Question Matching Technique to Find Answers

P.P.G Dinesh Asanka[1], A. S. Karunananda[2]
[1,2] Department of Computational Mathematics
University of Moratuwa, Moratuwa
Sri Lanka
dineshasanka@gmail.com[1]; asokakaru@yahoo.com[2]

Abstract

In today's business world, there are lots of knowledge systems and users need to find answers from existing knowledge bases. Due to the complexity of the systems and fat contents of them, it is difficult for users to find appropriate answers to their questions. Knowledge bases can be configured to mapped answers to questions. When end user enters a question, using natural language processing and text mining techniques, the question is matched with the existing question in the question bank in the knowledge base and matched answer and any other related contents (if exists) are provided to the end user. Proposed technique was evaluated with Sri Lankan 1978 constitution knowledge base and it was found that many questions can be matched with higher accuracy.

Index Terms **— Natural Language Processing, NLP, Text Mining, Term Frequency, TF-IDF, Cosine Distance, SoundEX.**

1. Introduction

There exists large number of knowledge bases in different domains and they have gathered lot of information over the years. Users frequently need to query those contents. Since most of these knowledge bases are unstructured and complex, retrieving relevant information is complex and will be a tedious process.

As there are no structural techniques to retrieve contents from large and complex knowledge bases, most users implement simple text retrieval techniques. This will return lot of documents and they might not be even matching the users' intended requests. So users might need to read almost all the documents and users have to verify whether the content is what s/he was looking for. Some users may implement technique like Microsoft Full Text Search [MSDN, 2012] which provides some features to retrieve documents with higher accuracy than a simple text search.

Every knowledge base has frequently asked questions a.k.a. FAQs. Therefore, if system can matched the user entered question to this FAQs, it can provide the answers to end users where in the knowledge base each question is linked to one or more answers. Sometimes FAQ can be linked to different data sets depending on the domain. In case of Sri Lankan legal domain, questions can be linked to legal cases, incidents and constitution contents. So when the question is matched, relevant cases, incidents and constitution content will be retrieved addition to the answer to the question. In this type of complexity, it will be difficult to use simple text matching technique.

Another issue with the simple text matching technique is that, there can be similar words. For example, user might be searching for a word called "amend", but there can be questions which has words like "amended", "amendment" or "amending" which should also be matched. However, with the simple text matching technique, those documents will not be matched and those documents will be ignored from the search.

Also, there are similar words which are relevant to each domain. For example, in Sri Lankan legal domain, 1978 constitution, President is referred as head of the state and head of the cabinet. So when users are raising questions, they may state about President but the question contains head of state which needs to be matched.

Also, in the knowledge bases, there are lot of key words are used. Users may raise a question stating PM which is an abbreviation for Prime Minister. Questions might be set up with Prime Minister and those questions needs to be mapped when user enters question with one or more abbreviations.

2. Current Work

In the research paper about Ontology-bases Question Answering System, architecture was proposed [Lee S, Ryu P., Choi K., 2006] to answer questions and it shows the path for ontology lookup for four types (concept

completion, example, enablement, and goal orientation). For each of the type, they have observed diverse search paths. For example, after searching a class for questioned entity, three kinds of paths are possible. The path selection is determined by query type. In determining search path, we should also consider the type of questioned entity (class/instance). It can be seen that the path for concept completion type is divided with two paths according to the results of class/instance discrimination. For the proposed question answer mechanism, questions should be arranged in a stucture. However, not all knowledge bases can be arranged in the given structure. Therefore, this method is more suitable for simple knwoledge bases not suited for legal domain queries.

In another short research paper named, Rich Lexical Knowledge based Q&A System for Ubiquitous Knowledge Service, it describes a simple domain specific (rice) architecture for ontological question answering system [Kawtrakul A. & Thunkijjanukij A, 2009].

Proposed question answering system is based on three sources of knowledge which interact:

- Lexical data and in particular lexical semantics and lexical inference.
- The domain data as represented by the rich conceptual functions, i.e. Rice Ontology.
- Some general purpose knowledge, useful for answering questions.

This is the closest architecture for the proposed research in this paper. However, evaluation results are not available with this research paper.

In another research paper, question and answering system was designed for Dinning Ontology [Palaniappan L., Sambasiva Rao N., 2010]. Proposed architecture of dining ontology as domain. User query will pass through annotator of named entities, then queries will be checked whether it is synchronized, after test by reasoner, the answer will be searched in the text pool and it is retrieved. Instead of going to database query, template process queries faster. Thus reusability is enhanced. In this research, another important mechanism is to match the questions in the Ontology with the question that users are entering.

Research paper by Jing Yu and et. al. titled Similarity Measure of Test Question Based on Ontology and Vector Space Model (VSM) is to identify same question from question bank [Yu J., Li D & et. al, 2014]. VSM is a common method for measuring text questions similarity in massive item bank system proposed by Salton in 1970s [Galton S. & Buckley G., 1988]. It is relatively legacy algorithm which was used in measuring text similarity. Though this algorithm is easy to be applied, it has ignored the relations among words in the documents and only uses words frequency. In this research paper, better solution was proposed.

Also, in another research [Asanka PPGD, 2013], cosine distance was used to identify closely matched documents using Term Frequence – Inverse Document Frequency (TF-IDF) technique. In this research paper, matrix was provided to find closely matching document using cosine distance. In the matrix, red cells indicate closely matching documents while white cells are the documents which are loosely matching documents. In this research, simple terms are used to match documents, hence the error rate is high in the proposed technique. However, the question matching technique is some what similar to doucument mapping technique which is helpful for this research.

3. Design

Question matching is done in two steps. First, all questions in the ontology needs to be analyzed and second step is to match the user question with the existing questions.

In the first step, there are three sub steps. Initial step will capture global terms for all questions with the frequency. String Representation and Soundex is used to capture global terms. In Natural Language Processing (NLP) , there are different text representation and depending on the text representation analysis is different [Zhai C., 2015]. String text representation is the most general representation and it is robust. Hence for this research, String representation is used.

There can be similar terms such as amend, amendable, amending, amendment, amendments with different formations. SOUNDEX algorithm is identified to find the similarities of similar words.

SOUNDEX is a phonetic algorithm for indexing names by sound as pronounce in English. The goal is for homophones to be encoded to the same representation, so that they can be matched despite minor differences in spelling [Poole D, 2012]. The algorithm mainly encodes consonants and a vowel will not be encoded unless it is the first letter. SOUNDEX is the most widely known of all phonetic algorithms [Poole D, 2014].

SOUNDEX function is used in Microsoft SQL Server 2014 [MSDN, SOUNDEX, 2012] to identify the SOUNDEX value of each term.

In the Global Terms capturing step, all the terms are captured and SOUNDEX value or each term is stored. While capturing Global Terms, frequent terms like, ‘a’,

'and', 'the', 'on', 'an', are ignored to improve the performance.
Next step is to find out, terms for each question which is called term lookup. There are several methods to calculate term frequencies such us simple count, log vector, 0/1 bit vector and BM25. For this research, word count method is selected as term frequency method.
Third step is to calculate TF-IDF for each question. Calculating Cosine Distance for documents using Term frequencies may not be accurate as there can be common terms across the documents. Therefore, Inverse document frequency is introduced. TF-IDF can be used to calculate the distance between the documents.
These three steps are shown in the Figure 1, where SQL Server Intergration Services (SSIS) was used.

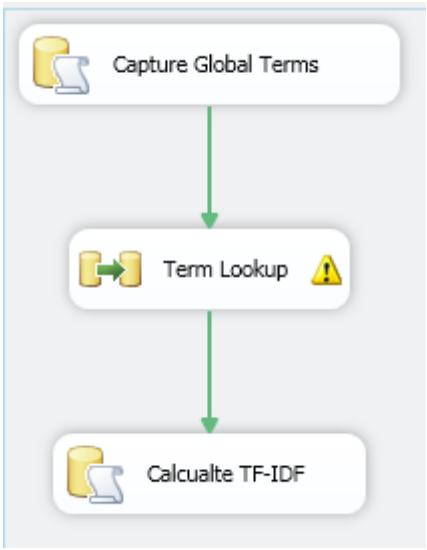

Figure 1. Calculate TF-IDF for Existing Questions in the Question Bank

For all the questions, TF-IDF is calculated before and saved in the database to improve performance of the searching mechanism.
Next part is, matching the user entered question with the question bank. When a user enters a question, first it needs to be checked for spelling. Though there is no special spell checker included, with the use of SOUNDEX error with wrong spelling is minimised.
There are several ways to calculate distance between two objects. There are Euclidean distance, Cosine distance, City-block (Manhattan) distance, Chebychev distance and Mahalanobis distance [Hair J.F. et. al., 2012]. Usually, Cosine Distance method is used to calculate the distance between two documents [Hand D. et al, 2007]. Apart from above distance measures, there are other measures like Chi-square, Lift, AllConf, Jaccard, Cosine, Kulczynski, MaxConf etc [Han J., 2015].
Question type also needs to be considered. There can be same terms used in multiple questions. Only the difference between those questions are objective of the question. That is which, what, how many etc. Every question in the question bank will have a question type. So when user enters the question, apart from the term matching, there need to have a question type matching as well.
Since there are two rankings, one ranking from terms and another ranking from question type. Question type ranking and term ranking is multiply to obtain the final ranking.

Final Ranking = Term Ranking * Question Type Ranking^ n

n - is decided by after carrying out few experiments and best n value is 0.5.

Typically, abbreviations and similar words are used as said before. For example, most people uses MP for Members of Parliament and PM for Prime Minister. Also, users use similar words. For example, President is Head of State and Head of the Cabinet. So when users enter President or Head of State both times it is referring to the same person or object. So Ontology is used to store those information which is described in implementation section later.

4. Implementation

Microsoft SQL Server 2014 was used to store data and Microsfot SQL Server Intergration Services (SSIS) was used to extract data. Proper indexes are used to enhace performance.
Frequently used words, abbrivations and similar terms are saved in the database. Since those are stored in the database, it can be easily modifed or new content can be added to those tables when required which has improved the usability.

5. Evaluation

Question matching can be divided into three areas such as normal question matching, questions with abbreviations and questions with similar words.
Evaluation was done by entering legal domain questions which are relevant to Sri Lanka. 1978 Sri Lankan constitution was used for the purpose of evaluation. Following scenarios were identified to evaluate the proposed system. Retrived questions are manually verified

to evaluate whether it is relevant. Most relevent questions are listed at the top in all cases.

Scenario 1: How to amend the constitution

When the above question is entered, following are the matching questions returned from the proposed system.

Questions Returned for Scenario 1

Question	**Is it Relevant?**
Who approves amendments to the constitution?	Yes
Which provisions are not amendable?	Yes
What are the details of the amendment proposal process?	Yes
How is the head of government selected?	No
What are the details for the amendment approval process?	Yes
How is the head of state selected?	No
How many executives are specified in the constitution?	No
What proportion of the vote is needed to approve a constitutional amendment?	Yes
How are members of the second chamber selected?	No
Does the constitution provide for at least one procedure for amending the constitution?	Yes

Out of the ten questions returned, seven questions are relevant to the question in senario 1 and no questions in the question bank are missed in the list. This means that, there is 70% of accuracy whereas the ranking is concerned, 67% (37 / 55) is the ranking relevancy. 37 is the total ranking for the correct questions while 55 is the total ranking for all the questions returned. Total ranking is calculated by considering the return order of the returned questions. For example, for the if the first question is relevent 10 ranking points were given.

Scenario 2: How to remove Members of Parliament?

When the above question is raised, only one question is retrieved which is "Are there provisions for dismissing members of parliament?". So there is 100% matching both with the number of questions and the ranking is concerned.

Scenario 3: How to remove MPs?

Scenario 3 is the same question as Scenario 2 with introducing on MPs instead of members of parliament. However, same question is returned. So it also has the 100% accuracy.

Scenario 4: What are the provisions to remove head of state?

Following are the questions returned for the above question.

Questions Returned for Scenario 4

Question	**Is it Relevant**
Are there provisions for dismissing the head of state?	Yes
Who can propose a dismissal of the head of state?	Yes
What are the details of the process to remove judges?	No
Who are the electors for the head of state?	No
Are there provisions for dismissing the head of government?	Yes
Who may remove the chief of the central bank?	No

Out of the six questions returned, three questions are relevant to the question and no questions in the question bank are not missing in the list. This means there is 50% of accuracy whereas the ranking is concerned and it has 62% (13 / 21) is the relevency ranking.

Scenario 5: What are the provisions to remove President?

Following are the questions returned for the above question. This scenario has similar words as of scenario 4.

Questions Returned for Scenario 5

Question	Is it Relevant
Are there provisions for dismissing the head of state?	Yes
Who can propose a dismissal of the head of state?	Yes
What are the details of the process to remove judges?	No
Who are the electors for the head of state?	No
Are there provisions for dismissing the head of government?	Yes

Out of the five questions returned, three questions are relevant to the question and no questions in the question bank was missed in the list. This means there is 60% of accuracy whereas the ranking is concerned it is 67 % (10 / 15) ranking relevancy.
When all the scenarios are considered ranking relevancy is at least 60% and in some scenarios it is 100 %.

6. Conclusion and Further Work

In the current system, user question needs to be matched to the question to get the answer and other related contents. This research can be further improved by providing an option to automatically find answer even if there is no matching question.
There are additional features introduced with SOUNDEX function like spell checking and also system has the capability of matching with abbreviations and equal terms. However, system does not have the ability of matching questions semantically. If that can be adopted, the system, system will be further enhanced. For example, when user enters query which includes, "person" system should be able to understand that citizen and person are equal in semantic. Currently only the words in local ontology will be matched, but system should be extended to get semantic words from defined third party Ontologies.
In this research, it was able to find a mechanism to map questions in the question bank to the user entered questions. For the selected sample scenarios, ranking relevancy is at least than 60% and in some scenarios it is 100 %.

References

[1] Asanka PPGD., "Finding Similar Text Files using Text Mining", 8th IEEE ICCSE 2013 Colombo.

[2] Galton S., Buckley G., "Term-weighting approaches in automation text retrieval", Information Processing and Management, vol. 24, no. 5, pp. 513-523, 1988.

[3] Hair J.F., Black W. C., Babin B. J., Anderson R. E. and Tatham R. L., in Multivariate Data Analysis, New Delhi, Pearson Education in South Asia, 2012, p. 599.

[4] Han J., Pattern Discovery in Data Mining, University of Illinois at Urbana-Champaign, Course Period 2015-Feb-09 – 2015-Mar-08.

[5] Hand D., Mannila H. and Smyth P., "Retrival by Content," in Principles of Data Mining, New Delhi, Prentice Hall of India Private Limited, 2007, p. 456–464.

[6] Kawtrakul A., Thunkijjanukij A., Khantonthong N., "Rich Lexical Knowledge based Q&A System for Ubiquitous Knowledge Service"., Department of Computer Engineering, Kasetsart University, Bankok, 2009,

[7] Lee S, Ryu P., Choi K., "Ontology-based Question Answering System", Korea Advanced Institute of Science and Technology., 2006.

[8] Microsoft Development Network, Query with Full-Text Search, MSDN, *https://msdn.microsoft.com/en-us/library/ms142583.aspx*, Accessed on 2014-12-04.

[9] Microsoft Development Network, *SOUNDEX (Transact-SQL), MSDN,* https://msdn.microsoft.com/en-us/library/ms187384.aspx?f=255&MSPPError=-2147217396*, Accessed on 2015-02-20.*

[10] Palaniappan L., Sambasiva Rao N., "An Ontology-based Question Answering Method with the use of Query Template", International Journal of Computer Applications (0975 – 8887), Volume 9– No.9, November 2010.

[11] Poole D., "Soundex - Experiments with SQL CLR", SQLServerCentral.com, http://www.sqlservercentral.com/articles/Programming/101836/, 2013/09/12, Accessed on 2015/06/12.

[12] Poole D., "Soundex - Experiments with SQLCLR Part 2", SQLServerCentral.com,

http://www.sqlservercentral.com/articles/soundex/120628/,, 2014/12/30, Accessed on 2015/06/12.

[13] Russel M.A., Whiz-Bang A., "Introduction to TF-IDF" in Mining the Social Web, New Delhi, O'Reilly Media, Inc., 2011, p. 201.

[14] Yu J., Li D., Hou J., Liu Y., Zhaoying Yang, "Similarity Measure of Test Question Based on Ontology and VSM", The Open Automation and Control Systems Journal 2014, 6, pp 262-267.

[15] Zhai C., Week 1: NLP, Text Representation, and Word Association Mining, Text Mining and Analytics, Department of Computer Science, University of Illinois at Urbana-Champaign., www.coursera.org., 2015.