IJCTA, 9(27), 2016, pp. 543-550 ISSN: 0974-5572
© International Science Press

A Refined Latent Semantic Analysis (LSA) Technique to Improve Accuracy Of Agriculture Data In Mobile Cloud Computing (MCC)

P. Calduwel Newton[1] and T. Cynthia[2]

ABSTRACT

The ultimate aim of this proposed idea is to develop a more accurate Question Answering mobile cloud application for agriculture domain using refined Latent Semantic Analysis (LSA) technique. Question Answering is an emerging research trend that helps the user to retrieve exact answers for the questions instead of relevant documents. Technically, Question Answering technique is a search program that retrieves answers from the database it has been provided. The proposed idea aims to help the farmers by answering all their queries regarding tomato plant diseases, symptoms, prevention and controlling the diseases. By querying the system through text or voice, the farmers get their answers accurately. The proposed technique applies Latent Semantic Analysis (LSA) technology that is used by Google and other search engines. The LSA technology retrieves the semantically related documents by eliminating the impact caused by synonymity and ambiguity of the words. A tool is developed to implement the basic features of the proposed technique. The dataset about the plant diseases and remedies are collected and processed manually from different web pages for testing to improve the accuracy of the proposed idea. The accuracy of the proposed technique is proven by subjective evaluation. Users query the developed tool by text or voice then rated based on accuracy of the answers they have retrieved. The proposed idea managed to achieve desirable percentage of accuracy.

Keywords: Agriculture tool, Tomato disease, Voice input, Latent Semantic Analysis, Question Answering System.

1. INTRODUCTION

Search engines give relevant documents to the user's query. Unlike search engines, QASs gives a short answer to the user with evidence. In wikipedia, question answering technique is defined as a discipline[1] within the fields of Information Retrieval(IR) and Natural Language Processing(NLP) that builds an automatic answering system asked by humans in natural language. International Business Machine's (IBM) Watson, START, Apple's SIRI and Ask.com are very popular question answering services. Domain of application is one of the dimensions in the growth of question answering technique. That means, QAS for specific domains like medicine, sports and agriculture improves the accuracy of QAS.

The proposed idea behind the tool is an accurate QAS for agriculture domain that can be easily accessed by their farmers with their smart phones. Farming sector in India is highly un-organized. Farmers are in need to adapt new technologies to meet the global food requirements. They need a lot of support, guidance and technical help from experts to cultivate seeds, to test soil, to select pesticides and to get weather reports. A study from Coaliation for Successful Workforce (CSAW)[2] said that there is a shortage of agriculture scientists. To compensate the shortage, many expert systems in agriculture, many mobile applications, many recommendation systems and QA techniques lend their helping hands to farmers to connect and communicate with experts across the globe.

1 Assistant Professor, Department of Computer Science, Government Arts College, Ayyarmalai, Kulithalai – 639 120, Karur-Dt, TamilNadu, India, *Email: calduwel@yahoo.com*

2 Assistant Professor, Department of Computer Applications, Bishop Heber College (Autonomous), Tiruchirappalli –620 017, TamilNadu, India, *Email: cynthiasamabi@gmail.com*

The tool uses Latent Semantic Indexing(LSI) technology or Latent Semantic Analysis(LSA) .LSA is chosen because it matches the question with the relevant documents that are close to answers in a more semantic manner. LSA prepares a *term-document-frequency* matrix and finds the closeness of documents relevant to the query by using *cosine similarity*. By considering semantics, the ambiguities of the terms are eliminated.

The proposed question answering technique for agriculture uses a data set that consists all the relevant details about tomato plant diseases answering along with the year of publishing in World Wide Web (WWW). This idea deals with tomato plants and it's diseases. It helps the farmers to query by typing or querying by speech to obtain the accurate answers. It is tested for accuracy by 100 users whom are varied by their age, occupation and gender. Many farmers, non-farmers, students and professors tried the proposed tool and checked the accuracy of the system.

2. RELATED WORKS

The proposed work is an inspiration of a recommendation system OAPRS that is designed by Qingtian Zang [3] et al., In their system, they designed a QAS for agriculture that prescribes the solution for farmer's query and retrieved some experts to communicate in online. In OAPRS, they used term-document-frequency matrix and cosine similarity to find the most relevant documents for the query. David Tobinski et.al[4] used LSA technology for automatic question scoring for open questions. Joao Carlos Alves dos Santos et al.,[5] used LSA technology in their work for automatic evaluation of answers. Their system is implemented in an online university to evaluate the answers automatically. The result is 84.94% accurate that is over performed human evaluation. Priyanka Singh et al.,[6] used the LSI technology to answer the unanswered questions in the public forum like Stack Overflow and Reddit.

When talking about QAS, MEANS[7] is QAS that combined Natural Language Processing (NLP) techniques to answer the medical problem related questions. It used ontology, a recent technology in semantic web to answer the questions. A survey[8] stated, QASs are classified based on the application domains, types of questions, types of data sources and the various techniques that are used to retrieve the answers. Based on application domain QASs, There is a QAS for Arabic language[9] where the questions are entered in Arabic.

Like application domains and linguistic dimensions, QASs are very popular applications in mobile devices. Apple's siri is a very successful QAS in Apple's iphone. Siri[10] is a revolution in gadget wise dimension of QASs and it is also a voice based QAS that used speech recognition and text-to speech technique. The users talk to siri to schedule their meetings, to place a call and so on. IBM's Watson, Google's Google Now, Microsoft's Cortana are also very popular QASs. Surprisingly, an image based QAS is developed by Johann Haswald et al., [11] which takes images as input and retrieves answering.

This paper proposes an idea for QAS in agriculture domain. In agriculture, to improve productivity and to prevent the plants from diseases, a QAS is developed by Takahiro Kawamura et al., [12]. There are many mobile applications available to get the up to date information about agriculture market details[13][14]. AGRI-QAS is an another answering system in agriculture domain that resolves the problems of farmers and deals factoid type questions. By referring the previous works in QASs, the proposed technique is developed in a unique manner to answer the queries of the tomato plant's diseases.

3. A REFINED LSA TECHNIQUE FOR AGRICULTURE DATA IN MCC

The proposed idea for Question Answering in agriculture domain is designed and developed for answering all the queries about tomato plant diseases. This technique concentrates on a single domain to improvise the accuracy. For answering the queries, a common Information Retrieval (IR) technique LSA with refinements is applied. The implementation of the idea behind the proposed technique is given in Figure 1. The refined LSA is a technique that finds relevant documents for the search term or query given by the user. When the query *Q* is given by the user, it

is passed to the Google search engine as search term. As a result, the proposed tool retrieves the first web page that is retrieved by the Google search engine. Document linearization is done in the retrieved web page's content by removing mark up tags and by removing special formatting, a plain text is obtained.

The plain text is tokenized as terms t_i by removing all the punctuations and stored. Here i={1,2..n}. The same text is tokenized as sentences and saved as *sent_token_list*. Each statement in *sent_token_list* is considered as documents d_i, where $i=\{1,2..n\}$. The terms t_i are pre-processed to minimize the number of terms, by removing stop words and by doing stemming. After that, LSA technique is implemented in terms t_i by finding the *term-document-frequency matrix (tdf)*. To find *tdf* matrix, *inverse document frequency (idf)* values for terms t_i are calculated by using the formula

$$idf=1+\log_2 \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ ti}$$

and *term frequency(tf)* for all the documents d_i in *sent_token_list* is calculated. By multiplying *tf* with *idf*, the tool obtains *tdf* matrix. The pre-processing and obtaining *tdf* is done for query *Q*. To find the similarity between the documents d_i and the query *Q*, *cosine similarity*[15] is calculated.

The higher similarity value shows the higher order of relevance to the query *Q*. The first, second and third maximum of similarity values $Max_k(CosSim(di,Q))$ where $k=\{1,2,3\}$ are taken. These three documents are considered as M_k where $k=\{1,2,3\}$ and compared with the text documents T that has answers a_x along with year of publishing Y in www. The text document *T* is manually gathered information from www. Each M_k is compared with each a_x in *T*. The five $Max_i(CosSim(M_k, a_x))$ is taken for further calculation to improve accuracy of the answer where i={1,2..5}. For each of the five maximum values, a_x that is compared with M_k is published in the current year, a weight $w_1=3$ is multiplied else if a_x is published in the previous year, $w_2=2$ is multiplied. The $Max(Maxi(CosSim(M_k, a_x)))$ is retrieved as answer *A*. To answer the voice based query, speech recognition technique captures the query and the steps that are mentioned above are followed to answer the voice based query.

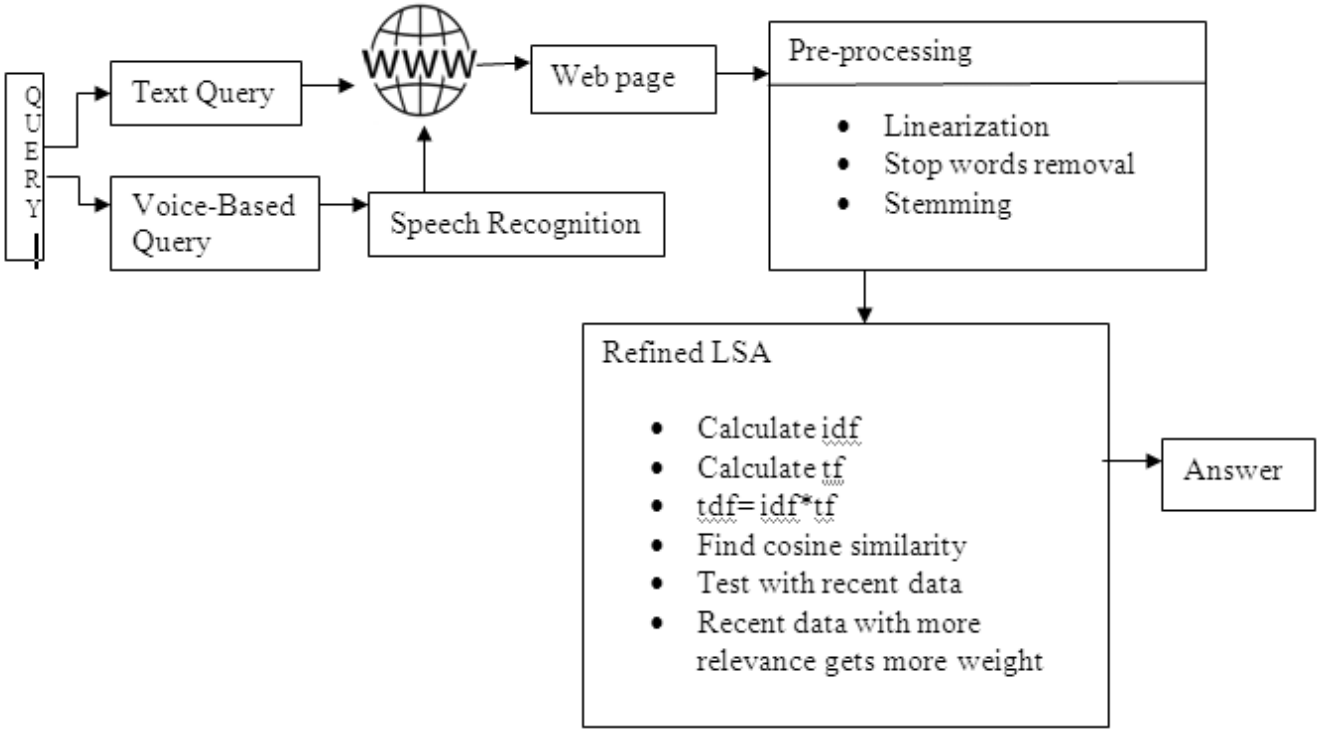

Figure 1. Refined LSA Technique for Agriculture in MCC

Various steps involves in refined LSA to enhance the accuracy of Question Answering is given below.

Input: Tomato_diseases related query to the World Wide Web

Output: Most relevant answer to the query

Steps:

1. When the query *Q* is given as input, it passes to the Google search engine as search term
2. For query *Q*, the contents of the first page retrieved by Google is linearized, tokenized as terms and unique terms are stored as terms t_i where i={1,2..3}
3. For query *Q*, the content of the web page is tokenized as sentences and stored as *sent_token_list*
4. Each sentence in *sent_token_list* is considered as document d_i
5. For each document d_i in *sent_token_list* is pre-processed by removing stop words and stemming to reduce the number of terms
6. For each term t_i, inverse document frequency idf is calculated using the formula

$$idf=1+\log_2 \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ ti}$$

7. For each term t_i in document d_i, term frequency *tf* is calculated by counting the number occurrences of t_i in d_i
8. For query *Q* and document d_i, term document frequency matrix *tdf* is calculated by multiplying *tf* and *idf*. *(tdf=idf*tf)*
9. To find similarity between document d_i and query *Q*, cosine similarity is calculated using the formula

$$CosSim(di,Q) = \frac{di.Q}{\|di\| * \|Q\|} \quad where\ 1<=\ CosSim(di,Q)\ <=0$$

Here $di.Q = di[0]*Q[0] + di[1]*Q[1] + ... + di[n]*Q[n]$,

$$||di|| = \sqrt{di[0]^2 + di[1]^2 + ... + di[n]^2} \quad and \quad ||Q|| = \sqrt{Q[0]^2 + Q[1]^2 + ... + Q[n]^2}$$

10. The three documents that has top three *Max(CosSim(di,Q))* values are considered as M_k where *k={1,2,3}*.
11. With text document *T*, that has collection of answers a_x and year of the answer's publication *Y*, each M_k is compared to find similarity i.e.*CosSim(Mk, a_x)* where *x={1.2..n}*
12. For the top five $Max_i(CosSim(M_k, a_x))$, year *Y* for a_x is checked to multiply *weight* $w_1=3$ for current year and $w_2=2$ for previous year with the similarity values.
13. The line that has the maximum value is displayed as answer *A* for *Q*.
14. For voice-based input, speech is recognized and the above steps are followed.

4. SIMULATION RESULTS AND FINDINGS

The proposed tool is designed with two buttons for getting text query and voice query. Once the button for text query is clicked, a text box that prompts the user for query is shown. After typing the query, by clicking submit button the answer is shown in the text box that is given below the submit button. The screenshots are shown in Figure 2 and Figure 3.

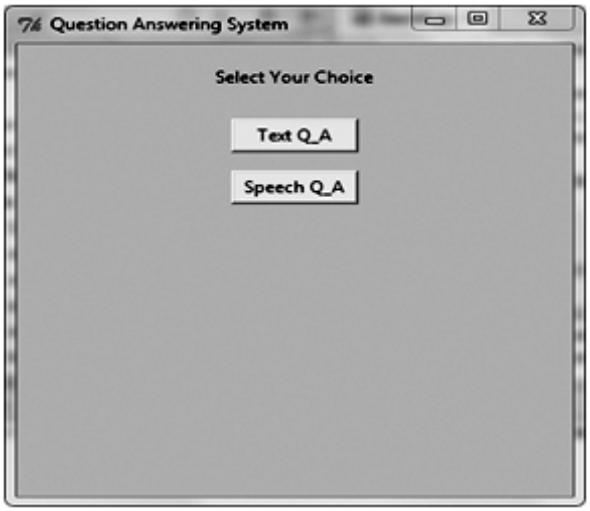

Figure 2. Screenshot for QA in MCC

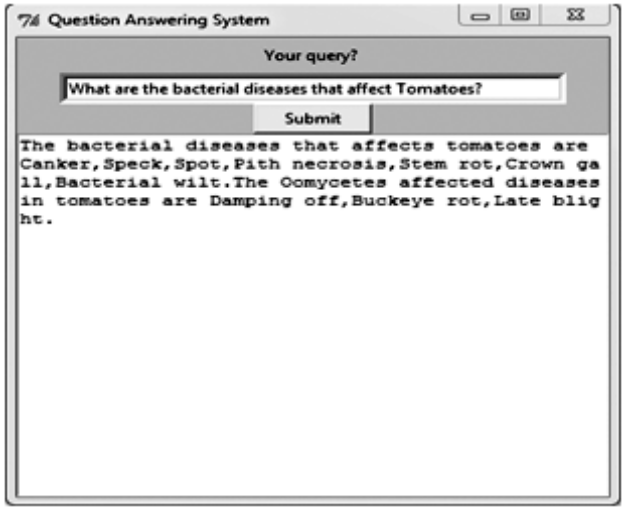

Figure 3. Screenshot for Text Query

The proposed tool is developed in Python and uses PyCharm 4.5.3 IDE. This tool implements the basic idea in QA, which is a part of Natural Language Processing(NLP) . The *nltk* package (natural language tool kit) for information extraction is imported. To design the Graphical User Interface(GUI), *Tkinter* package is imported. To get the top relevant web page from Google for the query, Google API is used along with a *urllib* package. *Beautifulsoup* package in python helps to retrieve the contents of the first web page from Google for the given query and helps for document linearization. Refined LSA technique needs the queries and documents to be pre-processed. To tokenize, tokenization package from *nltk*, to remove stop words, *stop words* package from corpus that is a part of *nltk*, for stemming, *PorterStemmer* package and *numpy* package for performing scientific calculation are imported.

Once the query is given by the user, the query is tokenized into terms, all the stop words are removed from the query and stemming is done with the query. Likewise, the linearization and lemmatization is done in the first web page that is retrieved by Google search engine for the query term. The *tfm* is built for both the query and the web page. By calculating *cosine similarity*, the top three highest ranking lines in the web page is considered as most relevant answer and compared with a text document that has multiple answers for the same query. These answers are stored along with their year of publishing in www. The three documents are compared with the text document and the similarity with more recent answer gets more weight. Because, the recent year data could have many updated information which enhances the accuracy of the answer. The high value shows more relevant and accurate and that is retrieved as answer. This tool is designed to answer voice based query input also. The sample screen shot for voice base query input is given in Figure 4 and Figure 5.

Figure 4. Voice-based query

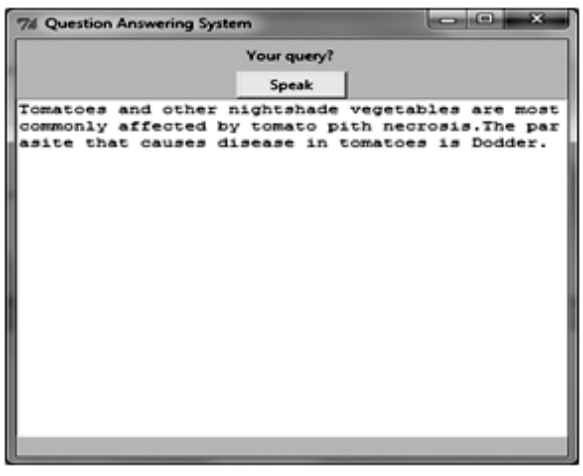

Figure 5. Answer for voice query

When the button for voice based query is clicked, the user is prompted to ask a question. A *speech recognition* package is included in the python program that recognizes voice based question and follows the discussed LSA technique to give the answer.

The refined LSA technique that is used in this implementation is a very commonly used technique in IR. The previous works used this technique to retrieve the most relevant document from the database. In this proposed work, refined LSA is programmed to retrieve the most relevant document from the web. So, the accuracy is guaranteed. This proposed technique cannot be easily measured, so subjective evaluation is conducted. Accuracy, tool design and user satisfaction of the proposed idea is measured. Accuracy is a specification that conforms to the correct value or a standard, tool design evaluates the performance and user friendliness of the tool and user satisfaction evaluates all aspects e.g., speed, usability of end user's interaction with this tool. 100 users tested the tool by giving various questions regarding tomato diseases. The most recommended task level questionnaire, Single Ease Question (SEQ) is given for evaluating accuracy and tool design. Likewise, a test level questionnaire, Standardized User Experience Percentile Rank Questionnaire (SUP-Q) is given to evaluate user satisfaction. The user rating starts from 1 for Very Poor, 2 for Poor, 3 for Fair, 4 for Good and 5 for Very Good. The aggregate data is given in the below tables Table 1. and Table 2.

Table 1. User Ratings for Text Query

Subjective Evaluation for Text Query (User rating 1 - Very poor to 5 - Very good)			
User Rating	*Answer Accuracy*	*Tool Design*	*User Satisfaction*
Very Poor	0	0	0
Poor	0	0	0
Fair	20	30	35
Good	65	60	55
Very Good	15	10	10

Table 2. User Ratings for Voice Query

Subjective Evaluation for Voice Query (User rating 1 - Very poor to 5 - Very good)			
User Rating	*Answer Accuracy*	*Tool Design*	*User Satisfaction*
Very Poor	0	0	0
Poor	0	0	0
Fair	30	40	10
Good	60	50	10
Very Good	10	10	10

The chart for the data is shown below in the Figure 6 and Figure 7.

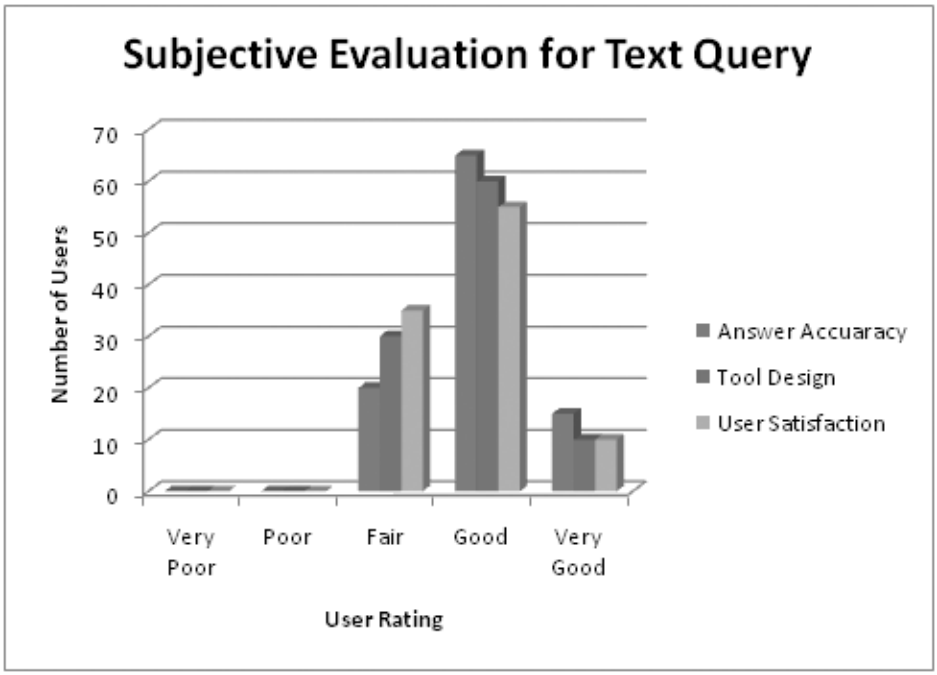

Figure 6. Subjective Evaluation of Text Query

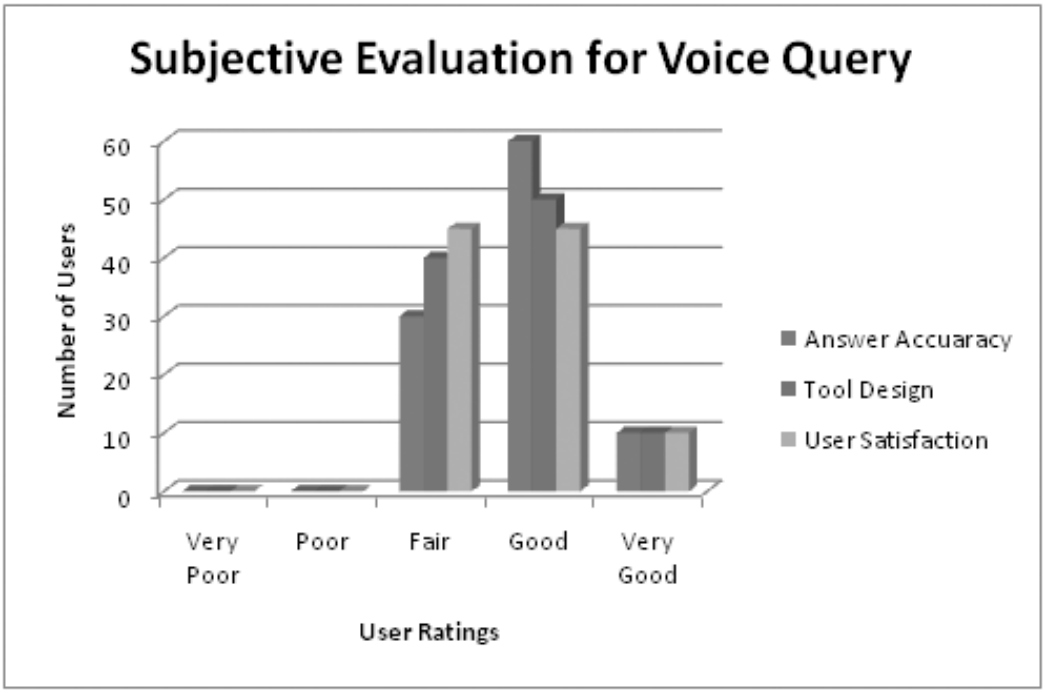

Figure 7. Subjective Evaluation of Voice Query

The users vary in different aspects including their living background, gender, occupation and age group. The subjective evaluation shows satisfactory results about the tool's accuracy in retrieving answers. The tool responses very well for factoid questions like "what" and "how" type of questions. But, the responses for Yes//No questions are not accurate. The tool needs to be improvised for all the question types. In future, this proposed QA technique will be implemented as a mobile web application and the computation sensitive part will be offloaded to the cloud server. The farmers can easily avail and access the benefits of the proposed system from their smart phones.

5. CONCLUSION

The proposed refined LSA technique for agriculture in MCC helps farmers to get the most relevant and accurate answers than millions of relevant documents. There are many QA web applications available for location finding and for medical field. In mobile phones, many mobile phone applications for agriculture that uses Short Message Service(SMS) to answer the farmer's queries. Most of the agriculture based applications in mobiles use SMS and some applications are automatic recommendation system that suggests and connects the experts through online. The proposed technique is very much accurate and useful for the farmers those who need only answers for their queries. The voice based query asking helps them a lot. This idea can be more specific and effective if it is designed for region like Tamilnadu by implementing the QAS in Tamil.

REFERENCES

[1] https://en.wikipedia.org/wiki/Question_answering

[2] http://www.highlandstoday.com/list/highlands-agri-leader-news/shortage-of-agricultural-scientists-is-a-real-problem-20140528

[3] Qingtian Zeng,, Zhichao Liang,, Weijian Ni, Hua Duan, "OAPRS: An Online Agriculture Prescription Recommendation System", *7th International Conference on Computer and Computing Technologies in Agriculture,*.327-336, 2014.

[4] David Tobinski, Oliver Kraft, "Latent Semantic Analysis as Method for Automatic Question Scoring**,"** *Proceedings of the First International Workshop on Artificial Intelligence and Cognition*, **1100**, 100-105.

[5] João Carlos Alves dos Santos,Eloi Luiz Favero, "Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers", *Journal of the Brazilian Computer Society,***1,** 2015.

[6] Priyanka Singh, Dr. Elena Simperl, "Using Semantics to Search Answers for Unanswered Questions in Q&A Forums", *Proceedings of the 25th International Conference Companion on World Wide Web*, 699-706, 2016.

[7] Asma Ben Abacha, Pierre Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies" *Information Processing & Management*,**51**, 570-594, 2015.

[8] Amit Mishra, Sanjay Kumar Jain, "A survey on question answering systems with classification", *Journal of King Saud University - Computer and Information Sciences*, 2014.

[9] Wissal Brini , Mariem Ellouze , Slim Mesfar , Lamia Hadrich Belguith," An Arabic question-answering system for factoid questions", *International Conference on Natural Language Processing and Knowledge Engineering*, 1-7, 2009.

[10] https://www.apple.com/in/business/docs/iOS_Security_Guide.pdf

[11] Johann Hauswald, Michael A. Laurenzano, Yunqi Zhang, Cheng Li, Austin Rovinski, Arjun Khurana, Ronald G. Dreslinski, Trevor Mudge, Vinicius Petrucci1, Lingjia Tang, Jason Mars," Sirius: An Open End-to-End Voice,*2th International Conference on Architectural Support for Programming Languages and Operating Systems*, **50**, 223-238, 2015.

[12] T.Kawamura, A. Ohsuga, "Question-Answering for Agricultural Open Data, Transactions on Large-Scale Data- and Knowledge-Centered Systems", **8960**, 15-28, 2015.

[13] Hetal Patel, Dr. Dharmendra Patel, Survey Of Android Apps For Agriculture Sector, *International Journal of Information Sciences and Techniques*, Vol.6, 1-5, 2016

[14] Kuldeep Sambrekar, V. S. Rajpurohit, A Proposed Model for Mobile Cloud Computing in Agriculture, *International Journal for Scientific Research & Development*,Vol. 2, Issue 07,429-432, 2014

[15] https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/