AN ONTOLOGY-BASED HIV/AIDS FAQ RETRIEVAL SYSTEM

Yirsaw Ayalew, Gontlafetse Mosweunyane, Barbara Moeng
Department of Computer Science, University of Botswana
Private Bag UB00704, Gaborone, Botswana
{ayalew, mosweuny, motswirib }@mopipi.ub.bw

ABSTRACT

This paper presents a discussion of the implementation of an ontology-based HIV/AIDS Frequently Asked Question (FAQ) retrieval system. The main purpose of the system is to provide an answer from an existing HIV/AIDS FAQ repository for any question on HIV/AIDS asked by any person. As the identification of the best possible answer requires the understanding of the semantics of both the question and the existing question-answer pairs in the FAQ, the use of ontology is very crucial. Ontologies have been widely used in natural language processing applications especially in Question Answering Systems. The ontology for the HIV/AIDS FAQ retrieval system has been built using Text2Onto tool which has been experimentally evaluated to be the most appropriate tool as reported in our earlier work.

Once the ontology is constructed, the next challenge is to make sure that the use of the domain ontology improves the performance of the FAQ retrieval System. For this purpose, we explored a number of approaches for computing semantic similarity between a user query and the existing question-answer pairs in the FAQ. Semantic similarity is computed based on inherent relationships between concepts using ontologies. Specifically, we use the semantic similarity metrics proposed by Thiagarajan et al. based on spreading activation networks (set based spreading). The results show an improvement in accuracy compared to the traditional information retrieval based question answering systems approaches.

KEY WORDS

FAQ retrieval system, Question answering system, HIV/AIDS ontology, Semantic similarity.

1. Introduction

HIV/AIDS has affected Sub-Saharan Africa more than any region in the world. Among the Sub-Saharan Africa countries, Botswana is one which is highly affected by this pandemic. To tackle this challenge, one strategy is to educate the population and increase awareness through the provision of access to information resources. Currently people can get information about HIV/AIDS from various sources including through FAQs website and HIV/AIDS Call Centres. FAQs and call centers provide question answer service.

Question answer services are becoming popular due to their ability to provide specific answers to users questions as opposed to list of potential answers as in search engines. In other words, users can directly obtain answers rather than a list of potentially relevant documents. For this reason, organizations provide FAQs to accommodate the common user questions about an organization's services or products or anything related to the particular organization. However, the FAQs that are provided on organizations websites require users to go through the FAQ question-answer pairs to find an answer for a question a user has. An ideal solution would be to allow users to pose just their questions and a system scan the FAQs and return the answer from the question-answer pairs for which the user's question and the question in the FAQ are identical or similar.

The purpose of an HIV/AIDS call center is to provide information appropriate to individual demands. In a call center setup, people call a toll-free line managed by the call centre and ask any questions related to HIV/AIDS they may have. The operator browses the HIV/AIDS frequently asked questions (FAQ) manual and provides the answer to the caller. If the answer is not in the manual the operator escalates the question to an HIV/AIDS specialist. The caller will be advised to call again at a later time. Once the answer is provided by the HIV/AIDS specialist, the question answer will be included in the FAQ manual. This setup, though helpful in many aspects, it still has a number of inconveniencies.

A more convenient solution would be to get the question answer service through mobile phones just by sending SMS (Short Message Service) questions. The ultimate goal of our research project is to develop a question answering (QA) system that can answer any question people may have about HIV/AIDS through standard mobile phones. With such a system, people can send SMS questions using mobile phones and get the answer as an SMS on their cell phone. In this paper, we focus on the development of an automated FAQ retrieval system (a special type of question answer service) on HIV/AIDS. One of the major tasks in an FAQ retrieval service is to find questions in the FAQ repository that are semantically similar to a user's question.

An automated FAQ retrieval system will automatically search the FAQ repository to see if the same or similar question exists in the repository. If the same or similar question is found, then the corresponding answer can be provided. However, determining the semantic similarity between a user question and questions in the FAQ repository is a difficult task. The difficulty is

DOI: 10.2316/J.2014.216.815-0008

due to the fact that the same question can be expressed using different words which have similar meanings. To address this issue, a number of approaches have been proposed to improve the accuracy of measures of similarity between user question and FAQ questions..

In question similarity computation, most FAQ retrieval systems employ either statistical similarity or semantic similarity or a combination of statistical and semantic similarity. The main difference among the different approaches is in the computation of semantic similarity. As ontologies are designed to capture inherent relationships among concepts, semantic similarity computation based on ontology has a potential to provide a more accurate measure of similarity. In this paper, we discuss the application of the semantic similarity metrics proposed by Thiagarajan et al. [1] based on spreading activation networks for HIV/AIDS FAQ retrieval. Spreading is the process of including the terms that are related to the original terms in an entity's description by referring to domain ontology.

The remainder of the paper is organized as follows: Section 2 presents related works on automated FAQ retrieval systems by emphasizing on the techniques used for the computation of semantic similarity. Section 3 provides a discussion of the semantic similarity measure used in this paper together with the ontology developed for the HIV/AIDS FAQ retrieval. A discussion of the results of our empirical study and issues associated with semantic similarity computation is provided in Section 4. Finally, the main points of the paper and highlights of our future works are presented in Section 5.

2. Related Work

In this section, we discuss research work carried out in the area of question answering systems specifically FAQ retrieval systems as the ultimate goal of our research is to develop automatic SMS-based FAQ retrieval system in the area of HIV/AIDS. Question answering systems can be either general purpose (i.e., open domain) or specialized (i.e., closed/restricted domain). Our focus will be on closed domain question answering systems as we focus on a specialized question answering system (i.e., FAQ) in the area of HIV/AIDS. In FAQ based question answering, the FAQ provides a ready made database of question-answer pairs. Therefore, the main task will be to find the closest matching question in the FAQ to retrieve the relevant answer for a given user question. The main difference among the different systems lies in the techniques used to evaluate the degree of similarity between user questions and questions in the FAQ repository. In this section, while discussing prior works, emphasis will be on the similarity measures employed by different systems. In addition, the related works that we discuss are closed domain FAQ retrieval systems. The technique used to compute the degree of similarity between a user question and questions in an FAQ repository is the key for accuracy of the answers.

Kothari et al. [2] provided an algorithm to determine the closest matching question from an FAQ repository to a user question by using a scoring function that assigns a score to each question in the FAQ repository. Their approach is that all the terms in the FAQ repository are put in a dictionary and then the degree of similarity between the terms of a user question and the terms in a dictionary is computed. The question in the FAQ that has the highest degree of similarity to the terms of the user question is retrieved. To accomplish this, the authors introduced a number of functions to compute similarity at a term level and at a question level. Their empirical evaluation of two FAQ repositories and comparison to Lucene's Fuzzy match feature indicates that their system can be very effective in automating SMS based FAQ retrieval.

Thiagarajan et al. [1] proposed an approach to compute semantic similarity between two entities (described using bag of words) using spreading process. Spreading is the process of including the terms that are related to the original terms in an entity's description by referring to domain ontology. Such spreading process results in an extended set of terms consisting of the original terms and those terms related to the original terms. The spreading process is meant to capture inherent relationships between concepts so that content matching is more accurate. The use of ontology becomes more useful in this approach as domain ontology holds knowledge about terms/concepts and their relationship with other terms/concepts. To determine similarity of entities, cosine similarity (commonly used in Information Retrieval) technique can be applied to the extended set. Empirical studies on user profile matching scenario show that this similarity computation provides more accurate measure of similarity compared to human-computed similarity.

Even though the spreading process was used for computing similarity between two user profiles, we found it useful for computing the similarity between a user query and FAQ questions.

Song et al. [3] introduced a technique for the computation of similarity of a user question and questions in FAQ repository by combining statistical and semantic similarity measures. They compute the two similarity measures separately before aggregating the result using a linear combination of the two similarity values.

For the computation of statistical similarity, they used dynamically formed vectors to avoid sparse vector space problem. For the computation of semantic similarity, they used bipartite mapping based on the hierarchical network structure of WordNet to compute semantic similarity using the minimum length path to measure the similarity of two words. The similarity of two questions is computed by summing up the semantic similarities between the words they have

For the evaluation of the effectiveness of their system, they used 58 modified questions from an FAQ repository which consists of 500 question-answer pairs. They considered only those questions which have similar questions in the FAQ repository. They considered only the top 1 answer for any given question. The results show a success of 64.3% which is quite good.

In Wang et al. [4] a semantic similarity measure was introduced based on domain ontology for an agricultural FAQ retrieval system. To compute the similarity between user query and FAQ questions, the questions are first classified into categories. Then the similarity between user query and the categories is determined based on the similarity between the concepts of the user query and the concepts of the category. Then comparison with the FAQ questions in that category is carried out.
Their experimental results show that their system performs better than keyword based retrieval. They achieved a performance of 78% recall and 82% rejection.

In [5] an SMS-based FAQ retrieval was proposed based on traditional information retrieval techniques such as unigram matching, bigram matching, and 1-skip bigram matching. These approaches are mainly focus on statistical similarity based on direct matches and use of synonym s and hyponyms from WordNet 3.0. They have also discussed the issue of cross-lingual FAQ retrieval. Even though the result was not as effective as one would expect in practice, the approach seems promising. One reason for low accuracy was due to the noisy nature of both SMS queries and the text of the FAQ answers.

Jeon et al. [6] introduced the idea of categorizing FAQ questions not only based on question similarity but also using the similarity of the answers. They assume that if two answers are similar enough then the corresponding questions should be semantically similar which warrants the inclusion of the questions in the same category. Their similarity measure is based on word translation probabilities using IBM machine translation model which allows computing semantic similarities between words. This approach allows to exploits the word relationships to retrieve similar questions from FAQ repository for each given question.

3. Ontology-based FAQ Retrieval System

FAQ retrieval system is a special type of QA system where answers are extracted from an existing FAQ knowledge base. Just like in the case of the call center, when a user sends a question to the system, the system will try to match to the equivalent question in the FAQ knowledge base. If there is the same or similar question in the FAQ, the corresponding answer will be sent back to the user. Therefore, the purpose of the ontology will be in assisting matching the incoming question with a question in the FAQ knowledge which requires understanding of the semantics of the incoming question. The keywords from the users question will be matched against classes, attributes and relations in the ontology in order to check if that question exists in the question-answer pair repository. Question similarity refers to the similarity between the keyword set of given question annotated by ontology and the pattern keyword set of FAQ question [7].

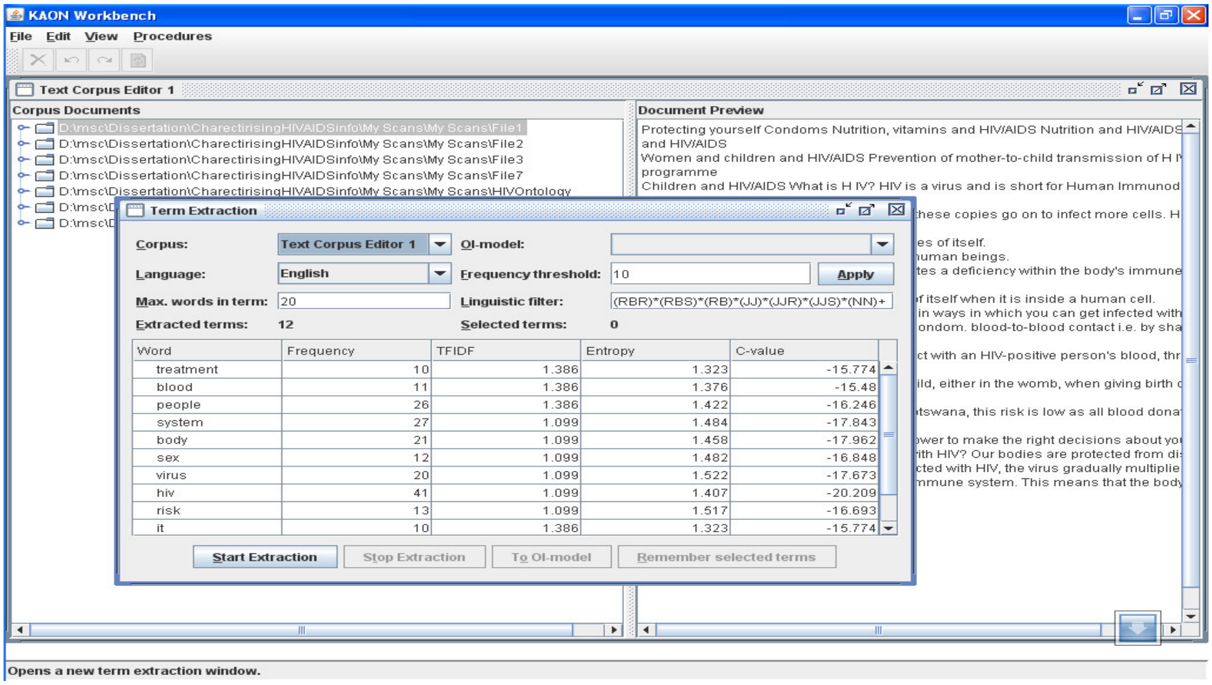
3.1 HIV/AIDS Ontology

As indicated in our previous work [8] ontology construction requires appropriate tools. For this purpose, we carried out an experimental evaluation of ontology construction tools and found that Text2Onto to be the most appropriate tool for our HIV/AIDS ontology construction.

Text2Onto extracts ontology elements automatically and ranks them according to their probability of being candidate terms. It extracts terms using TF/IDF (Term Frequency - Inverted Document Frequency) technique. TF/IDF is a statistical measure used to evaluate how important a word is to a document in a collection or corpus/knowledge source. The TF-IDF value increases proportionally to the number of times a word appears in a document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. The TF-IDF technique uses various mathematical forms to calculate the words that appear frequently and their relevance.

Figure 1 shows the ontology constructed using Text2Onto from the MASA booklet. We can see that treatment, blood, people, system, body, sex, virus, hiv, risk etc. are appearing at the top with higher TF-IDF values. Relations are extracted using association rules and text patterns. Association rules are used to find how terms are related by analysing data using if/then patterns, finding terms that appear frequently in the knowledge source, and generating rules by identifying the number of times the if/then statements have been found to be true.

The main relationship “subtopic-of” is automatically induced when subtopics are added to the ontology. Text2Onto can display concepts and their properties in the same window.



D:\msc\Dissertation\CharectirisingHIVAIDSinfo\My Scans\My Scans\File1
D:\msc\Dissertation\CharectirisingHIVAIDSinfo\My Scans\My Scans\File2
D:\msc\Dissertation\CharectirisingHIVAIDSinfo\My Scans\My Scans\File3
D:\msc\Dissertation\CharectirisingHIVAIDSinfo\My Scans\My Scans\File7
D:\msc\Dissertation\CharectirisingHIVAIDSinfo\My Scans\My Scans\HIVOntology
D:\msc\D
D:\msc\D

Protecting yourself Condoms Nutrition, vitamins and HIV/AIDS Nutrition and HIV/AIDS
and HIV/AIDS
Women and children and HIV/AIDS Prevention of mother-to-child transmission of H I\
programme
Children and HIV/AIDS What is H IV? HIV is a virus and is short for Human Immunod
these copies go on to infect more cells. H
es of itself.
human beings.
tes a deficiency within the body's immune
f itself when it is inside a human cell.
in ways in which you can get infected with
ondom. blood-to-blood contact i.e. by sha
ct with an HIV-positive person's blood, thr
ild, either in the womb, when giving birth o
tswana, this risk is low as all blood dona
ower to make the right decisions about yo
ith HIV? Our bodies are protected from di
cted with HIV, the virus gradually multiplie
mmune system. This means that the body

Corpus: Text Corpus Editor 1
OI-model:
Language: English
Frequency threshold: 10
Max. words in term: 20
Linguistic filter: (RBR)*(RBS)*(RB)*(JJ)*(JJR)*(JJS)*(NN)+
Extracted terms: 12
Selected terms: 0

Word	Frequency	TFIDF	Entropy	C-value
treatment	10	1.386	1.323	-15.774
blood	11	1.386	1.376	-15.48
people	26	1.386	1.422	-16.246
system	27	1.099	1.484	-17.843
body	21	1.099	1.458	-17.962
sex	12	1.099	1.482	-16.848
virus	20	1.099	1.522	-17.673
hiv	41	1.099	1.407	-20.209
risk	13	1.099	1.517	-16.693
it	10	1.386	1.323	-15.774

Opens a new term extraction window.

Figure 1: Text-To-Onto ontology representation

3.2 Semantic Similarity using Set Spreading

The similarity computation involves the use of the generated HIV-AIDS ontology. The inherent relationships between concepts in the ontology are used to enrich the concept vectors of both the query (i.e., user question) and the document being matched as in [1].

Both the query and the documents are represented as vectors using the Bag of Words (BOW) format. The query Q is represented by a set of terms (Qt$_1$, Qt$_2$, …,Qt$_n$} while the documents (questions) from the Question-Answer (QA) pairs is in the FAQ collection are also represented in a similar manner {Dt$_1$, Dt$_2$, …Dt$_n$}. The similarity computation entails assigning a weight $w_{t,\ d}$ to each term to reflect how important the term is in describing a document. The query Q can then be represented as {(Qt$_1$, w_{t1}), (Qt$_2$,w_{t2}), …, (Qt$_n$, w_{tn})} and the questions from the FAQ collection D can be represented as {(Dt$_1$, w_{t1}), (Dt$_2$,w_{t2}), …,(Dt$_n$, w_{tn})} The commonly used tf-idf weighting function [9], is used to determine the weight of each term as shown below.

$$\mathrm{w}_{t,d} = (1+\log_{10}\mathrm{tf}_{t,d})\times\log_{10}(N/\mathrm{df}_t)$$

tf$_{\mathrm{t,\ d}}$ - term frequency of the term in the document,
N - the number of documents and
df$_\mathrm{t}$ - the document frequency of the term

Cosine similarity between the query and the document is computed as follows:

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

q_i is the tf-idf weight of term i in the query and d_i is the tf-idf weight of term i in the document [9].

Thiagarajan et al. [1] proposed a scheme whereby terms are represented by either an *extended set* or a *semantic network* after spreading. The two schemes are similar in concept but differ in that one returns an extended set while the other returns a graph. The graph method has the advantage that different similarity measures can be applied based on the edges and paths in the graph.

In this paper, we apply set-based spreading. For each term representing a user question, conceptually related terms are derived from the ontology resulting in an extended set of terms for a particular question. The same can be done for the questions from the FAQ collection. Therefore, we can have two extended sets of terms extended with terms derived from the ontology that are related to the original terms. After this, we can apply the cosine similarity to determine the semantic similarity between a user question and questions from the FAQ collection.

For example, consider the two questions "*what is HIV?*" and "*what is AIDS?*". The set of terms representing "what is HIV?" will be <what, HIV>. Similarly, the set of terms representing "What is AIDS?" will be <what, AIDS>. Using the cosine similarity, the similarity between these two questions will be 0.43. This indicates a very low similarity between the two (based on terms that do not really contribute to the meaning of the main keywords *HIV* and *AIDS* in the two sets. Q will be represented as Q = {what 0.48, HIV 0.96} and D = {what 0.48, AIDS 0.96}.

Extending the term sets using the ontology constructed in our previous work as shown in Figure 2, results in a better semantic similarity. From Figure 2, we can see that different relationships such as *AIDS is-caused-by HIV virus* are established between the concepts. Using this relationship, the extended sets become Q = *{what, HIV, virus, AIDS}* and D = *{what, AIDS, HIV, Virus}*. The resulting similarity becomes 1 in this case since the two sets are now exactly the same.

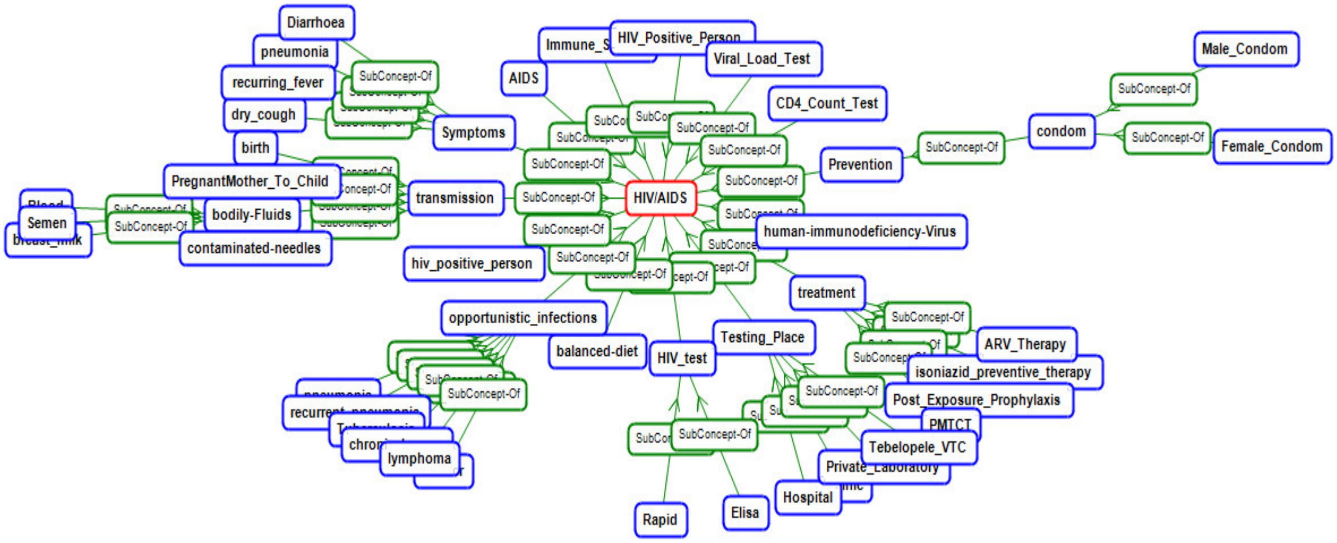


Figure 2: Part of the HIV-AIDS Ontology

Let's consider another example without including the weights for the purpose of clarity.

The questions Q_1 "*How is HIV transmitted*?" and Q_2 "*Which bodily fluids could contain the virus*?" will result in term sets Q_1 = *{How, HIV, transmitted}* and Q_2 = *{Which, bodily, fluids, could, contain, virus}*. The resulting Cosine similarity between Q1 and Q2 is 0.0. However, after extending the two sets with the sub concepts from the ontology highlighted in Figure 2, Q1 is represented as *{How, HIV, transmitted, birth, pregnant, mother, child, bodily, fluids, contaminated, needles, blood, semen, breast, milk}* and Q2 is represented as *{Which, bodily, fluids, could, contain, virus, blood, semen, breast, milk}*.

The cosine similarity of the two extended sets becomes 0.18.

Set-based spreading involves the use of ontology in RDF format. To be able to work with the ontology, there is a need to parse the RDF and navigate the relationships stored. The Apache Jena RDF API [10] has been implemented on top of the cosine similarity function to derive the relationships.

4. Discussion

The weight of the terms added through set spreading is calculated in the same way as the original terms. However, it will be necessary to control the type of relationships and the depth of relationships to avoid dilution of semantic similarity by least related concepts.

Another issue is the controlling of stop words. Different tools treat stop words differently which may have an effect on the resulting term set. This needs to be controlled to prevent similarity recognized based on common words in questions. The spreading function is an iterative function iterating through the ontology looking for related terms until the relatedness is exhausted. Trying to exhaust all the relationships may not be practical in some applications and we need to determine the level of iteration through experiments.

In this paper, we demonstrated the effectiveness of the set-based spreading process for the computation of semantic similarity between user question and questions in FAQ collection. However, it is important to investigate the improvement in effectiveness using a large collection of FAQs instead of few examples. In our future work, we planned to run the system on a large collection of FAQs.

In this paper, we focused on computing the similarity between a user question and questions in HIV/AIDS FAQ collection. However, it will be interesting to investigate to what extent semantic similarity can improve if we compare a user questions against question-answer pairs from FAQ collection instead of considering only questions from FAQ.

Currently, semantic similarity computation is based on Question to Question mapping. In the future, this will need to be extended to cater for Question to Question-Answer mapping. The results will inform the best method of finding an answer to a query from the FAQ knowledge-base.

5. Conclusion

In this paper, we presented an ontology based approach for FAQ retrieval using set spreading process for the computation of semantic similarity between a user question and questions from an FAQ collection. Through examples, we demonstrated the effectiveness of the set spreading process based on domain ontology (i.e., HIV/AIDS Ontology).

In the future, we will explore the effectiveness of graph-based spreading process for the computation of semantic similarity. In addition, we will investigate if mapping a user question to question-answer pairs in FAQ collection will improve the matching of user question and questions from FAQ.

References

[1] R. Thiagarajan, G. Manjunath, and M. Stumptner, Computing Semantic Similarity Using Ontologies,University of South Australia 2008.

[2] G. Kothari, S. Negi, T. A. Faruquie, V. T. Chakaravarthy, and L. V. Subramaniam, SMS based Interface for FAQ Retrieval, in *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore, 2009, pp. 852–860.

[3] W. Song, M. Feng, N. Gu, and L. Wenyin, Question Similarity Calculation for FAQ Answering, in *Third International Conference on Semantics, Knowledge and Grid*, 2007, pp. 298-301.

[4] F. Wang, G. Teng, L. Ren, and J. Ma, Research on Mechanism of Agricultural FAQ Retrieval Based on Ontology, in *9th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, 2008, pp. 955-958.

[5] I. Batyrshin, M. l. Mendoza, P. Pakray, S. Pal, S. Poria, S. Bandyopadhyay, and A. Gelbukh, SMSFR: SMS-Based FAQ Retrieval System, in *Advances in Computational Intelligence*. vol. 7630: Springer Berlin Heidelberg, pp. 36-45.

[6] J. Jeon, W. B. Croft, and J. H. Lee, Finding Similar Questions in Large Question and Answer Archives, in *International Conference on Information and Knowledge Management*, Bremen, Germany, 2005, pp. 1-7.

[7] Z. Remi, Towards Ontological Question Answering, in *ACL-2001 Workshop on Open-Domain Question Answering*, 2001.

[8] B. Moeng, Y. Ayalew, and G. Mosweunyane, Experimental Evaluation of HIV/AIDS Ontology Construction Tools, in *IASTED Health Informatics*, Gaborone, Botswana, 2012, pp. 339-346.

[9] C. D. Manning, P. Raghavan, H. Schuetze, *Introduction to Information Retrieval*: Cambridge University Press, 2008.

[10] Apache-Software-Foundation, A free and open source Java framework for building Semantic Web and Linked Data applications, 2014.