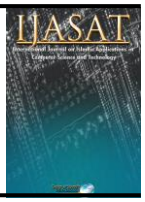

Semantic Web Application for Historical Concepts Search in Al-Quran.

Aliyu Rufai Yauri, Rabiah Abdul Kadir, Azreen Azman, Masrah Azrifah Azmi Murad
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia
rufaialeey@yahoo.com, {rabiah, azreen, masrah}@fsktm.upm.edu.my

ABSTRACT

With the growth of Islamic related documents on the web, several search engines exist today to pull out these documents such as data about Islamic history. However, most of the information retrieval on the web is based on keyword search and it is struggled with the ambiguity of natural language. To overcome this problem, a concept of semantic web is introduced by W3C consortium. The aim of this paper is to apply semantic web technology to retrieved historical concepts from the Holy Quran. The semantic web technology will improve the precision of knowledge retrieval from Quran. We used ontology and Jena inference engine to answer queries in natural language, which is related to the historical concepts in Quran. For the experiment, we used queries that were asked by the visitor of Islamic Research foundation website. The result shows that the precision is improved from traditional keywords search. This paper highlights the potential of semantic web technology in retrieving the related Islamic documents.

***Keywords*: Semantic Web, Ontology, Natural Language Queries, Inference Engine, Information Retrieval**

1. Introduction

Information technology has brought to the growth of Islamic related data on the web, databases and other digital and electronic devices. The web is a good source of information that is related to different Islamic topics (Subaimi bin Muhammad Sarif et al., 2010). Users prefer to access the information or knowledge of Islamic history in the Holy Quran that is related to their query via the web against the traditional way of watching or visiting Islamic scholars.

However, Islamic related information on the web and other storage mechanisms are not organized in chronical way. It is rather spiral or event based such as in the Quran (Naseem Shazadi et al., 2011). The web search systems that are based on keyword search have low precision, as a result of dealing with the ambiguity of natural language. Users may have to do several crawling on the search result before being able to find the related document of what he/she was intended to find. To deal with the limitation of the current web searching, semantic web technology was introduced by the World Wide Web Consortium. Semantic Web is referred to web of linked data, where data are structurally represented, based on Resource Description Format (RDF). The Goal of semantic web is made knowledge widely available and increases the utility of some knowledge by enabling advanced applications for

searching, browsing and evaluation (Pascal Hitzler et al., 2002). The main back bone of semantic web is ontology (Aliyu Rufai Yauri et al., 2013). Ontology can be seen as concepts, entity or objects that exist in domain.

This paper uses semantic web technology application to present a framework that identifies historical concepts from the holy Quran based on user natural language query. Applying semantic web technology to various heterogeneous Islamic information sources will ease access to Islamic data with higher precision.

We used existing Quran ontology, annotated the concepts to provide relationship between the concepts and store in knowledgebase. Knowledgebase contains annotated ontology that is represented in RDF. Natural language analysis to the user query has been done in order to match the query with the knowledge base for historical concepts retrieval. Although, recently researchers have intensify efforts in this area of research, however most of the existing works focus on certain surah or chapter of Quran with simple queries. However, this research covers the whole contents of Quran and analyzes user's complex Query containing word, phrase, sentence or paragraph.

The rest of this paper is organized as follows. Section 2 contains review of the current work. Section 3 introduces the methodology of the system. Section 4 presents the experiment and result; followed by evaluation in section 5. Finally, section 6 gives the conclusion and discusses future work.

2. Related Works

Semantic technology is an extension of the current web where data are linked to provide more defined meaning in order for human and computers work together (T. Berners-Lee et al. ,2001). One of the core elements of semantic web is the ontology. Ontology conceptualizes a domain into a machine-readable format. Ontology is a mechanism through which knowledge is represented in form concepts, nodes that linked relationships between these concepts, and restrictions.

Ontology-based information extraction system, is seen as a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontology (Daya C. Wimalasuriya et al., 2010). Current work on application of semantic web on Quran domain is mostly focusing on ontology creation itself, retrieving mechanisms but mostly base on section of Quran and single sentence Queries. Quran ontology creation was found in (Maha Al-Yahy et al., 2010), where the work uses automatic extraction method to acquire ontology from Quran and Hadith domain text. They focus mainly on salat or prayer concepts found in Quran and hadith. (Ontology Research team, UITM, 2010) presents a methodology for building ontology from Quran and hadith domain. They automatically generate ontology instances form unstructured document of Al-Quran, hadith and other related Islamic knowledge domain. Their system extracts concept and build the taxonomy of Islamic Knowledge. The approach was the integration of ontology learning, ontology population and text mining framework for the extraction of information from Islamic knowledge sources.

Another ontology development model can be seen in (Maha Al-Yahy et al., 2010). They presented a design and implementation of ontology model that uses time noun from the holy Quran to derive ontology structure.

Dataquest is a framework for modeling and retrieving knowledge from distributed knowledge sources primarily related to the Holy Quran (Qurat ul Ain et al., 2011). But this work was based on keyword search which allows users to enter a keyword or navigate through the form. It does not accept user Query with complex phrases.

Another related work is in (Sumayya Baqai et al., 2009). The work focuses on bilingual (English/Arabic) comprehensive search tool for the Holy Quran and the work was also using

keyword search. The users need to get the correct keyword in order to retrieve the desired information. A work that comes close to this work is in Leveraging Semantic Web technology for standardized knowledge modeling and retrieval of holy Quran and religious text that was presented in (Ivan Habernal et al., 2013). The system was designed for proving Quran knowledge sharing, storing, modeling, reasoning and retrieval from diverse Islamic domain sources. However, this work is differs from their work based on the ability of the system to answer a query with complex phrases. The users are allowed to ask any query in natural language by entering many words or sentences, as he/she desired.

3. System Overview

The system focus on search for historical information that is mentioned in Quran semantically. The users can ask any question concerning historical concept or object from Quran using complex phrases in natural language. The proposed system is able to process users' query and retrieve the related knowledge by matching the processed user query against the knowledge base. User's query, is processed by the Query Processing Model. The knowledge base contains Quran ontology that is represented in RDF/XML serialization format. The retrieving of answer for the user query is performed by Semantic Search Processing model. Figure 1, below depicts the flow of the system.

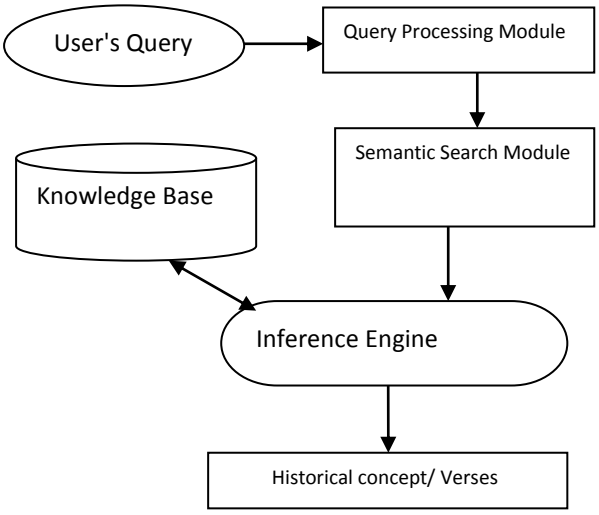

Fig 1.System Framework

3.1. Knowledge Based and Ontology

The knowledge base is designed to store Quran ontology which includes historical concepts in Quran and the annotation of these concepts. The main importance of a well-designed ontology is its precise semantics. Well-designed ontology, allows for complex semantic construct to be used in order to infer over the knowledge base. Our ontology engages with the historical concept that exists in the holy Quran (example: *names of prophets mentioned in Quran*). We annotate the concepts to provide relationship between concepts. This enables the use of semantic construction, in order to infer over the knowledge base.

Ontology in knowledge base is represented in RDF format. RDF represents ontology concepts in graphical form. The graph tied the concepts in triple form [*subject, predicate, object*]. Subject and object are brother concepts, whereas predicate shows the relation of them. Relationships are in two forms, namely object property and data property.

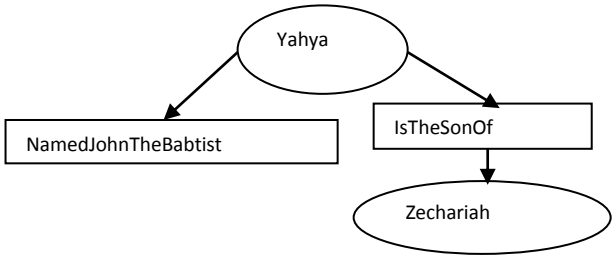

Fig 2. Concept Annotation

Figure 2, above shows the graphical representation of triples which can be seen below:

Subject *predicate* *Object*
Yahya ⟶ *IsThefatherOf* ⟶ Zechariah

The predicate of *IsThefatherOf* relates the concepts of *Yahya* and *Zechariah*, which are both concepts in the Quran ontology.

Subject *predicate*
Yahya ⟶ *NamedJohnTheBabtist*

The predicate *NamedJohnTheBabtist* is literal of *Yahya*, which gives more description to *Yahya*. The literal is known as data property, which gives more description to the concept Yahya. Below is an example of the query that can be asked to the system, based on the above semantic relation of the data. *Example of query:*

"I want to know who is the son of Zachariah?"

Or

"Who was the person that built the great wall to stop Gog and magog?"

In the knowledge base we have a collection of triples that are constructed to infer on these concepts, therefore in both queries "Yahya" will be retrieved. The system also will present verses in Quran that mention "prophet Yahya" to the user. Since ontology stores the knowledge base are represented in triple form, user's natural language query also need to be represented in triple form in order to gain access to knowledge in the knowledge base. We will see in details how user's natural language query is processed into triple representation in the paper.

3.2. Query Processing Module

Query processing module analyzes the user's query and reformulates it into the same representation as in knowledge base, which is in RDF.

Example of query:

"I want to know who the son of Zachariah is."

What the system does is to implement the natural language processing to normalize the query. The query will go through several processes in NLP, such as stop word removal, lemmatization and part of speech tagging. Since the historical concepts are nouns, the idea behind part of speech tagging is to identify noun concepts from user query word and to implement it. Look up keyword in the knowledge based was executed to see nouns from the user query that exist in the knowledge base. From the query above, *Zachariah* is a noun and it exists in the knowledge base as a concept. The system will automatically identify *Zachariah* as a concept.

The system will automatically search into the knowledge base, identifies concepts that have relations with *Zachariah* and uses n-gram maximum likelihood estimate model. The

model avails us to compute the estimate for the parameters of an N-gram model by normalizing count from training set, and normalizes them so that they lie between 0 and 1. We compute the probability of this concept that is related with *Zachariah* based on the remaining user tokens.

From the above example of query, only *Zachariah*, is identified as a concept in the knowledge base. Therefore, we are left with the following tokens:
"I want to know who is the son of"

Then, we use these remaining tokens to compute the possibility of word/phrase/sentence that form the relations between *Zachariah* with another concept in the knowledge base. We use n-gram maximum entropy model for this task. We rank the possible outcomes and take word/phrase/sentence with higher score for further processing.

After the computation, the system was able to identify [**?,** IsthesonOf, Zachariah] with isthesonOf giving the result the highest score. The system automatically parses the formulated triple to Semantic Search Model in order to retrieve the answer.

3.3. Semantic Search Module

The main function in this module is to ensure that the best possible reformulated query is parsed to the inference engine and retrieve the answers. For inference, the system uses Jena API which is a Java API inference engine for semantic web applications (Apache Jena, 2010). It contains interfaces for representing RDF graph containing concepts, relations, literals, and all the other key concepts of RDF.

Jena read RDF graph from the knowledgebase and match against the user query. This enables the retrieval of concepts located in the knowledgebase.
Example: With the computed triple by the system as follows [?, IsthesonOf, Zachariah], we have a relation and an object identified by the system. However, we do not know what the subject is. Then, we parse it to the Semantic search module to infer *Subject* that has relation: *IsTheSonOf*
Withthe Object *Zachariah*, the system parses to Jena to infer the right answer.

Firstly, Jena read the entire knowledge base, i.e. it reads all the RDF graph and stores in a Model; it uses different accessing methods to access information stored in the Model. System can retrieve either the subject, a relation which is the property, or the object depending on the type of query semantic search model pass to the Jena API inference engine.

In our case, semantic search parses the triple [?, IsthesonOf, Zachariah] to Jena inference engine. Jena reads the model and search any match form in the knowledge base where [Yahya, IsTheSonOf, Zachariah]. Then, Jena infers that *Yahya* is the son of *Zachariah*. In the knowledge base we have annotated each concept with corresponding verses in the Quran that is mentioned as individual. In this case, therefore, the system will return *Yahya* and all the related verses, [Q: 3:39, 6:85, 19:7, 19:12, and 21:90]

4. Experimental results

We used a total sum of 30 queries for this experiment. Table1 shows examples of the queries and answers that we were able to retrieve.

Table1: Query sets and Result sample

Query	Concept	Verses retrieved
I want to know, who is the father of Prophet Yahya?	Zechariah	6verses(Q3:37,3:38, 6:85,19:2,19:7, 21:89)
Please can you tell me who the son of prophet Zachariah is?	Yahya	5verses (Q: 3:39,6:85, 19:7, 19:12, 21:90)
I heard that Maryam was staying in the mosque along side with her guardian or father, who is the guardian or father of Maryam?.	Imran	3 verses (Q:3:33,3:35,66:12)
Who was the person sent to people of Thamud?	Salih	8verses(Q: 7:73,7:75, 7:77,11:61,11:62, 11:66, 11:89:26:142)
Quran mentioned different people that include prophets and non-prophet. I would like to know whose name was mentioned the most in Quran?.	Musa	136verse (example:Q2:51,2:53,2:54,2:55)

The system's evaluation was based on the relevance of the results retrieved by the system.We used popular precision and recall technique (Daniel Jurafsky et al., 2009).

$$precision = \frac{|\{relevantdocuments\} \cap \{retrieveddocuments\}|}{|\{relevantdocuments\}|} \quad \ldots\ldots (1)$$

$$recall = \frac{|\{relevantdocuments\} \cap \{retrieveddocuments\}|}{|\{totalrelevantdocuments\}|} \quad \ldots\ldots (2)$$

Precision and recall methods was used to measure effectiveness of the search system. Recall measure of how many of the relevant documents were retrieved, while precision measure of how many of retrieved document were relevant.

Table 2.Results

Queries	Average Precision	Average Recall
30 Queries	0.92	1.0

The system shows that application of semantic web technology Islamic related document improves precision (as shown in Table 2) against the traditional keyword search. The system has return result of 92% precision and 100% recall when applied on queries to retrieve historical concepts from the Holy Quran.

5. Conclusions and Feature Work

In this paper, we presented semantic web application for historical concepts search in Quran. The system analyzes the natural language queries and match them against the knowledge base to retrieve historical data from the holy Quran by employing semantic web applications capability such as ontology annotations and use of inference engine to infer over given facts. The system is able to accept user query of any length and retrieve answer with good precision and recall.

However, the system has not highlighted what should be done, if the system is not able to get correct translation of user query into knowledge base representation. This problem will be focused on in future research.

Reference

Subaimi bin Mhd Sarif, Yusuf bin Ismail. (2010).Data mining through Internet Search Engine: The case of for Islamic Management Material. Knowledge management International Conference, 25-27 may, 2010. Kuala Terengganu.

Naseem Shazadi, Atta-ur-rahman, Adil Shaheen. (2011). Semantic Network based Semantic Search of religious Repository. International Journal of computer Application (0975-8887).

Pascal Hitzler, Markus Krotzch, sebastin Rudolph. (2002). Foundation of Semantic Web Technologies.Boka Raton London New York. Chapman & Hall/CRC.

Aliyu Rufai Yauri, RabiahA.Kadir, AzreenAzman, Masrah Azrifah Azmi Murad. (2013).Quranic verse extraction base on concepts using OWL-DL ontologies. Research journal of Applied sciences, Engineering and Technology.

T. Berners-Lee, J.Hinder, O.Lassila, (2001).Semantic Web, Scientific American, 2001.

Daya C. Wimalasuriya and Dejing Dou. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 2010 36: 306 originally published online 19 March 2010.

Maha Al-Yahy, Hend Al-Khalifa , Hend Al-Khalifa , Alia Bahanshal, Iman Al-Odah , Nawal Al-Helwah.(2010). An Ontological Model for Representing Semantic Lexicons: An Application on Time Nouns in the Holy Quran. The Arabian Journal for Science and Engineering, Volume 35, Number 2C, 2010.

Ontology Research team, UITM. (2011). Ontology Extraction Tool for Quran and Hadith using Fuzzy-Swam Algorithms. Available at http://webs.cs.utm.my/ontologyGroup/index.php/proposal1(accessed 25 Sep 2011).

Qurat ul Ain, Amna Basharat.(2011). Ontology driven Information Extraction from the Holy Qur'an related Documents. 26th IEEEP Students' Seminar 2011 Pakistan Navy Engineering College National University of Sciences & Technology

SumayyaBaqai, AmnaBasharat, Hira Khalid, Amna Hassan, ShehneelaZafar. (2009) .Leveraging Semantic Web Technologies for Standardized Knowledge Modeling and Retrieval from the Holy Quran and Religious Texts.ACM 978-1-60558-642-7/09/12 2009.

Ivan Habernal, Miloslav Komopik. (2012). SWSNL: Semantic Web Search Using natural language. Expert Systems with Application 40(2013) 3649-3664.

Apache Jena.http://jena.apache.org/. Accessed 12 may 2013.

Daniel Jurafsky, James H. Martin (2009). Speech and Language Processing. USA: Pearson Education International.