Ontology-Based Question Answering System in Restricted Domain

Rosmayati Mohemad, Noor Maizura Mohamad Noor, Noraida Haji Ali and Ee Yong Li
Software Technology Research Group, School of Informatics and Applied Mathematics,
Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.
rosmayati@umt.edu.my

***Abstract*—The complexity of natural language presents difficult challenges that traditional Questions and Answers (Q&A) system such as Frequently Asked Questions, relied on the collective predefined questions and answers, unable to address. Traditional Q&A system is unable to retrieve exact answer in response to different kind of natural language questions asked by the user. Therefore, this paper aims to present an architecture of Ontology-based Question Answering (OQA) system, applied to library domain. The main task of OQA system is to parse question expressed in natural language with respect to restricted domain ontology and retrieve the matched answer. Restricted ontology model is designed as a knowledge base to assist the process based on the effective information derived from the questions. In addition, ontology matching algorithm is developed to deal with the question-answer matching process. A case study is taken from the library of Sultanah Nur Zahirah of Universiti Malaysia Terengganu. A prototype of Sultanah Nur Zahirah Digital Learning ONtology-based FAQ System (SONFAQS) is developed. The experimental result shows that the architecture is feasible and significantly improves man-machine interaction by shortening the searching time.**

***Index Terms*—Knowledge Engineering; Library; Ontology; Question Answering System.**

I. Introduction

The explosive growth of information available on World Wide Web has been attracting many people to rely on Question and Answers (Q&A) sites for querying answers to their questions. One of the common and straightforward traditional Q&A is Frequently Asked Questions. Q&A is identified as the strategies to overcome technical limitation where it provides a textual collection of expert answers to a list of common questions in about specific domain that users might frequently ask. According to Romero et al. [1], Q&A is used to reduce cost of technical support.

Most of traditional Q&A systems, however, require user to search for an answer to a particular question manually or has to browse through a long list of Q&A collection to find relevant questions and answers. The systems provide no effective mechanisms to assist the user in obtaining useful information from Q&A knowledge database [2]. In addition, current Q&A systems do not have any features to handle questions with incomplete information, thus reducing the retrieval accuracy of the matched answer. The retrieval result is less accurate since the systems only match directly with the questions in the database without any inference. This causes the inability of specifying a query that can meet more user's requirements than the extracted answer, thus hindering the relevant information to be retrieved [1]. Moreover, the manual categorization of Q&A systems is based on human judgment where the questions and answers are compiled in a database and reused the answer when similar questions occur, thus causes inconsistencies in matching process [3].

The current way that is being adapted to retrieve relevant information in accord to user's requirements is by integrating ontology in Q&A system [4, 5]. The purpose is to accumulate knowledge and semantically analyze the questions provided by the user. In this study, the concept of ontology is employed as the key technique to support user queries in Q&A system. Ontology is an emerging technology for representing a particular domain knowledge in a meaningful way that can be understood and manipulated by machine by structuring knowledge into a formal conceptualization. The main advantage of using ontology is to provide accurate answers by analyzing the questions in natural language [6].

In this paper, an Ontology-based Question Answering (OQA) System is proposed to handle current constraints in traditional Q&A systems. By using the ontology approach, it can contribute to shorten the searching time and improve the accuracy rate of retrieving the answers. Here, the prototype of Sultanah Nur Zahirah Digital Learning ONtology-based FAQ System (SONFAQS) is developed using real case study in library domain, taken from the library of Sultanah Nur Zahirah of Universiti Malaysia Terengganu. The main goal is to assist the librarians by automatically searching the questions and answers according to users' specific aspect.

The paper proceeds in the following manner. Section II presents the related works. Meanwhile, the architecture is detailing out in the Section III. Section IV reports the implementation of the Ontology-based Question Answering prototype. Finally, Section V concludes with a summary of this paper and future research directions.

II. Related Works

The importance of ontology in categorizing and structuring domain knowledge is exploited in Q&A systems. Most ontology-based Q&A systems categorize the specific domain knowledge into ontology structures and then, a list of questions and answers is composed based on the created ontology. This can be seen through a number of Q&A systems developed and researched in previous studies using open domain ontology such as AQUA [7], QASYO [6], Pythia [8] and NLQA [9]. Meanwhile, in a restricted domain, various and diverse ontology-based Q&A systems have been presented such as in medical [10], biology [11] and physic [12].

AQUA is an experimental Q&A system developed by Vargas-Vera and Motta [7] based on the combination of

natural language processing, ontology, logic and information retrieval techniques. The system has been tested to answer questions about academic people and organizations. Here, ontology is used to formulate the natural language query in the ontological structures. In addition, most similar Q&A system has been proposed by Moussa and Abdel-Kader [6], namely QASYO where it used YAGO ontology as the background knowledge.

Meanwhile, Pythia is an ontology-based Q&A system provided by Unger and Cimiano [8]. It is able to parse constructed complex natural language questions and then can subsequently translate into formal queries with respect to a grammar that has been composed in the ontology. NLQA is another Q&A system that has been proposed by Athira et al. [9]. Natural language technique was used to analyze a single complex question using both syntactic and semantic methods and decompose it into a set of less complex queries using ontology and morphological expansions.

These ontology-based Q&A systems highlight the essential role of ontology to improve the retrieval accuracy of the answers based on the questions asked in either in open domain or restricted domain. In order to realize the implementation of ontology-based Q&A system, it also involves natural language processing and information retrieval technologies. However, as far as we are concerned, none of the previous researches provide a conceptual architecture to serve as a comprehensive guideline for building the ontology-based Q&A system in library domain.

III. Architecture of Ontology-Based Question Answering System

An architecture of Ontology-based Question Answering (OQA) system is proposed as the main guideline in designing OQA system for retrieving short answer using natural language query in a specific domain, composed in the ontology. Figure 1 shows the OQA architecture. The proposed architecture is focused on natural language questions.

This section describes in detail the components of the OQA architecture. It is characterized by the three following essential components; the question parsing, the ontology building and the ontology matching. These three components are the mechanisms to retrieve the relevant answer from natural language question asked. Each component of the architecture is fully elaborated in the subsequent subsections.

A. Question Parsing

The first component of OQA architecture is the question parsing, in which the natural language question is resolved into logical syntactic form. The component performs three processes; tokenizing, removing stop word and stemming. An indicator list that consists of stop words is provided for this component to assist parsing process. String tokenizing and Porter Stemmer algorithms are adopted in this component. The reason this component is necessary is to reduce the high dimensionality problem of processing natural language questions. With this component, any irrelevant words are removed. The outcome resulting from this component is parsed question that has been stored as a set of tokens.

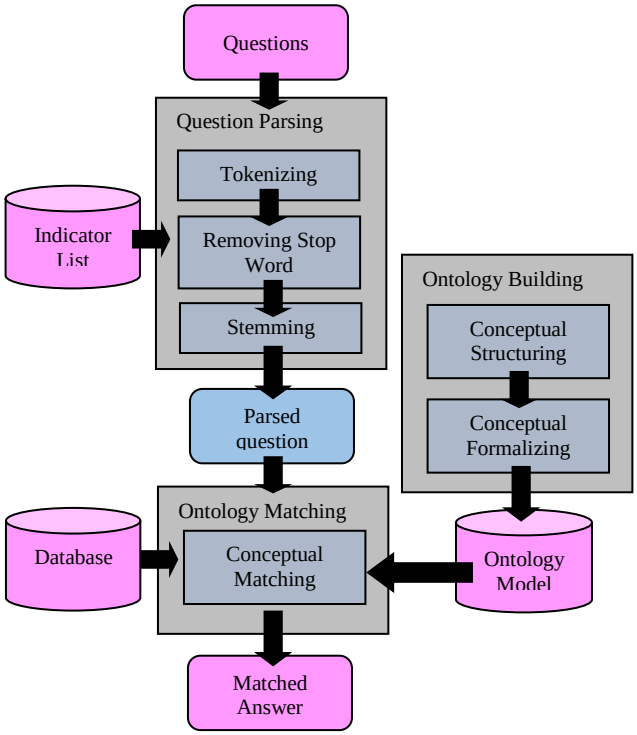

Figure 1: Architecture of Ontology-based Question Answering (OQA) System

String tokenizing algorithm is employed to tokenize the input question by splitting up the sentence into words and removing whitespace between each word. Figure 2 depicts the excerpt of string tokenizing algorithm. A set of tokenized words then is used as input for the next process, which is removing the stop word. Here, the tokenized words are compared with the stop words list. Any tokenized word that is matched with the stop words list is removed from the tokenized words. Figure 3 describes the details of stop word removal algorithm.

1. Read input question
2. Call StringTokenizer class
3. Split question into token
4. Remove whitespace token
5. Write the extracted token into new file
6. Check is there any other token
7. If yes, write the other token into next line
8. If no, end

Figure 2: String Tokenizing Algorithm.

1. Read extracted token
2. Compare the token with stop word list
3. If match, ignore and proceed to the next token
4. If not match, write the token word into new text file
5. Repeat step 2-4 for the next token
6. End

Figure 3: Stop Word Removal Algorithm

The last process in the question parsing component is stemming. Here, Porter Stemming algorithm as proposed by Hooper and Paice [13] is applied. Figure 4 shows the excerpt of Porter Stemming algorithm. There is a suffix rule lies inside the algorithm. If the tokenized word extracted from the previous process matches to a suffix rule, the word is then tested with the conditions that lie inside the rule. Once the condition passed and accepted, the suffix is removed, and the

word is passed to the next step to test with other suffix rules.

1. Read the extracted token and store into an array
2. Call and pass the array value to the stem class
3. Go through the first rule of the Porter Stemmer
4. Match the word with the rule
5. If match, test the word with the conditions in the rule
6. If condition passed, remove the suffix
7. Proceed to the next rule
8. Repeat step 4-7 for the second
9. Write the stemmed word into new file
10. End

Figure 4: Porter Stemming Algorithm.

B. Ontology Building

The second component of the architecture deals with the ontology building. The outcome of this component is a list of concepts and properties formalized in an ontology model. The model is required to support conceptual matching. The ontology building component consists of two main processes; conceptual structuring and conceptual formalizing.

Conceptual structuring process is intended to organize and structure the knowledge acquired from domain experts of a particular domain into a list of concepts during knowledge acquisition activity. List of keywords are identified based on predetermined questions and answers obtained from the domain experts. Once a list of keywords is built, these keywords are then transformed into formalized model through conceptual formalizing. Here, ontology model can be formalized using existing ontology tools such as Protégé, TopBraid Composer and others.

C. Ontology Matching

The architecture's third component is the ontology matching to detect matched answers based on parsed queries and ontology model through the conceptual matching process. Java Regex API is used as the method for conceptual matching. Regex stands for Regular Expressions is the way to describe a set of strings based on the common characteristics shared by other string in the set [14]. Here, Pattern and Matcher classes are used. A Pattern object is a compiled representation of a regular expression or text while a Matcher object is the engine that interprets the pattern and performs matching operation against a string of input text. Figure 5 explains the details algorithm for conceptual matching.

1. Connect to database
2. Retrieve all elements from database and store into an array
3. Create empty ontology model
4. Read ontology file into the empty model
5. Store subclasses in the ontology model into an array
6. Compile the subclass as the pattern for matching
7. Read the keywords and store in an array
8. Match each keyword with the pattern created in step 6
9. If match,
   a. store subclass of the matched keyword
   b. store the instances of the subclass
10. Match the instances with retrieved elements
11. If match, retrieve the questions and answers of the matched instance
12. End

Figure 5: Conceptual Matching Algorithm

First, an empty model is created using Jena library together with Java Programming Language to model the ontology created. During this step, the classes in the ontology model are listed out and saved as a text file for matching purpose later.

Then, by using the Pattern and Matcher class in the Java Regex library, the stemmed keyword file is then read and matched by matching the stemmed keywords with the class file created earlier. If there is any match found, the matched class is further extracted again to retrieve any subclass of the matched class.

Next, the retrieved subclasses are matched with the keyword list to find any possible keyword matched, and if there is any matched, the instances of the matched subclass are listed out. Instances here are the answers for each respective class and the subclass. Then, the instances are saved in a text file. After the concept or keywords had been matched with the ontology model, the answers that matched with the concepts or keywords are then extracted from the database.

IV. ONTOLOGY-BASED QUESTION ANSWERING PROTOTYPE

This section discusses a prototype that has been developed based on the of Ontology-Based Question Answering architecture proposed. A prototype of Sultanah Nur Zahirah Digital Learning ONtology-based FAQ System (SONFAQS) is tested using a real case study in library domain, taken from the library of Sultanah Nur Zahirah of Universiti Malaysia Terengganu.

A. Library Ontology Model

Library ontology model is the resulting outcome of ontology building. Here, a group of domain experts, consisting of librarians is identified to assist in conceptual structuring process. List of frequently-asked questions and possible answers are analyzed with the help of librarians. There are five main activities that need to be completed during conceptual structuring; building glossary of terms, building concept taxonomies, building ad hoc binary relation, building concept dictionary, and defining in details the ad hoc binary relations, instance attributes, class attributes and constant that are identified in the concept dictionary. Figure 6 depicts the semantic network of the relationship between concepts in library ontology model.

Conceptual formalizing involves the construction of library ontology model using the TopBraid Composer (TBC). Based on each question and the answer respectively, the main keywords that can represent the question and answer is analyzed and generated. Then, the keywords are used to construct the subclasses, instances, and relationship or so-called properties in the ontology.

B. SONFAQS prototype

SONFAQS prototype is developed using Java Server Pages (JSP). Figure 7 shows the main interface of the prototype. Here, the user is allowed to input any natural language question on library domain, i.e., "*How to return book via book drop?"*. Then, the question is being passed through the question parsing engine. Firstly, the engine will tokenize, remove any stop words and stem the question in order to identify main keywords to be passed.

Subsequently, the parsed question in the form of keywords will be matched to pre-defined concepts in the library ontology model in order to extract the relevant answer for the particular question asked. If any matched keyword found, the answers and questions that are under the matched keyword are retrieved and displayed. Figure 8 depicts the interface for displaying the answer retrieved from the question asked. The

result shows that there are two questions and answers that have the same keyword.

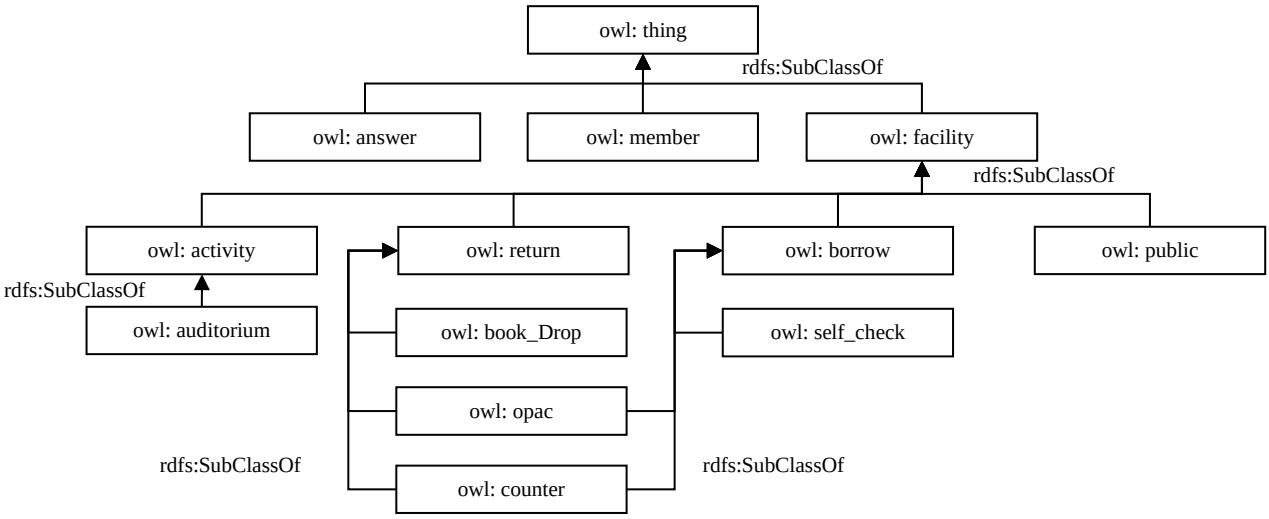

Figure 6: Ontology building process using TopBraid Composer

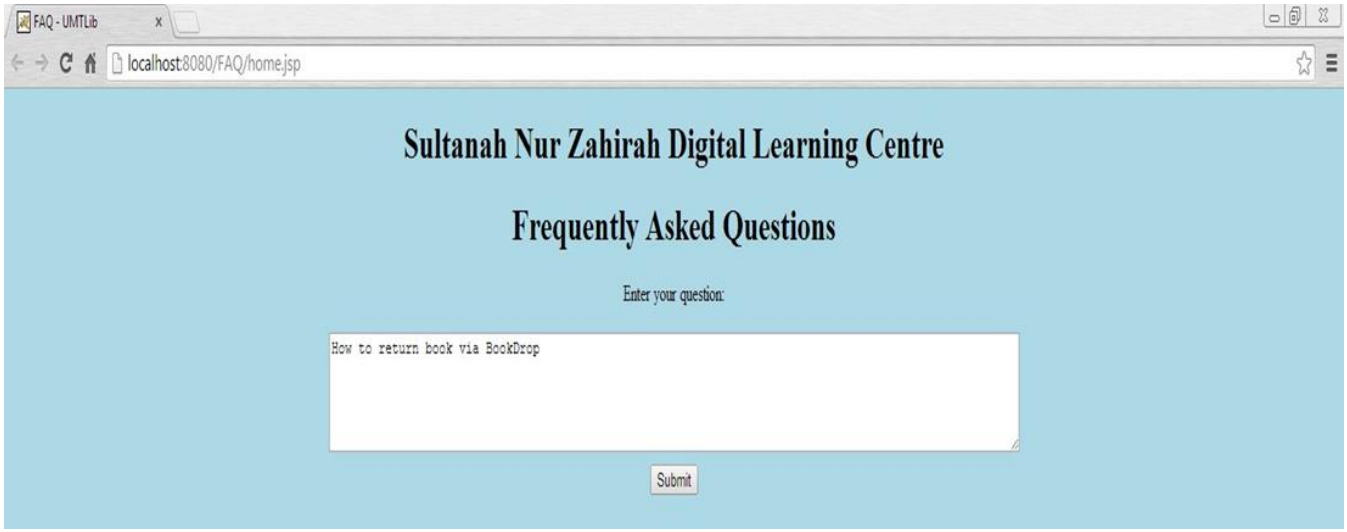

Figure 7: The Main Interface of Sultanah Nur Zahirah Digital Learning ONtology-based FAQ System (SONFAQS).

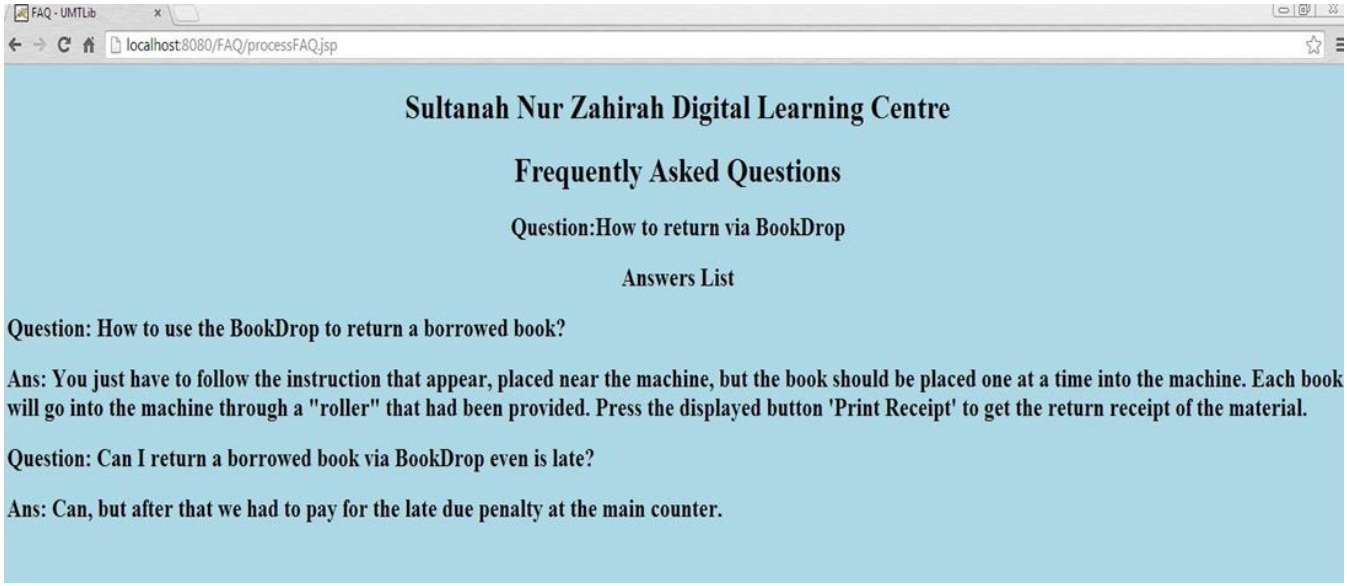

Figure 8: The Retrieved Questions and Answers

V. Discussion and Conclusion

This paper presents an architecture of ontology-based question answering system. In this section, a comparison is done between the old FAQ system used in the library of Sultanah Nur Zahirah with SONFAQS. For the old system, hands-on approach is still being used to find any answers for the questions asked by the users. Therefore, the process of finding answers is tedious and time-consuming. The implementation of SONFAQS replaced the hands-on approach where the users or librarians just need to input any questions in natural language form, and the answer is searched automatically. Hence, the SONFAQS helps in reducing times and increase in efficiency on handling the Q&A process. The comparison is summarized in Table 1.

Table 1
Comparison of the existing system with SONFAQS.

Old FAQ System	SONFAQ System
A hands-on approach whereby the librarian needs to find manually for the answers to the asked questions.	System approach whereby the system helps the user to find answers respectively to the questions asked.
Longer time for answers searching.	Shorter time for answers searching.
Less efficient in handling the Q&A process.	More efficient in handling the Q&A process.

For future works, we will consider to use a better stemmer algorithm, implement queries to construct ontology model through prototype and queries that can identify words with multiple meanings or synonyms, and make a comparison on them.

Acknowledgment

This research is supported by the Malaysian Ministry of Higher Education, Fundamental Research Grant Scheme (FRGS) vote 59395.

References

[1] M. Romero, A. Moreo, and J. L. Castro, "A cloud of FAQ: A highly-precise FAQ retrieval system for the Web 2.0," *Knowledge-Based Systems*, vol. 49, no. 0, pp. 81-96, 2013.

[2] S.-Y. Yang, F.-C. Chuang, and C.-S. Ho, "Ontology-supported FAQ processing and ranking techniques," *Journal of Intelligent Information Systems*, vol. 28, no. 3, pp. 233-251, 2007.

[3] R.-S. Shaw, C.-F. Tsao, and P.-W. Wu, "A study of the application of ontology to an FAQ automatic classification system," *Expert Systems with Applications*, vol. 39, no. 14, pp. 11593-11606, 2012.

[4] G. Suresh kumar, and G. Zayaraz, "Concept relation extraction using NaÃ¯ve Bayes classifier for ontology-based question answering systems," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 1, pp. 13-24, 2015.

[5] Y. Cao, F. Liu, P. Simpson *et al.*, "AskHERMES: An online question answering system for complex clinical questions," *Journal of Biomedical Informatics*, vol. 44, no. 2, pp. 277-288, 2011.

[6] A. M. Moussa, and R. F. Abdel-Kader, "QASYO: A question answering system for YAGO ontology," *International Journal of Database Theory and Application*, vol. 4, no. 2, pp. 99-112, 2011.

[7] M. Vargas-Vera, and E. Motta, "AQUA - Ontology-based question answering system," in *MICAI 2004: Advances in Artificial Intelligence*, R. Monroy, G. Arroyo-Figueroa, L. E. Sucar, and H. Sossa, Eds. Springer Berlin Heidelberg, 2004, pp. 468-477.

[8] C. Unger, and P. Cimiano, "Pythia: Compositional meaning construction for ontology-based question answering on the semantic web," in *Natural Language Processing and Information Systems*, *Lecture Notes in Computer Science*, R. Muñoz, A. Montoyo, and E. Métais, Eds. Springer Berlin Heidelberg, 2011, pp. 153-160.

[9] P. M. Athira, M. Sreeja, and P. C. Reghuraj, "Architecture of an ontology-based domain-specific natural language question answering system," *International Journal of Web & Semantic Technology*, vol. 4, no. 4, pp. 31-39, 2013.

[10] A. Ben Abacha, and P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," *Information Processing & Management*, vol. 51, no. 5, pp. 570-594, 2015.

[11] M. Neves, and U. Leser, "Question answering for Biology," *Methods*, vol. 74, pp. 36-46, 2015.

[12] A. Abdi, N. Idris, and Z. Ahmad, "QAPD: An ontology-based question answering system in the physics domain," *Soft Computing*, In Press, pp. 1-18, 2016.

[13] R. Hooper, and C. Paice, "The Lancaster Stemming Algorithm," University of Lancester, 2005, Avaible at http://www.comp.lancs.ac.uk/computing/research/stemming/. [Retreived 09 November, 2014]

[14] Oracle. "The Java Tutorials. Regular Expressions," 2014. Available at http://docs.oracle.com/javase/tutorial/essential/regex/intro.html. [Retrieved 9 June 2014].