A Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus

Mohammad Alqahtani
University of Leeds
scmmal@leeds.ac.uk

Eric Atwell
University of Leeds
E.S.Atwell@leeds.ac.uk

The Holy Qur'an is the most important resource for the Islamic sciences and the Arabic language (Iqbal et al., 2013). Muslims believe that the Qur'an is a revelation from Allah that was given 1,356 years ago. The Qur'an contains about 80,000 words divided into 114 chapters (Atwell et al., 2011). A chapter consists of a varying number of verses. This holy book contains information on diverse topics, such as life and the history of humanity and scientific knowledge (Alrehaili and Atwell, 2014). Corpus linguistics methods can be applied to study the lexical patterns in the Qur'an; for example, the Qur'an is one of the corpora available on the SketchEngine website. Qur'an researchers may want to go beyond word patterns to search for specific concepts and information. As a result, many Qur'anic search applications have been built to facilitate the retrieval of information from the Qur'an. Examples of these web applications are Qurany (Abbas, 2009), Qur'an Explorer (Explorer, 2005), Tanzil (Zarrabi-Zadeh, 2007), Qur'anic Arabic corpus (Dukes, 2013), and Quran.com.

The techniques used to retrieve information from the Qur'an can be classified into two types: semantic-based and keyword-based. Semantic-based search techniques are concept-based which retrieves results by matching the contextual meaning of terms as they appear in a user's query, whereas the keyword-based search technique returns results according to the letters in the word(s) of a query (Sudeepthi et al., 2012). The majority of Qur'anic search tools employ the keyword search technique.

The existing Qur'anic semantic search techniques include the ontology-based technique (concepts) (Yauri et al., 2013), the synonyms-set technique (Shoaib et al., 2009), and the cross language information retrieval (CLIR) technique (Yunus et al., 2010). The ontology-based technique searches for the concept(s) matching a user's query and then returns the verses related to these concept(s). The synonyms-set method produces all synonyms of the query word using WordNet and then returns all Qur'anic verses that contain words matching any synonyms of the query word. Cross language information retrieval (CLIR) translates the words of an input query into another language and then retrieves verses that contain words matching the translated words.

On the other hand, keyword-based techniques include keyword matching, the morphologically-based technique (Al Gharaibeh et al., 2011), and use of a Chabot (Abu Shawar and Atwell, 2004). The keyword matching method returns verses that contain any of the query words. The morphologically-based technique uses stems of query words to search in the Qur'an corpus. In other words, this technique generates all other forms of the query words and then finds all Qur'anic verses matching those word forms. The Chabot selects the most important words such as nouns or verbs from a user query and then returns the Qur'anic verses that contain any words matching the selected words.

There are several deficiencies with the Qur'anic verses (Aya'at) retrieved for a query using the existing keyword search technique. These problems include the following: some irrelevant verses are retrieved, some relevant verses are not retrieved, or the sequence of retrieved verses is not in the right order (Shoaib et al., 2009). Misunderstanding the exact meaning of input words forming a query and neglecting some theories of information retrieval contribute significantly to limitations in the keyword-based technique (Raza et al.). Additionally, Qur'anic keyword search tools use limited Islamic resources related to the Qur'an. This affects the accuracy of the retrieved results.

Moreover, current Qur'anic semantic search techniques have limitations in retrieved results. The main causes of these limitations include the following: semantic search tools use one source of Qur'anic ontology that does not cover all concepts in the Holy Qur'an, and Qur'anic ontologies are not aligned to each other, leading to inaccurate and uncomprehensive resources for Qur'anic ontology.

To overcome the limitations in both semantic and keyword search techniques, we designed a framework for a new semantic search tool called the Qur'anic Semantic Search Tool (QSST). This search tool aims to employ both text-based and semantic search techniques. QSST aligns the existing Quranic ontologies to reduce the ambiguity in the search results.

QSST can be divided into four components: a natural language analyser (NLA), a semantic search model (SSM), a keywords search model (KSM), and a scoring and ranking model (SRM). NLA tokenizes a user's query and then applies different natural language processing techniques to the tokenized query. These techniques are the following: spelling correction, stop word removal, stemming, and part of speech tagging (POS). After that, the NLA uses WordNet to generate synonyms for the reformatted query words and sends these synonyms to the SSM and the KSM. The SSM searches in the Qur'anic Ontology database to find the related concepts of the normalised query and then returns results. At the same time, KSM retrieves results based on words matching the input words. SRM refines the results retrieved from both KSM and SSM by eliminating the redundant verses. Next, SRM ranks and scores the refined results. Finally, SRM presents the results to the user.

References

Abbas, N. H. 2009. *Quran 'search for a concept' tool and website*. MRes thesis, University of Leeds.

Abu Shawar, B. and Atwell, E. 2004. An Arabic chatbot giving answers from the Qur'an. Proceedings of TALN. **4**(2), pp.197-202.

Al Gharaibeh, A. et al. 2011. The usage of formal methods in Quran search system. In: Proceedings of international conference on information and communication systems, Ibrid, Jordan. pp.22-24.

Alrehaili, S. M. and Atwell, E. 2014. Computational ontologies for semantic tagging of the Quran: A survey of past approaches. In: LREC 2014 Proceedings.

Atwell, E. et al. 2011. An artificial intelligence approach to Arabic and Islamic content on the internet. In: Proceedings of NITS 3rd National Information Technology Symposium.

Dukes, K. 2013. *Statistical parsing by machine learning from a classical Arabic treebank*. PhD thesis.

Explorer, Q. 2005. *Quran Explorer* [Online]. [Accessed 26 October 2014]. Available from: http://www.quranexplorer.com/Search/Default.aspx

Iqbal, R. et al. 2013. An experience of developing Quran ontology with contextual information support. *Multicultural Education & Technology Journal*. **7**, pp.333-343.

Raza, S.A. et al. An essential framework for concept based evolutionary Quranic search engine (CEQSE).

Shoaib, M. et al. 2009. Relational WordNet model for semantic search in Holy Quran. Emerging Technologies, 2009. ICET 2009. International Conference on, 2009. IEEE, 29-34.

Sudeepthi, G. et al. 2012. A survey on semantic web search engine. *International Journal of Computer Science*, **9**.

Yauri, A. R. et al. 2013. Quranic verse extraction based on concepts using OWL-DL ontology. *Research Journal of Applied Sciences Engineering and Technology*. **6**, pp.4492-4498.

Yunus, M. et al. 2010. Semantic query for Quran documents results. Open Systems (ICOS), 2010 IEEE Conference on, 2010. IEEE, 1-5.

Zarrabi-Zadeh, H. 2007. *Tanzil* [Online]. [Accessed 26 October 2014]. Available from: http://tanzil.net/