International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 24 Issue 9 – JUNE 2017.

Developing a Semantic Architecture for Question Answering Selection and Retrieval Systems

J.kirupavathy[#1] , Dr.J.Jebamalar Tamilselvi[#2] and M.Kumaran[#3]

[#]*Jaya Engineering College, Thiruninravur-602 204,Tamilnadu, India*

***Abstract*— Recent developments made in the web services have applied to the Information retrieval tasks. Semantic matching is a critical task for many applications in several Natural Languages processing like question answering scheme, etc. Keyphrases is the subfield that contains metadata that summarizes and characterize the documents. Though, previous techniques were introduced a key phrase extraction model, still the issues like word mismatching, misidentification of the words are not yet focused. In this paper, we have proposed an efficient keyphrase extraction model that efficiently retrieves the relevant data in lesser time. We have constructed machine learning models which build an index for every keyword. Firstly, the keyword is allowed for stemming process that eliminates the stopwords in the sentences. Then, the stemmed words is further allowed to build into normalized words that combines with Medinet and Wordnet. By doing so, we have achieved faster-response time for query retrieval process of the Question Answering scheme. Experimental results have shown the efficiency of the proposed system.**

***Index Terms*— Information retrieval, machine learning model, meta data, Semantic matching and normalized words.**

I. Introduction

Semantic matching is a critical task for many applications in natural language processing (NLP), such as information retrieval [1], question answering [2] and paraphrase identification. Taking question answering as an example, given a pair of question and answer, a matching function is required to determine the matching degree between these two sentences. Recently, deep neural network based models have been applied in this area and achieved some important progresses. A lot of deep models follow the paradigm to first represent the whole sentence to a single distributed representation, and then compute similarities between the two vectors to output the matching score. In general, this paradigm is quite straightforward and easy to implement, however, the main disadvantage lies in that important local information is lost when compressing such a complicated sentence into a single vector.

A central topic in developing intelligent search systems is to provide answers in finer-grained text units, rather than to simply rank lists of documents in response to Web queries. This can not only save the users' efforts in fulfilling their information needs, but also will improve the user experience in applications where the output bandwidth is limited, such as mobile Web search and spoken search. Significant progress has been made at answering factoid queries [18], such as "how many people live in Australia?", as defined in the TREC QA track. However, there are diverse Web queries which cannot be answered by a short fact, ranging from advice on fixing a mobile phone, to requests for opinions on some public issues. Retrieving answers for these "non-factoid" queries from Web documents remains a critical challenge in Web question answering (WebQA).

Key phrases such as named entities (person, location and organization names), book and movie titles, science, medical or military terms and other, are usually among the most information-bearing linguistic structures. Translating them correctly will improve the performance of cross-lingual information retrieval, question answering and machine translation systems [4]. However, these key phrases are often domain-specific, and people. Some name and terminology is a single word, which could be regarded as a one-word phrase. Instantly. We create new key phrases which are not covered by existing bilingual dictionaries or parallel corpora, therefore standard data-driven or knowledge-based machine translation systems cannot translate them correctly. As an increasing amount of web information becomes available, exploiting such a huge information resource is becoming more attractive. searched the web for parallel corpora while [5] extracted translation pairs from anchor texts pointing to the same webpage. However, parallel webpages or anchor texts are quite limited, and these approaches greatly suffer from the lack of data [6].

The rest of the paper is organized as follows: Section II describes related work; Section III describes the proposed work; Section IV describes the experimental analysis and concludes in Section V.

II. RELATED WORK

This section depicts the existing approaches carried out in the field of key phrases extractions.

A. *Preliminaries*

Automatic key phrase extraction systems have been evaluated on corpora from a variety of sources ranging from long scientific publications to short paper abstracts and email messages. There are at least four corpus-related factors that affect the difficulty of key phrase extraction.

Length: The difficulty of the task increases with the length of the input document as longer documents yield more candidate keyphrases. For instance, each Inspec abstract has on average 10 annotator-assigned keyphrases and 34 candidate keyphrases [6]. In contrast, a scientific paper typically has at least 10 keyphrases and hundreds of candidate keyphrases, yielding a much bigger search space. Consequently, it is harder to extract keyphrases from scientific papers, technical reports, and meeting transcripts than abstracts, emails, and news articles.

Structural consistency: In a structured document, there are certain locations where a keyphrase is most likely to appear. For instance, most of a scientific paper's keyphrases should appear in the abstract and the introduction. While structural information has been exploited to extract keyphrases from scientific papers (e.g., title, section information), web pages (e.g., metadata), and chats (e.g., dialogue acts), it is most useful when the documents from a source exhibit structural similarity. For this reason, structural information is likely to facilitate keyphrase extraction from scientific papers and technical reports because of their standard format (i.e., standard sections such as abstract, introduction, conclusion, etc.). In contrast, the lack of structural consistency in other types of structured documents (e.g., web pages, which can be blogs, forums, or reviews) may render structural information less useful.

Topic change: An observation commonly exploited in keyphrase extraction from scientific articles and news articles is that keyphrases typically appear not only at the beginning [8] but also at the end of a document. This observation does not necessarily hold for conversational text (e.g., meetings, chats), however. The reason is simple: in a conversation, the topics (i.e., its talking points) change as the interaction moves forward in time, and so do the keyphrases associated with a topic. One way to address this complication is to detect a topic change in conversational text [9]. However, topic change detection is not always easy: while the topics listed in the form of an agenda at the beginning of formal meeting transcripts can be exploited, such clues are absent in casual conversations (e.g., chats).

Topic correlation: Another observation commonly exploited in keyphrase extraction from scientific articles and news articles is that the keyphrases in a document are typically related to each other [10]. However, this observation does not necessarily hold for informal text (e.g., emails, chats, informal meetings, personal blogs), where people can talk about any number of potentially uncorrelated topics. The presence of uncorrelated topics implies that it may no longer be possible to exploit relatedness and therefore increases the difficulty of keyphrase extraction.

B. *Existing approaches*

Generally, the keyphrase extraction executes in the following steps:

- Extracting a list of words/phrases that serve as candidate keyphrases using some heuristics.
- Determining which of these candidate keyphrases are correct keyphrases using supervised or
- Unsupervised approaches.

1) *Selecting candidates words or phrases*

As noted before, a set of phrases and words is typically extracted as candidate keyphrases using heuristic rules. These rules are designed to avoid spurious instances and keep the number of candidates to a minimum. Typical heuristics include (1) using a stop word list to remove stop words, (2) allowing words with certain partof-speech tags (e.g., nouns, adjectives, verbs) to be candidate keywords (3) allowing n-grams that appear in Wikipedia article titles to be candidates and (4) extracting n-grams or noun phrases that satisfy pre-defined lexico-syntactic pattern(s) [11].

Many of these heuristics have proven effective with their high recall in extracting gold keyphrases from various sources. However, for a long document, the resulting list of candidates can be long [12]. Consequently, different pruning heuristics have been designed to prune candidates that are unlikely to be keyphrases.

a) *Supervised approaches:*

Research on supervised approaches to keyphrase extraction has focused on two issues: task reformulation and feature design.

b) *Task reformulation:*

Early supervised approaches to keyphrase extraction recast this task as a binary classification problem [13]. The goal is to train a classifier on documents annotated with keyphrases to determine whether a candidate phrase is a keyphrase. Keyphrases and non-keyphrases are used to generate positive and negative examples, respectively. Different learning algorithms have been used to train this classifier, including naive Bayes [14]

c) *Feature selection*

Structural features encode how different instances of a candidate keyphrase are located in different parts of a document. A phrase is more likely to be a keyphrase if it appears in the abstract or introduction of a paper or in the metadata section of a web page. In fact, features that encode how frequently a candidate keyphrase occurs in various sections of a scientific paper (e.g., introduction, conclusion) and those that encode the location of a candidate keyphrase in a web page (e.g., whether it appears in the title)[16] have been shown to be useful for the task.

Syntactic features encode the syntactic patterns of a candidate keyphrase. For example, a candidate keyphrase has been encoded as (1) a PoS tag sequence, which denotes the sequence of part-of-speech tag(s) assigned to its word(s); and (2) a suffix sequence, which is the sequence of morphological suffixes of its words. However, ablation studies conducted on web pages and scientific articles reveal that syntactic features are not useful for keyphrase extraction in the presence of other feature types.

2) *Unsupervised approaches*

a) Graph based ranking

Intuitively, keyphrase extraction is about finding the important words and phrases from a document. Traditionally, the importance of a candidate has often been defined in terms of how related it is to other candidates in the document. Informally, a candidate is important if it is related to (1) a large number of candidates and (2) candidates that are important. Researchers have computed relatedness between candidates using co-occurrence counts and semantic relatedness and represented the relatedness information collected from a document as a graph [14].

This instantiation of a graph-based approach overlooks an important aspect of keyphrase extraction, however. A set of keyphrases for a document should ideally cover the main topics discussed in it, but this instantiation does not guarantee that all the main topics will be represented by the extracted keyphrases [17]. Despite this weakness, a graph-based representation of text was adopted by many approaches that propose different ways of computing the similarity between two candidates.

b) Topic based clustering

Another unsupervised approach to keyphrase extraction involves grouping the candidate keyphrases in a document into topics, such that each topic is composed of all and only those candidate keyphrases that are related to that topic. There are several motivations behind this topic-based clustering approach. First, a keyphrase should ideally be relevant to one or more main topic(s) discussed in a document. Second, the extracted keyphrases should be comprehensive in the sense that they should cover all the main topics in a document. Below we examine three representative systems that adopt this approach.

KeyCluster: The author in [15] adopts a clustering-based approach (henceforth KeyCluster) that clusters semantically similar candidates using Wikipedia and co-occurrence-based statistics. The underlying hypothesis is that each of these clusters corresponds to a topic covered in the document, and selecting the candidates close to the centroid of each cluster as keyphrases ensures that the resulting set of keyphrases covers all the topics of the document.

III. PROPOSED WORK

This section depicts the working of the enhanced semantic architecture of the keyphrase extraction system. The thought of this proposed system arise from these issues:

- Recognition of key terms
- Misidentification of the words
- Lack of artificial intelligence
- Lack of machine learning
- Time delay for answers

To overcome from the above mentioned issues, we have proposed ranking based relevant answering systems.We have build keyphrase extraction technique that efficiently support the multiple languages. The proposed keyphrase extraction process consists of following modules:

A. *Q and A application:*

The previous web applications posted the questions and it's answered by other users. This kind of action leads to greater redundancy and non-trusted system. In the perspective of medical practitioners, this system imposes non-trusted environment. In order to resolve this, we have built an efficient Q and A scheme that presents faster response to the answered questions and makes the user-friendly environment.

B. *Key concept detection:*

The reason behind this fast-answering system is the deployment of Natural Language Processing (NLP). The objective of the NLP system is to efficiently return the answers from the relevant key terms. It specifically deals with the Parts of Speech Tagging (POST) that analyzes the phrases and nouns of the given terms. Before processing, stemming process is involved to eliminate the stopwords. This step investigates on the specific keywords from the given base words.

C. *Bridging the answers:*

Based on the given base words, the proper meaning will be analyzed with the help of English dictionary and medical terms. Normalization is the process executes after the completion of stemming process. A domain specific knowledge is given in the normalization process. The relevant answers are obtained from the Local Mining Database using the normalized words.

D. *Machine learning and Language translation:*

Machine learning process operates from the use of local mining and global learning techniques. Eventually, the local mining database is updated for every given new base words. The global learning system contains a vast amount of medical related queries and terms. This will acts as backend system to retrieve the related resource to the query. An index is constructed for every keyword, so as to retrieve the words easily and at less time. If the resource is unavailable, the query will be answered later.

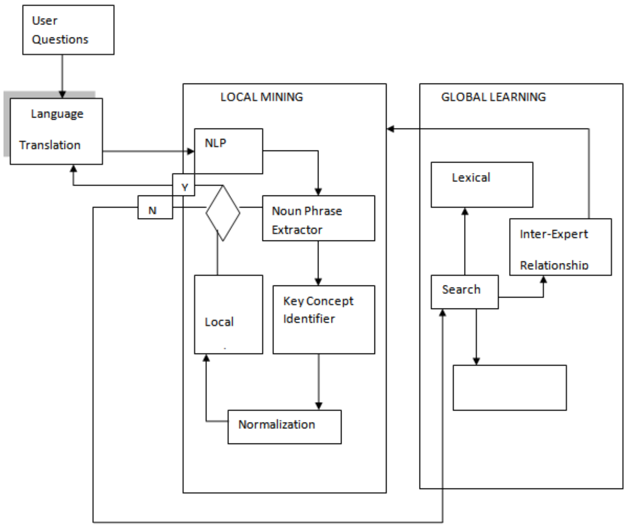


Fig.1. Proposed architecture diagram

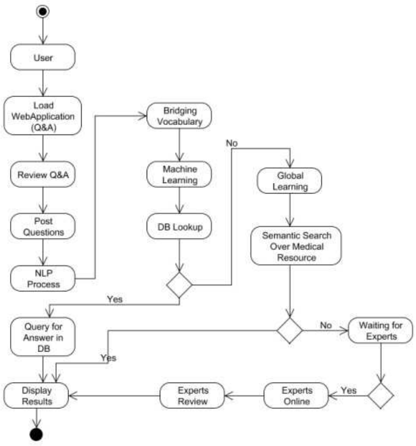
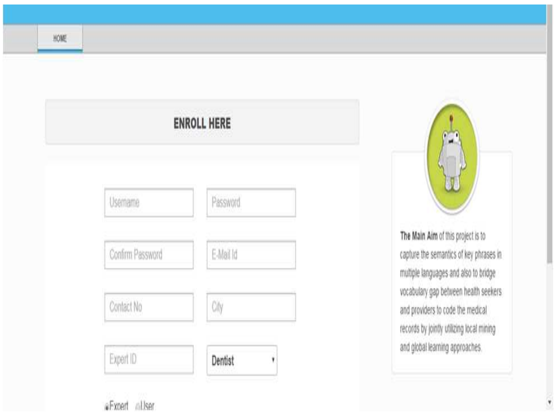


Fig.2. Activity diagram for the proposed framework

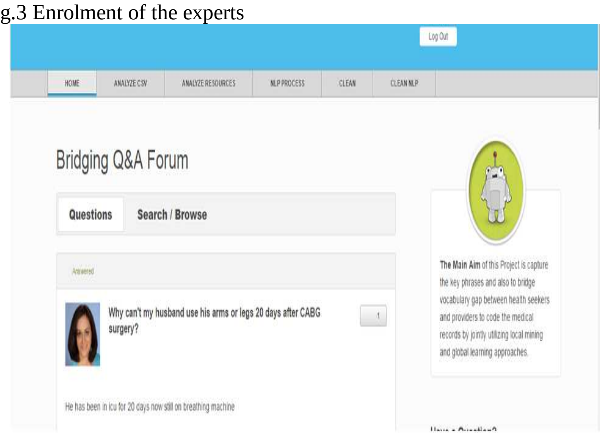
IV. EXPERIMENTAL ANALYSIS

This section depicts the experimental analysis of our proposed techniques via Java programming language. In order to obtain the proper meaning of the relevant terms, we have used WordNet and Medinet. The following screens are the deployment of our proposed technique.



The Main Aim of this project is to capture the semantics of key phrases in multiple languages and also to bridge vocabulary gap between health seekers and providers to code the medical records by jointly utilizing local mining and global learning approaches.

Fig.3 Enrolment of the experts

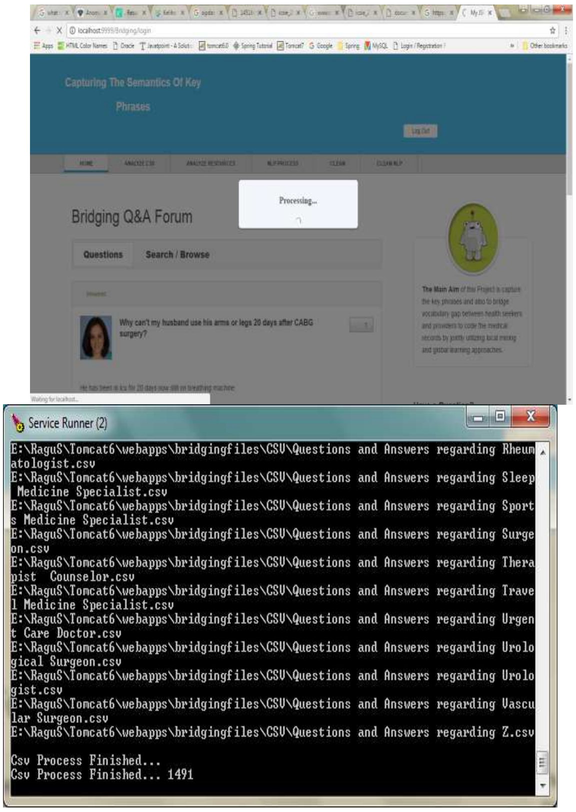


Why can't my husband use his arms or legs 20 days after CABG surgery?

He has been in icu for 20 days now still on breathing machine

The Main Aim of this Project is capture the key phrases and also to bridge vocabulary gap between health seekers and providers to code the medical records by jointly utilizing local mining and global learning approaches.

Fig.4 Development of Q and A forum



Why can't my husband use his arms or legs 20 days after CABG surgery?

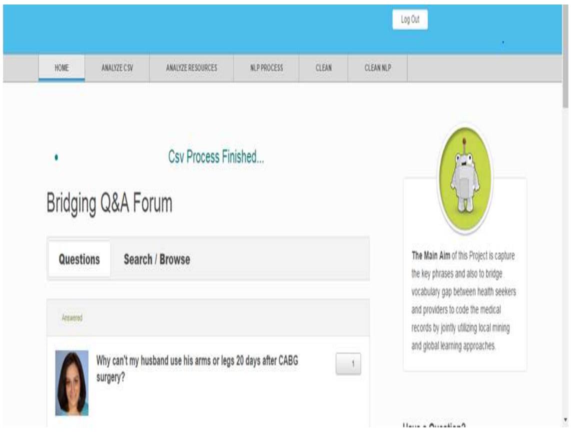
He has been in icu for 20 days now still on breathing machine

The Main Aim of this Project is capture the key phrases and also to bridge vocabulary gap between health seekers and providers to code the medical records by jointly utilizing local mining and global learning approaches.

```
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Rheum
atologist.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Sleep
 Medicine Specialist.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Sport
s Medicine Specialist.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Surge
on.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Thera
pist  Counselor.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Trave
l Medicine Specialist.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Urgen
t Care Doctor.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Urolo
gical Surgeon.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Urolo
gist.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Vascu
lar Surgeon.csv
E:\RaguS\Tomcat6\webapps\bridgingfiles\CSV\Questions and Answers regarding Z.csv

Csv Process Finished...
Csv Process Finished... 1491
```

Fig.5 Processing of the posted queries

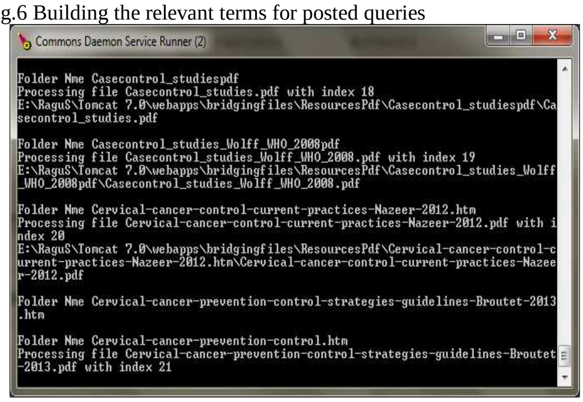


Csv Process Finished...

Why can't my husband use his arms or legs 20 days after CABG surgery?

The Main Aim of this Project is capture the key phrases and also to bridge vocabulary gap between health seekers and providers to code the medical records by jointly utilizing local mining and global learning approaches.

Fig.6 Building the relevant terms for posted queries



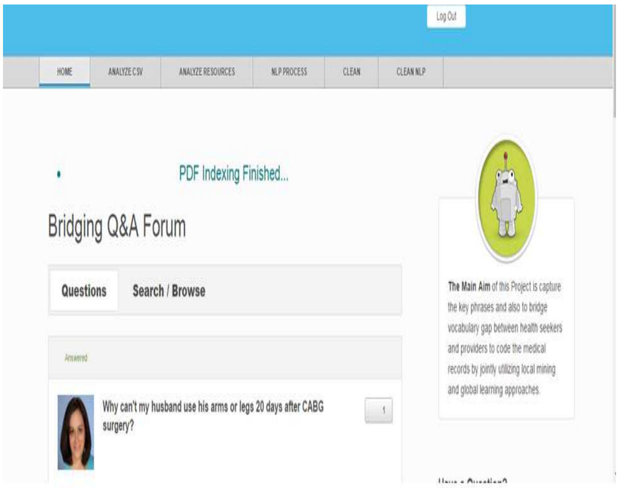
```
Folder Nme Casecontrol_studiespdf
Processing file Casecontrol_studies.pdf with index 18
E:\RaguS\Tomcat 7.0\webapps\bridgingfiles\ResourcesPdf\Casecontrol_studiespdf\Ca
secontrol_studies.pdf

Folder Nme Casecontrol_studies_Wolff_WHO_2008pdf
Processing file Casecontrol_studies_Wolff_WHO_2008.pdf with index 19
E:\RaguS\Tomcat 7.0\webapps\bridgingfiles\ResourcesPdf\Casecontrol_studies_Wolff
_WHO_2008pdf\Casecontrol_studies_Wolff_WHO_2008.pdf

Folder Nme Cervical-cancer-control-current-practices-Nazeer-2012.htm
Processing file Cervical-cancer-control-current-practices-Nazeer-2012.pdf with i
ndex 20
E:\RaguS\Tomcat 7.0\webapps\bridgingfiles\ResourcesPdf\Cervical-cancer-control-c
urrent-practices-Nazeer-2012.htm\Cervical-cancer-control-current-practices-Nazee
r-2012.pdf

Folder Nme Cervical-cancer-prevention-control-strategies-guidelines-Broutet-2013
.htm

Folder Nme Cervical-cancer-prevention-control.htm
Processing file Cervical-cancer-prevention-control-strategies-guidelines-Broutet
-2013.pdf with index 21
```



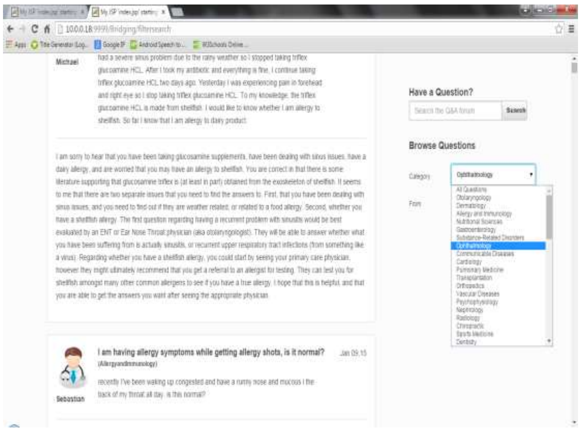
PDF Indexing Finished...

Bridging Q&A Forum

Why can't my husband use his arms or legs 20 days after CABG surgery?

The Main Aim of this Project is capture the key phrases and also to bridge vocabulary gap between health seekers and providers to code the medical records by jointly utilizing local mining and global learning approaches.

Fig.7 Extracting the data into structural form.



had a severe sinus problem due to the rainy weather so I stopped taking triflex glucoamine HCL. After I took my antibiotic and everything is fine, I continue taking triflex glucoamine HCL two days ago. Yesterday I was experiencing pain in forehead and right eye so I stop taking triflex glucoamine HCL. To my knowledge, the triflex glucoamine HCL is made from shellfish. I would like to know whether I am allergy to shellfish. So far I know that I am allergy to dairy product.

I am sorry to hear that you have been taking glucosamine supplements, have been dealing with sinus issues, have a dairy allergy, and are worried that you may have an allergy to shellfish. You are correct in that there is some literature supporting that glucosamine triflex is (at least in part) obtained from the exoskeleton of shellfish. It seems to me that there are two separate issues that you need to find the answers to. First, that you have been dealing with sinus issues, and you need to find out if they are weather related, or related to a food allergy. Second, whether you have a shellfish allergy. The first question regarding having a recurrent problem with sinusitis would be best evaluated by an ENT or Ear Nose Throat physician (aka otolaryngologist). They will be able to answer whether what you have been suffering from is actually sinusitis, or recurrent upper respiratory tract infections (from something like a virus). Regarding whether you have a shellfish allergy, you could start by seeing your primary care physician, however they might ultimately recommend that you get a referral to an allergist for testing. They can test you for shellfish amongst many other common allergens to see if you have a true allergy. I hope that this is helpful, and that you are able to get the answers you want after seeing the appropriate physician.

I am having allergy symptoms while getting allergy shots, is it normal?
(Allergyandimmunology)

recently I've been waking up congested and have a runny nose and mucous I the back of my throat all day. is this normal?

Fig.8 Obtaining the relevant results for the given key terms.

V. CONCLUSION

Automatic keyphrase extractions have been widely studied by the research communities. The state of the art reveals that still the performance of the keyphrase extraction is not yet achieved successfully, in accord to user's requirements. An automatic mining of data from the relevant document is known as keyphrase extraction system. In this paper, we have proposed intelligent keyphrase extraction techniques that automatically extract the relevant keywords from the given set of documents. We have built an efficient Q and A scheme that posts and answers the questions in a rapid time. It combines with Medinet and Wordnet corpus to order and retrieve the data in a stipulate period of time. Document keyphrases have enabled fast and accurate searching for a given document from a large text collection, and have exhibited their potential in improving many natural language processing (NLP) and information retrieval (IR) task. Experimental results have shown the efficiency of our proposed system.

REFERENCES

[1] Wei Nan Zhang et al, "Capturing the Semantics of Key Phrases Using Multiple Languages for Question Retrieval", IEEE Transactions on Knowledge and Data Engineering, 28 (4), 2016.

[2] X. Cao, G. Cong, B. Cui, C. S. Jensen, andQ. Yuan, "Approaches to exploring category information for question retrieval in community question-answer archives", ACM Trans. Inf. Syst., vol. 30,no. 2, p. 7, 2012.

[3] J. H. Park and W. B. Croft, "Query term ranking based ondependency parsing of verbose queries," inProc. ACL, 2010,pp. 829–830.

[4] J. Allan, J. P. Callan, W. B. Croft, L. Ballesteros, J. Broglio, J. Xu,and H. Shu, "INQUERY at TREC-5," inProc. TREC, 1996, pp. 119–132.

[5] J. P. Callan, W. B. Croft, and J. Broglio, "TREC and tipster experiments with INQUERY," inProc. Inf. Process. Manage., 1995,pp. 327–343.

[6] M. Bendersky and W. B. Croft, "Discovering key concepts in ver-bose queries," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res.Develop. Inf. Retrieval, 2008, pp. 491–498.

[7] K. Collins-Thompson and J. Callan, "Query expansion using random walk models," inProc. 14th ACM Int. Conf. Inf. Knowl. Man-age., 2005, pp. 704–711.

[8] J. Xu and W. B. Croft, "Query expansion using local and globaldocument analysis," inProc. 19th Annu. Int. ACM SIGIR Conf. Res.Develop. Inf. Retrieval, 1996, pp. 4–11.

[9] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions inlarge question and answer archives," in Proc. 14th ACM Int. Conf.Inf. Knowl. Manage., 2005, pp. 84–90.

[10] G. Zhou, L. Cai, J. Zhao, and K. Liu, "Phrase-based translationmodel for question retrieval in community question answerarchives," inProc. 49th Annu. Meeting Assoc. Comput. Linguistics:Human Lang. Technol. - Vol. 1, 2011, pp. 653–662.

[11] P. Resnik and N. A. Smith, "The web as a parallel corpus,"Com-put. Linguist., vol. 29, no. 3, pp. 349–380, Sep. 2003.

[12] P. D. Turney, "Learning algorithms for keyphrase extraction,"Inf.Retr., vol. 2, pp. 303–336, 2000.

[13] A. Hulth, "Improved automatic keyword extraction given morelinguistic knowledge," in Proc. Conf. Empirical Methods NaturalLang. Process., 2003, pp. 216–223.

[14] M. Bendersky and W. B. Croft, "Modeling higher-order termdependencies in information retrieval using query hypergraphs,"inProc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012,pp. 941–950.

[15] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Query expansion bymining user logs,"IEEE Trans. Knowl. Data Eng., vol. 15, no. 4,pp. 829–839, Jul. 2003.

[16] D. Buscaldi, P. Rosso, and E. S. Arnal, "A WordNet-based queryexpansion method for geographical information retrieval,"Work.Notes Clef Workshop, 2005.

[17] S. Riezler, A. Vasserman, I. Tsochantaridis, V. O. Mittal, and Y.Liu, "Statistical machine translation for query expansion inanswer retrieval," inProc. 45th Annu. Meeting Assoc. Comput. Linguistics, 2007, pp. 464–471.

[18] J. Gao and J.-Y. Nie, "Towards concept-based translation modelsusing search logs for query expansion," inProc. 21st ACM Int.Conf. Inf. Knowl. Manage., 2012.

[19] R. A. Baeza-Yates and B. A. Ribeiro-Neto, Modern InformationRetrieval - The Concepts and Technology Behind Search, 2nd ed. Har-low, England, U.K: Pearson Education Ltd., 2011.

[20] C. Bannard and C. Callison-Burch, "Paraphrasing with bilingualparallel corpora," inProc. ACM Int. Conf. Inf. Knowl. Manage.,2005, pp. 597–604

BIOGRAPHY

#1 J.Kirupavathy., received BE degree in CSE from Sakthi engineering college, India in 2011 and her interests are data mining, Networking and Data Structure.

***2 Dr.J.JebamalarTamilselvi** received her Ph.D. in 2009 from the Department of Computer Applications at Karunya University, Coimbatore, INDIA. She received her B.Sc. (Computer Science) from Manonmanium Sundaranar University of Tamil Nadu, INDIA in 2003 and MCA Degree from Anna University, Coimbatore, Tamil Nadu, INDIA in 2006. Her area of interest includes Data cleansing approaches, Data Extraction, Data Integration, Data Warehousing and Data Mining. She is a life Member of International Association of Engineers (IAENG), International Association of Computer Science and Information Technology (IACSIT), and the Society of Digital Information and Wireless Communications. Reviewer and Member of International Journal of Engineering Science and Technology (IJEST) Member and Convergence Information Technology (JCIT). Her research has been accepted and published in 17 international journals, and 12 national and international conferences. She had been awarded the P.K Das Memorial Best Faculty Award in 2014 by the Nehru Group of Institutions, Coimbatore and the Education and Research Award in 2015 by the Karunya University, Coimbatore.

$3 M.Kumaran He has completed his Under Graduate from University of Madras and Post Graduate from Anna University. He has 18 years of experience in teaching. His area of interest is Open Source system Developments, Information Security and Protocol Management.