl}\right)}\right)$$

For example, The Ontology scoring of BM25 will be calculated for terms like 'rukun' and'sujud', based on their related terms. For terms without ontological relationships, the standard BM25 ranking function will be used. Documents retrieved will be ranked based on their priority using the Final BM25 Ranking Score, accumulating the original score and Ontology Score of Documents. (see Eq. 1.1).

$$\text{Final BM25 Ranking Score} = score(D,Q) + \sum ONT(score(D,Q)) \quad (1.1)$$

Where, ONT (score (D, Q)) is a Accumulated of Ontology Score of Documents for given query.

3.2 Pre-processing of Ontology Indexes

The pre-processing of an ontology was developed for the domain of Malay Translated Hadith of Sahih Bukhari, consisting of 32 Hadith Concepts. The ontology was classified under language expressivity and formality, with a focus on software ontology. The ontology was designed using a top-down approach, starting with the domain and expanding with related terms. The ontology serves as the knowledge base of the domain, and a.txt file was created to store all ontology terms. The file lists children on the left column and parents on the right, mapping each other. For example, the concept of 'Solat' can be mapped to 'rukun' if the child is 'niat', and'sunat' if the child is 'hukum'.

4. Result and Discussion

4.1 Performance Evaluation

The experiment uses the Malay corpus, consisting of 2026 Malay Translated of Sahih Bukhari Hadith text document and Hadith Concept Ontologies of 32 topics from the text. Experts Dr. Zulhilmi bin Mohamed Nor from Universiti Sains Islam Malaysia and Dr. Ahmad Yunus Mohd Noor from Universiti Kebangsaan Malaysia help developed the ontology and tohey also referenced to the book Al-Fiqh Al-Manhaji Mazhab Al-Syafie, translated to Malay by Zulkifli *et al*., [22]. The researchers also used 30 queries from eight topics to observe the results of the proposed system. Topic 'Umum' had 11 queries, while other topics had two to four queries with an average of three queries for each topic. The relevance of the document was determined by Hadith experts such as, Dr Zulhilmi bin Mohamed Nor from Universiti Sains Islam Malaysia and Dr Ahmad Yunus Mohd Noor from Universiti Kebangsaan Malaysia. (P@10), (%no) and *AP* have been using as the evaluation metric for this experiment.

No.	TOPIC	*FBMIR* P@10	*BM25* P@10	*Vector* P@10	*FBMIR* %no	*BM25* %no	*Vector* %no	*FBMIR* MAP	*BM25 MAP*	*Vector* MAP
1	*Makanan*	0.377	0.8	0.466	65	62.3	20	53.4	0.824	0.502
2	*Adab*	0.444	1	0.75	55.6	0	25	0.421	1	0.715
3	*Solat*	0.549	0.476	0.444	45.1	52.4	55.6	0.622	0.486	0.565
4	*Iman*	0.404	0.466	0.3	59.6	53.4	70	0.361	0.463	0.265
5	*Bersuci*	0.522	0.937	0.513	47.8	6.3	48.7	0.576	0.968	0.534
6	*Ibadah*	0.693	0.35	0.541	30.7	75	45.9	0.727	0.328	0.577
7	*Sirah*	0.677	0.733	0733	32.3	26.7	26.7	0.776	0.830	0.905
8	*Umum*	0.534	0.333	0.353	46.6	66.7	74.7	0.582	0.338	0.392

Fig. 2. Evaluation results of information retrieval models on Malay Hadith Corpus

The experiment showed that OBMIR slightly outperformed BM25 original score and Vector Space Model (VM) on three topic sets of query results, such as "Solat", "Ibadah", and "Umum", on the P@10, %no measures, and MAP. BM25 original score outperformed OBMIR for all results on the P@10, %no measures, and MAP in four topic sets, while the Vector model slightly outperformed OBMIR for all results on the P@10, %no measures, and MAP for the topic of "Sirah". Overall, OBMIR outperformed the other two models in terms of individual query results, yielding better results in 15 queries out of 30 queries.

4.2 Significance, Limitation and Future Works

From the results, research indicates ontology application as a positive ranking indicator for BM25 improves document retrieval accuracy, benefiting Islamic Knowledge students and the public, enhancing search query results. Hadith Concept Ontologies also can aid Muslims and Islamic Knowledge students in finding related knowledge on specific topics for their study purposes. However, Ontology application may decrease performance for certain queries, particularly those retrieving good results with BM25 original score, due to system recall and potential impact on previous relevant documents. For further work, the system's performance could be enhanced by incorporating fusion selection, filtering, or term selection into the OBMIR algorithm for future work. Moreover, the ontology can be enhanced by adding relevant terms for better document retrieval and strengthening Hadith Concept Ontologies, benefiting all Muslims, especially Islamic Knowledge students. Other than that, researchers aim to enhance the results using Fuzzy logic technique by incorporating both positive and negative ranking indicators.

5. Conclusion

In overall, researchers developed a ranking function using Hadith Concept Ontologies in Malay language and the BM25 model of IR system. The OBMIR system outperformed the original BM25 model in 15 queries, with better results in P@10, % no measures, and MAP. Future work may include improving Hadith Concept Ontologies, using another corpus for positive and negative ranking indicators, and using Fuzzy Logic techniques.

Acknowledgment

All authors are grateful for College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Pahang Branch for financial support.

References

[1] Manning, D, Raghavan & Schutze. *An Introduction to Information Retrieval,* Online edn, Cambridge University Press, Cambridge, England, 2009. https://dl.acm.org/doi/abs/10.5555/1394399

[2] Meisalo, Veijo, Erkki Sutinen, and Jorma Tarhio. "The possibilities of information retrieval in education." In *Current research on mathematics and science education, Proc. The XXI annual symposium of The Finnish Association of Mathematics and Science Education Research (eds. A. Laine et al.)*, pp. 528-549. 2004.

[3] Zeng, Yuqun, Xusheng Liu, Yanshan Wang, Feichen Shen, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, and Hongfang Liu. "Recommending education materials for diabetic questions using information retrieval approaches." *Journal of medical Internet research* 19, no. 10 (2017): e342.https://doi.org/10.2196/jmir.7754

[4] Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: a guide to create your first ontology." *Knowledge Systems Laboratory Technical Report KSL-01-05, Stanford University. 25p* (2001).

[5] Drew Desilver, David Masci. World's Muslim population more widespread than you might think, viewed 20 November 2018, <http://www.pewresearch.org/fact-tank/2017/01/31/worlds-muslim-population-more-widespread-than-you-might-think/>.

[6] Ott, Niels, and Detmar Meurers. "Information Retrieval for Education: Making Search Engines Language Aware." *Themes in Science and Technology Education* 3 (2010): 9-30. https://www.learntechlib.org/p/148641/

[7] Gruber, Thomas R. "A translation approach to portable ontology specifications." *Knowledge acquisition* 5, no. 2 (1993): 199-220. https://doi.org/10.1006/knac.1993.1008

[8] Ahmad, Nor Diana, Brandon Bennett, and E. S. Atwell. "Semantic-based ontology for Malay Qur'an reader." In *IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies*. Leeds, 2016.http://doi.org/10.7763/IJFCC.2013.V2.229.

[9] Yahya, Zulaini, Muhamad Taufik Abdullah, Azreen Azman, and Rabiah Abdul Kadir. "Query translation using concepts similarity based on Quran ontology for cross-language information retrieval." *Journal of Computer Science* 9, no. 7 (2013): 889. https://doi.org/10.3844/jcssp.2013.889.897

[10] Saad, Saidah, Naomie Salim, Hakim Zainal, and Zurina Muda. "A process for building domain ontology: An experience in developing Solat ontology." In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pp. 1-5. IEEE, 2011. https://api.semanticscholar.org/CorpusID:14155037

[11] Salim, Naomie. Methodology of ontology extraction for islamic knowledge text, in *Seminar Kebangsaan Pendidikan Sains dan Matematik, UTM, 2008*. https://api.semanticscholar.org/CorpusID:17562730

[12] Ayalew, Yirsaw, Barbara Moeng, and Gontlafetse Mosweunyane. "Experimental evaluation of ontology-based HIV/AIDS frequently asked question retrieval system." *Health Informatics Journal* 25, no. 4 (2019): 1434-1450. https://doi.org/10.1177/1460458218775147

[13] Nada, Yasser A., Khaled M. Fouad, Hossam Meshref, and Y. M. Assem. "New System for Adaptive Information Retrieval Based on Fuzzy Sets." *International Journal of Applied Engineering Research* 12, no. 19 (2017): 8160-8173.https://api.semanticscholar.org/CorpusID:39114306

[14] Deshmukh, Deepenti H., Poonam B. Lohiya, and Shwetambari G. Pundkar. "Information retrieval system based on ontology." In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 1061-1065. IEEE, 2016. https://doi.org/10.1109/SCOPES.2016.7955603

[15] Yi, Myongho. "Effective medical resources searching using an ontology-driven medical information retrieval system: H1N1 case study." *The Electronic Library* 30, no. 2 (2012): 248-263.https://doi.org/10.1108/02640471211221368

[16] Comfort T Akinribido, Babajide S Afolabi, Bernard I Akhigbe, Ifiok J Udo. A Fuzzy Ontology Based Information Retrieval System for Relevant Feedback, *International Journal of Computer Science, World Academy of Science, Engineering and Technology,* Vol.8, No.1, pp. 382-389, 2011.

[17] Green, Jason M., Jaturon Harnsomburana, Mary L. Schaeffer, Carolyn J. Lawrence, and Chi-Ren Shyu. "Multi-source and ontology-based retrieval engine for maize mutant phenotypes." *Database* 2011 (2011): bar012.https://doi.org/10.1093/database/bar012

[18] Lin, J. N. K. "Fuzzy Ontology-Based System for Product Management and Recommendation." *International Journal of Computers* 1, no. 3 (2007). https://api.semanticscholar.org/CorpusID:18564424

[19] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. *Modern Information Retrieval the concepts and technology behind search,* 2 edn, Pearson Education Limited, England, pp. 57-128, 2011.

[20] Khatoon, Thayyaba, and A. Govardhan. "Query Expansion with Enhanced-BM25 Approach for Improving the Search Query Performance on Clustered Biomedical Literature Retrieval." *Journal of Digital Information Management* 16, no. 2 (2018). https://doi.org/10.6025/jdim/2018/16/2/85-98

[21] Runjie Zhu. Improvement In Probabilistic Information Retrieval Model -Rewarding Terms With High Relative Term Frequency, M.S thesis, York University, Toronto, Canada, 2018.

[22] Zulkifli bin Mohamad al-Bakri. Al-Fiqh al-Manhaji Mazhab al-Syafie, Jabatan Kemajuan Islam Malaysia (JAKIM), 2016.