International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Machine Learning Approach for Question Answering in Information Retrieval

G. N. Prithy*1, S. Selvi2

*1Computer Science , R. M. K Engineering College, Chennai , Tamil Nadu , India

2Computer Science , R. M. K Engineering College, Chennai , Tamil Nadu , India

ABSTRACT

CQA (Community Question Answering) is a major challenge nowadays due to the popularity and advantages of CQA archives over the web. This paper deals with the methods to solve lexical gap problem in question retrieval and providing multiple domain based CQA archives. The aim of question answering in CQA is to find the existing questions that are similar to the question being asked but this has become a big challenge due to the lexical gap problem. In this paper, we have proposed to learn the word embedding and the category under which the question is being asked by using natural language processing. Three methods are being used. One is local mining and the other is global mining. In local mining the question s answer is checked in the local database, if found the answer is retrieved back else the global mining process starts in which the answer is checked in other sites information and the relevant answer is retrieved back. Third concept is the expert level where the answer for the query is got from an expert.

Keywords : CQA, Natural Language Processing, Mining.

I. INTRODUCTION

Data mining is a computing process of discovering patterns in large data sets by using methods at the intersection of machine learning, statistics, and database systems. It is an important process where intelligent methods are applied to extract the data patterns. It is also an interdisciplinary sub field of the computer science. The overall aim of the data mining process is to extract meaningful information from a data set and convert them into an understandable structure for further use. Apart from raw analysis step, it also involves database and data management aspects, data preprocessing, model and inference considerations, interesting metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the process of analysing the "knowledge discovery in databases" or KDD. The terms related are data fishing,data dredging, and data snooping.These terms refer to the use of data mining methods to sample parts of a bigger population data set which are (or may be) too small for making any reliable statistical inferences about the validity of the patterns discovered. These methods can be used to create new hypotheses for testing them against the larger data populations. The knowledge discovery in databases (KDD) process commonly has the following stages: (1) Selection (2) Preprocessing (3) Transformation (4) Data mining (5) Interpretation/evaluation.

Data Mining has six major classes of tasks: (1)Anomaly detection(outline/change/deviation detection) identification of unusual data sets, which could be interesting or data errors that might require further new investigation. (2)The Association rule learning (dependency modeling) Searches for finding relationships between variables. For example, a supermarket might collect data regarding their customer purchasing habits by implementing association rule, the supermarket can find out which products are bought together frequently and use this information for making marketing of the goods. This is also referred to as market basket analysis.(3) Clustering is the process of discovering the groups and structures in the data that are similar in some way or another "", without using known structures in data. (4) Classification is the process of generalizing known structure for applying it to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or "spam".(5) Regression attempts to find out a function that models the data with the least error ,that is used for estimating the relationships between data or data sets. (6)Summarization is the last step which is used for providing a more compact representation of the data set, which includes visualization and report generation.

A. CLUSTERING

Clustering is the process of grouping a particular set of data or objects based on their characteristics and aggregating them according to the similarities between them. In data mining, this methodology does data partitioning by implementing a specific join algorithm, which is best suitable for the desired information analysis. This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object must belong to a cluster in a determined degree. More specific divisions can be possible to create like objects that may belong to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

There are several various ways for implementing this partitioning, based on the distinct models. Distinct algorithms are applied to each model, differentiating its properties and results. These models are differentiated by their organization and type of the relationship between them. The most important ones are: Centralized - each cluster is represented by a single vector mean, and each of the object value is compared to these mean values Distributed. the cluster is built by using statistical distributions Connectivity .The connectivity on these models are based upon distance function between elements Group algorithms have only group information Graph , cluster organization and relationship between members is defined by a graph linked structure Density members of the cluster are grouped by the regions where observations are dense and similar

B. Pre-processing

Preprocessing of data is a data mining technique which involves converting raw data into an understandable form. Real- world data is most of the incomplete, inconsistent, and lacks in certain behaviors or trends, and is likely to posses many errors. Preprocessing of data is a method which has proven to resolve such issues. Data preprocessing prepares the raw data for further processing. The Data goes through various series of steps during preprocessing: (1) Data Cleaning: Data is cleansed through processes like filling in missing values, smoothening of noisy data, or resolving the inconsistencies present in the data. (2) Data Integration: Inn this Data with various forms of representations are put together and conflicts within those data are resolved. (3) Data Transformation:In this the Data is first normalized, aggregated and generalized. (4) Data Reduction: This step aims to provide a reduced representation of the data in a data warehouse. (5) Data Discretization: It Involves in

reducing the number of values of a continuous attribute by dividing the range of attribute intervals.

C. NLP

Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics .It is concerned with the interactions between computers and human (natural) languages.It is particularly, concentrates on programming the computer to effectively process large natural language corpora. Challenges associated with natural language processing are natural language understanding, natural language generation (frequently from formal, machine-readable logical forms), machine perception,connecting language and dialog systems.Even some combinations of them are possible. part-of-speech tagging (POS tagging or POST), is also called as grammatical tagging or word-category disambiguation.It is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context i.e., identification of its relationship with the adjacent and related words in a phrase, sentence, or paragraph. Simplified form of this part-of-speech is commonly taught to school children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

II. RELATED WORK

There has been a significant increase in the number of community-based question and answer services on the Web, where people answer other people's questions. These services build up large archives of questions and answers rapidly, and these archives serve as a valuable linguistic resource.[1] Proposed Finding Similar Questions in Large Question and Answer Archives. One of the major tasks in a question and answer service is to find questions in the archive that are semantically equivalent to a user's question. This enables retrieval of high quality answers from the archive and removes the time lag related with the community-based system. In this paper, we discuss all possible methods for question retrieval that are based upon using the similarity between answers in the archive to estimate probabilities for a translation-based retrieval model. This model is capable of finding semantically similar questions with relatively little word overlap.

[10] proposed improving Question Retrieval in Community Question Answering Using World Knowledge Community question answering (CQA), which provides a platform for people with diverse background to share information and knowledge ,has become an increasingly popular research topic. In this paper, we focus on the task of question retrieval. The main problem in question retrieval is to measure the similarity between the queried questions and the questions which have been solved by other users previously and already existing. The traditional methods measure the similarity based on the bag-of-words (BOWs) representation. This representation neither captures dependencies between related words, nor handles synonyms or polysemous words. In this work, we first propose a way to build a concept 6 thesaurus based on the semantic relations extracted from the world knowledge of Wikipedia. Then, we develop a unified framework to leverage these semantic relations in order to enhance the question similarity in the concept space. Experiments conducted on a real CQA data set revealed that with the help of Wikipedia thesaurus, the performance of question retrieval was improved in comparision with the traditional methods.

[7] proposed Learning the Latent Topics for Question Retrieval in Community QA .Community-based Question Answering (CQA) is a popular online service where users can ask and answer questions on any topics. This paper deals with the problem of question retrieval. Question retrieval in CQA aims to find historical questions that are semantically equivalent or relevant to the queried questions. Although the translation-based language model has gained the state-of- the-art performance for question retrieval, they ignore the latent topic information in calculating the semantic similarity between questions.

In this paper, we propose a topic model incorporated with the category information into the process of is covering the latent topics in the content of questions.

Then we combine the semantic similarity based latent topics with the translation-based language model into a unified framework for question retrieval. Experiments are carried out on a real world CQA data set from Yahoo! Answers. The results show that our proposed method can significantly improve the question retrieval performance of translation- based language mode.

A. Role of NLP in retrieval

In [8] the requirements for an educational Question Answering (QA) system operating on social media content has been analysed. As a result of which they have identified a set of advanced natural language processing (NLP) technologies to address the issues in educational Question Answer. We conducted an inter-annotator agreement study on 7 subjective question classification in the Yahoo! Answers social Q&A site and propose a simple, but effective approach to automatically identify subjective questions. We also developed a two-stage QA architecture for answering learners questions. In the first step, we aim at re-using human answers to already answered questions by employing question paraphrase identification. In the second step, we apply information retrieval techniques to perform answer retrieval from social media content. We show that elaborate techniques for question preprocessing are crucial

In [2] Retrieval in a question and answer archive involves finding good answers for a user's question. The proposed model works in contrast to typical document retrieval models. The retrieval model for this task exploits question similarity as well as ranks the answers associated with them. A retrieval model that combines translation-based language model for the question part with a query likelihood approach for the answer part has been proposed. It incorporates word-to-word translation probabilities learned through exploiting different information sources. The Experiments performed on this method showed that the proposed translation based language model for the question part performs better when compared to the baseline methods significantly. By combining with the query likelihood language model for the answer part, substantial additional effectiveness improvements have been obtained.

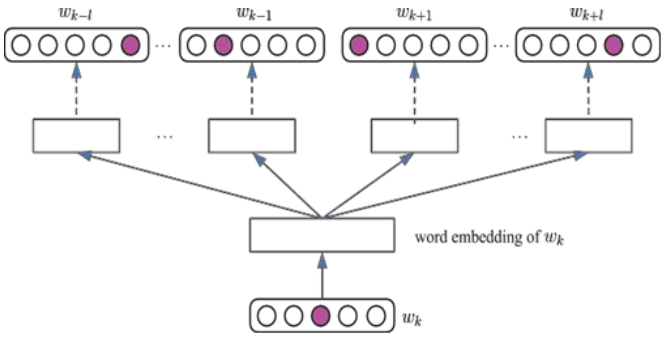

Figure 1. The skip-gram model, which predicts surrounding words given the current word

They have considered the context-aware predicting model. They have more specifically used the Skip-gram model and continuous bag-of-words model for learning word embed- dings, because they were much more efficient and memory-saving than other approaches.

Let wk represent the kth words in the given words sequence w_1; w_2; ; w_N. For simple explanation, they took the example of a Skip-gram model to describe the details. In the Skip-gram model (see Fig. 1), a sliding window is employed on the input text stream to generate the training data, and l indicates the context window size to be 2l 1. In each of the slide window, the model aims to use the central word wk as the input to predict the context words. Let Md×N denote the learned embedding matrix, where N is the vocabulary size and d is the dimension of word embeddings. Each column of M denotes the embedding of a word. Let wk is first mapped to its embedding ewk by selecting the corresponding column vector of *M*. The probability

of its context word $w_{k+j,}$ is then computed using a log-linear *softmax* function as given below.

$$P(w_{k+j}\backslash wk;\theta)=\frac{\exp(e^{T}{}_{wk+j}{}^{e}w_{k})}{\sum^{N}{}_{w=1}\exp(e^{T}{}_{w}e_{wk})}$$

Where u are the parameters we should learned, *k* 1 *d*, and *j€[- 1,1]* . Then, the log-likelihood over the entire training data can be computed as

$$J(\theta)=\sum_{(wk,wk+j)}\log p(w_{k+j} \mid wk;\theta)$$

B. Word Embedding Learning

Word embedding learning has gained wide interests for a variety of NLP tasks (e.g., Chinese word segmentation and part- of-speech tagging (POST) , named entity recognition , dependency parsing, sentiment analysis ,information extraction , question answering etc.). The idea is that the related words tend to be close to each other and with the vector representation. They also demonstrated that the learned word embedding representations could also capture meaningful syntactic and semantic regularities. Among the various word embedding learning methods .

To find out the possible errors for back propagation, we need to compute the derivative of p(wk+j|wk;θ) ,whose computation cost is proportional to the vocabulary size N. Since N is often very large, it is hard to directly find the derivative. To solve this problem, Mikolov proposed a simple negative sampling method. This method generates r noise samples for each of the input word for estimating the target word, in which r is a very small number compared to N. Hence, the training time produces a linear scale to the number of noise samples and it also becomes independent of the vocabulary size. Let us assume that the frequency of the word w is u(w) , then the probability of sampling w is usually set to p(w) α u(w)3/4 .Even though we are using the skip-gram model to illustrate this approach, similar framework can also be developed on the basis of any other word embedding models.

III. OUR APPROACH

The proposed system is a Efficient Q and A Scheme which could give Instant Answers Analyzing the Users Objective behind the Question(Fig.2). Users can ask questions in multiple domain, here health care is more focused. The User posts Questions for instant answers is processed by a natural language processing technique so that the proper meaning would be revealed. The NLP Process comprises a several steps. Of which Parts Of Speech Tagging (POST) results in Phrases and Nouns Extraction. The Keywords thus Extracted is subject to Stemming Process which eliminates the Stop words in the sentence and also trims the keyword for Base Word. The Proper meanings will be analyzed with English Dictionary and the Medical Terms will be Normalized based Domain Specific Knowledge. Different domain Terminologies were Collected and grouped so that the checking with the synonyms of keywords could result in Normalization. The Normalized words will be checked for Contradictions with medical terminologies and the related answers will be queried from Local Mining Database.The use of Local Mining and Global Learning techniques in medical domain. Local Mining database gets updated by the Global learning data once user posts a newer Kind of Query to the Answering System. The Global learning Comprises a large collection of Medical Related Resources in its back end which helps to retrieve a related resource to the Query based on terminology keywords. This Search is completely indexed an thus the retrieval time is faster. In case of resource insufficiency the Query and the Question will be left in pending state till a expert arrives. Once Experts reviewed the query the answers not only dispatches to the Medical Seekers and also updates the Local Mining Database for future instant retrieval to the related Query from other Users, other domain answers retrieved from data set.

A. Q and A Application

Generally, In Existing Web Applications the Questions posted by the users are answered by the Other User which might result in redundancy and user unbelievability especially for medical related doubts, clarifications and questions. In medical sites

a medical Experts who can give believable answers should be available all the time which is practically not possible and time Consuming .So we build a Efficient Q and A Scheme which could give Instant Answers Analyzing the Users Objective behind the Question.

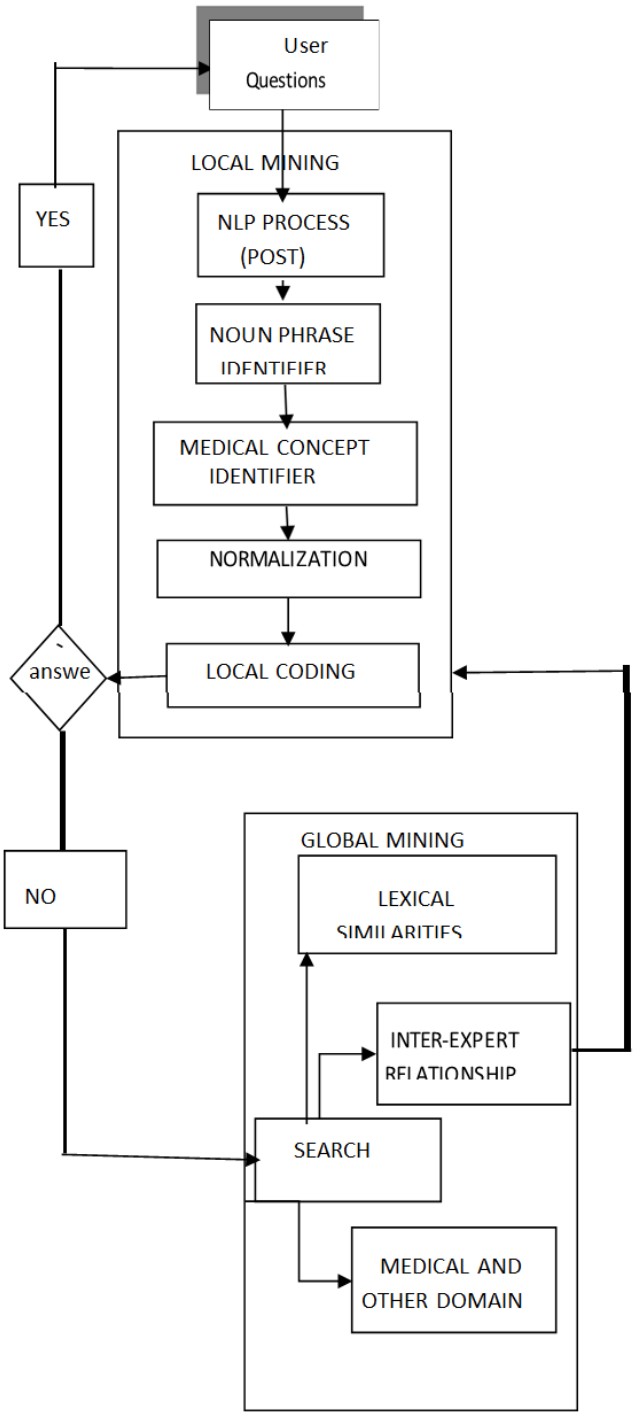

Figure 2. Architecture Diagram

B. NLP Process

The User posts Questions for instant answers is processed by a natural language processing technique so that the proper meaning would be revealed. The NLP Process {Table.1) comprises a several steps. Of which Parts Of Speech Tagging (POST) results in Phrases and Nouns Extraction. The Keywords thus Extracted is subject to Stemming Process which eliminates the Stop words in the sentence and also trims the keyword for Base Word.

C. Bridging Lexical Gap

Proper meanings for each of the word will be analyzed with a Machine Learning in Our Approach is achieved by the use of Local Mining and Global Learning techniques in medical domain. Local Mining database gets updated by the Global learning data once user posts a newer Kind of Query to the Answering System. The Global learning Comprises a large collection of Medical Related Resources in its back end which helps to retrieve a related resource to the Query based on terminology keywords. This Search is completely indexed an thus the retrieval time is faster. In case of resource insufficiency the Query and the Question will be left in pending state till a expert arrives. Once Experts reviewed the query the answers not only dispatches to the Medical Seekers and also updates the Local Mining Database for future instant retrieval to the related Query from other Users, other domain answers retrieved from data set.

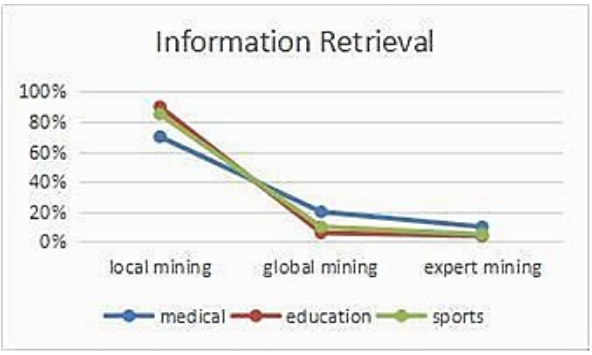

Figure 3. Percentage of information retrieved using local, global and expert mining

TABLE 1. Retrieval Results for "What is causing this very bad pain in my right hip above the pelvic bone?"

Type of Mining	Rank	Related Questions	Extracted Noun,phrase Or domain of expert
Local Mining	1st 6t 13th	Pelvic pain cause sand treatments in women and men 5 common causes of hip pain in women Pain above right hip-causes ,symptoms,treatment	Pelvic pain Hip pain Right hip.
Global Mining	1st 3th 4th	Dull pain on right side above hip Right side pain above hip below ribs Pain in lower right abdomen near hip bone	Pain,hip Right,pain Pain,hip
Expert Mining		What is causing this very bad pain in my right hip above the pelvic bone	Arthrologist

IV. CONCLUSION AND FUTURE WORK

This paper presents a community question answer retrieval by using two enhanced model such as basic category and enhanced category model. Where as user posted question retrieved based multiple domain category and we concentrate all the domain answer especially on medical sites. Our scheme is able to produce promising performance as compared to the prevailing coding methods. More importantly, the whole process of our approach is unsupervised and holds potential to handle large-scale data. Local Mining Gives direct Answers and Global Learning is implemented as a Search Engine. Machine Learning improves system performance. The future work can focus on implementing much more domains, suggesting a nearby hospital to the user ,providing appointment with the expert etc.

V. REFERENCES

[1]. Guangyou Zhou and Jimmy Xiangji Huang, Modelling and learning distributed word representation with metadata for question retrival , in proceedings of IEEE Transactions on Knowledge and Data Engineering (Volume: 29, Issue: 6, June 1 2017).

[2]. X. Xue, J. Jeon, and W. B. Croft, Retrieval models for question and answer archives, in Proceedings of the SIGIR, 2008, pp. 475 482.

[3]. L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, Knowledge sharing and yahoo answers: Everyone knows something, in Proceedings of the WWW, 2008, pp. 665 674.

[4]. H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, Searching questions by identifying question topic and question focus, in Proceedings of ACL, 2008, pp. 156 164.

[5]. J.-T. Lee, S.-B. Kim, Y.-I. Song, and H.-C. Rim, Bridging lexical gaps between queries and questions on large online q&a collections with compact translation models, in Proceedings of the EMNLP, 2008, pp. 410 418.

[6]. D. Bernhard and I. Gurevych, Combining lexical semantic resources with question & answer archives for translation-based answer finding, in Proceedings of the ACL, 2009, pp. 728 736.

[7]. L. Cai, G. Zhou, K. Liu, and J. Zhao, Learning the latent topics for question retrieval in community QA, in Proc. 5th Int. Joint Conf. Natural Language Process., 2011, pp. 273 281..

[8]. K. Zhang, W. Wu, H. Wu, Z. Li, and M. Zhou, Question retrieval with high quality answers in community question answering, in Proceedings of the CIKM, 2014, pp. 371 380.

[9]. S. Robertson, S.Walker, S. Jones, M. Hancock-Beaulieu and M. Gatford, Okapi at TREC-3, in Proceedings of TREC, 1994, pp. 109 126.

[10]. G. Zhou, Y. Liu, F. Liu, D. Zeng, and J. Zhao, Improving question retrieval in community question answering using world knowledge, in Proc. 23rd Int. Joint Conf. Artif. Intell., 2013, pp. 2239 2245.

[11]. J. Jeon, W. B. Croft, and J. H. Lee, Finding similar questions in large question and answer archives, in Proceedings of the CIKM, 2005, pp.84 90 [12]X. Cao, G. Cong, B. Cui, C. S. Jensen, and Q. Yuan, "Approaches to exploring category information for question retrieval in community question- answer archives," ACM Trans. Inf. Syst., vol. 30, no. 2, 2012, Art. no. 7.

[12]. Z. Ji, F. Xu, and B. Wang, "A category-integrated language model for question retrieval in community question answering," in Proc. 8th Asia Inf. Retrieval Societies Conf., 2012, pp. 14-25.

[13]. A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. 49th Annual Meeting Assoc. Computer. Linguistics, 2011, pp. 142-150.

[14]. J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in Proc. 48th Annual. Meeting Assoc. Computer. Linguistics, 2010, pp. 384-394.

[15]. C. Xu, et al., "RC-NET: A general framework for incorporating knowledge into word representations," in Proc. 23rd ACM Int. Conf. Inf. Knowl. Manage., 2014, pp. 1219-1228.

[16]. M. Yu and M. Dredze, "Improving lexical embeddings with semantic knowledge," in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, 2014, pp. 545-550.

[17]. X. Zheng, H. Chen, and T. Xu, "Deep learning for Chinese word segmentation and POS tagging," in Proc. Conf. Empirical Methods Natural Language Process., 2013, pp. 647-657.

[18]. R. Collobert, J. Weston, L.Bottou, M.Karlen, K Kavukcuoglu,and P.P.Kuksa "Natural language processing (almost) from scratch," J. Mach. Learn. Res. , vol. 12, pp. 2493-2537, 2011.

[19]. M. Bansal, K. Gimpel, and K. Livescu, "Tailoring continuous word representations for dependency parsing," in Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, 2014, pp. 809-815.

[20]. R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive auto encoders for predicting sentiment distributions," in Proc. Conf. Empirical Methods Natural Language Process., 2011, pp. 151-161.