JSCDM

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/jscdm
e-ISSN : 2716-621X

Journal of Soft Computing and Data Mining

Query Expansion for Quran French Text Retrieval Using Semantic Search

Nuhu Yusuf[1,2], Mohd Amin Mohd Yunus[1*], Norfaradilla Wahid[1]

[1]Faculty of Computer Science & Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, 86400, MALAYSIA

[2]Mgt. & Information Technology Department,
Abubakar Tafawa Balewa University Bauchi, Gubi, Bauchi, 234, NIGERIA

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2020.01.02.003
Received 20 October 2020; Accepted 30 November 2020; Available online 15 December 2020

Abstract: Quran translation search are now gaining interest of many researchers in the field of Quran informatics. Current Quran translation search performance still needs to be improving. Despite the fact that many researchers contribute on its performance, few considered translation performance on semantic search for query expansion in French text retrieval. In this study, the performance of Quran translation search is proposed using semantic search. The experiment was carried out using 6236 verses of the holy Quran translated by Hamidullah. The results show that the proposed query expansion based on semantic search perform best and achieved almost 36% improvement. Experiment on the Quran ontology document should serve as the further research direction.

Keywords: Information retrieval, text retrieval, Quran search, semantic search, query expansion

1. Introduction

Information retrieval concerns with the finding information need from a document collection [1], [2]. Text information retrieval is one of the branches of information retrieval which deals with the text information need that users retrieved. It is widely used in various domains to retrieve information using different languages [3], [4]. Such languages among others are English, Arabic, French, Malay and German. Text information retrieval plays significant roles in getting users information needs based on their queries. In particular, it becomes difficult to retrieve user information needs with higher precision percentage [5], [6], [7]. Therefore, a specific language vocabularies usage from text retrieval ensures users relevant information need [8], [9].

French text retrieval uses French language vocabularies in queries to obtain users information need from the available collections [10]. To improve the quality of French text retrieval, the query must be expanded with additional information for better search results [11], [12]. This is because; additional terms can provide more information to query without changing its meaning [13].

Recently, semantic search techniques have been presented in many papers to improve text retrieval and expand the query [14], [15]. Consequently, the current techniques have covered lexical [16], ontology [17] and machine learning [18]. French text retrieval frequently assumed terms within sentence query distribution are important. For better search results, query expansion must be considered the context of terms as appeared in the sentence. However, there is still need to revisit how the context of words helps to improve French text retrieval as only a few contributions were made on the field and these contributions don't consider improvement in terms of query expansion with explicit relevant judgment from experts.

*Corresponding author: gi170012@siswa.uthm.edu.my

This paper improves Quran French text retrieval based on query expansion with latent semantic analysis. The contributions of this paper are as follows:

1. Propose semantic search query expansion which considered the context of terms appeared in a sentence
2. Compare the results with other similar semantic search approaches.

2. Related Works

Text information retrieval use to retrieved users' information needs. However, the quality of the retrieved results still not up to satisfaction, as irrelevant results also retrieved. Many research efforts have been considered to improve the quality of search results. Sharma et al. [19] present a new framework for search improvement using ontology. However, the paper doesn't thoroughly discuss how semantic were computed. In addition to that, Herrera et al. [20] utilize semantic annotation based on domain ontology. Hidden terms relationship between two documents can provide better improvement. Jiang et al. [17] present how to develop terms relationship using ontology. The ontology contained both terms synonyms and antonyms [21]. Moreover, Zhu et al. [22] suggest that semantic similarity improvement can effectively produce high-quality matching results. Yusuf et al [23] expand the query using English word synonyms while Soudani et al [8] believed Arabic word synonyms can provide better semantic information retrieval. This paper will also utilize the semantic approach within words context in a sentence.

Pseudo-relevant feedback is a query expansion technique that most of the recent query expansion methods used. However, it is only assumed top documents are important to users need [5], [9]. Specifically, the top documents may contain irrelevant documents. Explicit relevant feedback requires expert evaluation on the quality of retrieved results. Although explicit feedback has also based on the top document, it is good in expanding the query. Implicit relevant feedback is the technique that captures user feedback automatically [4], [13]. Other techniques such as Wikipedia [24]and thesaurus [25] still play vital roles in query expansion. This paper considered explicit feedback to expand the query. It considered the experts from Quran translations domain to provide their input on search results.

3. Methods

The proposed semantic search method is a distributional model which changes Quran French text retrieval into the matrix. Given a Quran French text query $Y = \{y_1, y_2, y_3 \ldots . . y_n\}$ where n is n x n matrix. The row-column represents the target term distribution [26]. To compute the co-occurrence of words, we considered the distance of a term within a particular angle. Therefore, the cosine similarity has been used based on equation (1):

$$Similarity = \frac{\sum_{i=1}^{n} C_i D_i}{\sqrt{\sum_{i=1}^{n} C_i^2}\sqrt{\sum_{i=1}^{n} D_i^2}} \quad (1)$$

Where $C_{i\ and}\ D_i$ are terms vectors of C and D respectively

Furthermore, to rank our documents, we utilize that similarity obtained in equation (1) which is based on Quran French text retrieval. We used the Okapi BM25 ranking document which comprises term frequency and inverse document frequency. The BM25 ranking is given in equation (2):

$$Score(D,Q) = \log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \cdot \frac{f(q_i,D) \cdot (K_i + 1)}{f(q_i,D) + K_i \cdot \left(1 - b + b\frac{|D|}{avgdl}\right)} \quad (2)$$

Where D represent the document and Q represent the query term. Also, k and b represent free parameters while N represents the total number of documents within the corpus.

To evaluate the ranking results, we used Mean Rank Reciprocal (MRR) in addition to mean average precision (MAP) which is the most widely used. The equations (3) and (4) represent MRR and MAP respectively:

$$MRR = \frac{1}{2}\sum_{q=1}^{N}\frac{1}{rank_q} \tag{3}$$

$$MAP = \frac{1}{Q}\sum_{Z=1}^{Q}\frac{1}{D_z}\sum_{i=1}^{DZ}P(d_i) \tag{4}$$

Our proposed approach was tested using Hamidullah Quran dataset collected from Hamid [27]. We also used thirty-six queries for the experiment. Yusuf et al. [28] explained different types of queries that can be used in retrieving information [29], [30].

4. Results and Discussion

This section presents the results of the proposed query expansion method based on semantic search and compared with similar models as shown in Table 1. In term of MAP result, Table 1 shows the MAP using BM25, the proposed method and the Yusuf et al. The proposed method outperforms other methods in retrieving relevance results. It achieves **35.92%** improvements as compared to BM25 with 18.61% and Yusuf et al (citation) with17.75%. Also, for MRR, Table 1 shows that our proposed method obtained higher MRR results with 22.05% as against BM25 with 10.15 and Yusuf et al with 10.58.

Table 1 - Results comparisons

Methods	MAP (%)	MRR (%)
BM25	18.61	10.15
Proposed Method	35.92	22.05
Yusuf et al	17.75	10.58

5. Conclusion and Future Work

This paper used improved the query expansion method for Quran French text information retrieval using semantic search. Six thousands two hundred and thirty- six (6236) verses have been considered for our investigation. The results show that our proposed query expansion method performed better compared to other models. Further study should consider developing Quran ontology and then use to expand the query of the holy Quran

Acknowledgement

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

The authors would like to thank the Center for Graduate Studies Universiti Tun Hussein Onn Malaysia (UTHM), the Faculty of Computer Science & Information Technology UTHM and indeed the Abubakar Tafawa Balewa University (ATBU) Bauchi for their support during this research paper.

References

[1] Yusuf, N., Yunus, M. A. M., Wahid, N., Nawi, N. M., Samsudin, N. A., & Arbaiy, N. (2020). Query Expansion Method for Quran Search Using Semantic Search and Lucene Ranking. *Journal of Engineering Science and Technology*, *15*(1), 675-692.

[2] Bounhas, I., Soudani, N., & Slimani, Y. (2020). Building a morpho-semantic knowledge graph for Arabic information retrieval. *Information Processing & Management*, *57*(6), 102124.

[3] Ghembaza, M. I. E. K. (2019). Specialized Quranic Semantic Search Engine. *International Journal of Computer Science and Information Security (IJCSIS)*, *17*(2).
[4] Atwan, J., & Mohd, M. (2017). Arabic Query Expansion: A Review. *Asian Journal of Information Technology*, *16*(10), 754-770.
[5] Soudani, N., Bounhas, I., & Slimani, Y. (2019). MOSSA: a morpho-semantic knowledge extraction system for Arabic information retrieval. *International Journal of Knowledge and Web Intelligence*, *6*(2), 106-141.
[6] Khan, S. Z., Rahman, M. M., Sadi, A. S., Anwar, T., Mohammed, S., & Chowdhury, S. (2017). The Quranic Nature Ontology: From Sparql Endpoint to Java Application and Reasoning. *International Journal of Innovative Computing*, *7*(2).
[7] Ababneh, A., Lu, J., & Xu, Q. (2016, August). Arabic Information Retrieval: A Relevancy Assessment Survey. ISD.
[8] Safi, H., Jaoua, M., & Belguith, L. H. (2016, June). PIRAT: a Personalized Information Retrieval system in Arabic Texts based on a hybrid representation of a user profile. In *International Conference on Applications of Natural Language to Information Systems* (pp. 326-334). Springer, Cham.
[9] Lahbari, I., El Alaoui, S. O., & Zidani, K. A. (2018). Toward a new arabic question answering system. *Int. Arab J. Inf. Technol.*, *15*(3A), 610-619.
[10] Soudani, N., Bounhas, I., & Slimani, Y. (2016, November). Semantic information retrieval: A comparative experimental study of NLP tools and language resources for arabic. In *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)* (pp. 879-887). IEEE.
[11] Samy, H., Hassanein, E. E., & Shaalan, K. Arabic Question Answering: A Study on Challenges, Systems, and Techniques. *International Journal of Computer Applications*, *975*, 8887.
[12] Moussallem, D., Wauer, M., & Ngomo, A. C. N. (2018). Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, *51*, 1-19.
[13] Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. *Database*, *2018*.
[14] Bakari, W., Bellot, P., & Neji, M. (2016). Researches and Reviews in Arabic Question Answering: principal approaches and systems with classification.
[15] Ray, S. K., & Shaalan, K. (2016). A review and future perspectives of arabic question answering systems. *IEEE Transactions on Knowledge and Data Engineering*, *28*(12), 3169-3190.
[16] Petrasova, S., Khairova, N., & Lewoniewski, W. (2018, August). Building the semantic similarity model for social network data streams. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)* (pp. 21-24). IEEE.
[17] Jiang, S., Hagelien, T. F., Natvig, M., & Li, J. (2019, January). Ontology-based semantic search for open government data. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)* (pp. 7-15). IEEE.
[18] Banisakher, D., Reyes, M. E. P., Eisengberg, J. D., Allen, J., Finlayson, M. A., Price, R., & Chen, S. C. (2018, July). Ontology-Based Supervised Concept Learning for the Biogeochemical Literature. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 402-410). IEEE.
[19] Sharma, S., Kumar, A., & Rana, V. (2017, December). Ontology based informational retrieval system on the semantic web: Semantic Web Mining. In *2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)* (pp. 35-37). IEEE.
[20] Herrera, N. P., Gomez, F. L., Bucheli, V. A., & Pabón, O. S. (2018). Semantic annotation and retrieval of scientific documents in a big data environment, 7 (6).
[21] Kherwa, P., & Bansal, P. (2017, September). Latent Semantic Analysis: An Approach to Understand Semantic of Text. In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)* (pp. 870-874). IEEE.
[22] Zhu, B., Li, X., & Sancho, J. B. (2017, December). A novel asymmetric semantic similarity measurement for semantic job matching. In *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)* (pp. 152-157). IEEE.
[23] Yusuf, N., Amin, M., Yunus, M., & Wahid, N. (2019). Query expansion based on explicit-relevant feedback and synonyms for English Quran translation information retrieval. *Int. J. Adv. Comput. Sci. Appl.*, *10*(5), 227-234.
[24] Jiang, Y., Bai, W., Zhang, X., & Hu, J. (2017). Wikipedia-based information content and semantic similarity computation. *Information Processing & Management*, *53*(1), 248-265.
[25] Jana, A., & Goyal, P. (2018). Can network embedding of distributional thesaurus be combined with word vectors for better representation?. *arXiv preprint arXiv:1802.06196*.
[26] Algburi, M. A., Mustapha, A., Mostafa, S. A., & Saringatb, M. Z. (2019, September). Comparative Analysis for Arabic Sentiment Classification. In *International Conference on Applied Computing to Support Industry: Innovation and Technology* (pp. 271-285). Springer, Cham.
[27] Hamid, Z. Z. (2007). Quran translations. *Tanzil Documents*.
[28] Yusuf, N., Yunus, M. A. M., & Wahid, N. (2019). A comparative analysis of web search query: informational vs navigational queries. *Int. J. Adv. Sci. Eng. Inf. Technol*, *9*(1), 136-141.

[29] Mohammed, M. A., Gunasekaran, S. S., Mostafa, S. A., Mustafa, A., & Abd Ghani, M. K. (2018, August). Implementing an agent-based multi-natural language anti-spam model. In *2018 International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR)* (pp. 1-5). IEEE.

[30] Basir, N., Nabila, N. F., Zaizi, N. J. M., Saudi, M. M., Ridzuan, F. H. M., & Pitchay, S. A. (2017). Intelligent Quranic ontology retrieval. *Advanced Science Letters, 23*(5), 4449-4453.