Information Retrieval for Medical Queries Using Data Mining

V. Sabapathi*, S. Sasikumar, M. Ganesh, J. Yuvaraj, P.R. Harishraj

Department of Computer Science and Engineering, Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Avadi, Tamilnadu

***Corresponding author: E-Mail: sabapathi2000@gmail.com**

ABSTRACT

There is a rapid improvement in web development in last 10years. So technologies have improved a lot. Our project is based on search engine concept for medical queries. User will have lot of queries in medical field. They won't get the answer what they actually need. Medical queries are more sensitive so we have to provide accurate answers. For that we are developing this search engine with the knowledge of Medical Health Representative (MHR).Large number of med-lib files are added. It contains pre inserted question and answers that are frequently asked by user. We are using local mining and global learning approaches. User questions will be sensed and analyzed to give most relevant answers that's what all search engines not have. Analyzing the key phrases using semantics includes NLP process, noun and verb and key concept identifier and lexical similarities etc. User will get the answers finally what they need with number of ranked list of answers that our server give based on the question user asked. Ranking based on the important keyword in the question.

KEY WORDS: Key concept detection, noun phrase extraction, query/question expansion, question retrieval.

1. INTRODUCTION

Group Question Answering (CQA) administrations have developed as prominent choices for online data procurement. As indicated by Google Trends, all the over three CQA administrations had more than 10 million hunts and visits in 2011. Over circumstances, an enormous measure of top notch question and reply (QA) sets has been collected as far reaching learning bases of human knowledge. It helps clients to look for exact information by acquiring right answers specifically, as opposed to perusing through substantial positioned arrangements of results. Subsequently to recover pertinent inquiries and their comparing answers turns into an essential undertaking for data procurement. Here we characterize address recovery in CQA benefits as an assignment in which new inquiries are utilized as questions to discover pertinent inquiries for which the appropriate responses are as of now accessible. For effortlessness and consistency, we utilize the expression "inquiry" to mean new inquiries postured by clients and "question" to indicate those addressed inquiries accessible in the CQA files.

Address recovery in CQA is not quite the same as general Web look Dissimilar to the web crawlers that arrival a not insignificant rundown of positioned records, address recovery gives back a few inquiries with conceivable answers specifically. Mean-while, address recovery can likewise be considered as a traditional Question Answering (QA) issue, yet the concentration of the QA errand is changed from answer extraction, answer coordinating and answer positioning to hunting down important inquiries with great prepared answers.

One noteworthy test is the word verboseness in the inquiries where imperative words might be encompassed by other extra words. As Park and Croft depicted, these extra words will probably befuddle the momentum web search tools instead of help them. For instance, in an inquiry: "Why are you less inclined to come down with a bug or influenza in spring summer and pre-winter than winter months?", a portion of the words are key terms for question recovery, for example, "contract a bug" and "winter months", some of them are corresponding words which are less imperative and may bring about perplexities for recovery models, for example, "spring summer and harvest time". The other real test is the word confuse between the inquiries and the hopeful inquiries for recovery.

2. MATERIALS AND METHODS

In this section the necessary techniques are provided in detail for the design, and the implementation of our search engine and integration system using data mining is explained briefly.

Key Concept Detection: The User asks 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.

Although, noun phrases have been verified to be reliable in key concept detection in information retrieval we also consider verb phrases. We observed that in CQA questions, verb phrases are important information carriers. Questions like "Why do people get colds more often in lower temperature?" and "Why are you less likely to catch a cold or flu n spring summer and autumn than winter months?" are two similar questions that share less common noun phrases, but their verb phrases are paraphrases. The above examples illustrate that verb phrases are as important as noun phrases in question retrieval. Hence, we use noun phrases and verb phrases extracted from the query question.

Query Expansion: A viable strategy to handle the word crisscross issue in data recovery is inquiry development proposed a connection based inquiry extension strategy to concentrate development terms from hunt log information. The separated terms were then coordinated into the first question in a unified positioning model to enhance the execution of Web inquiry. They then investigated the word relations in the entire corpus as worldwide data. At long last they joined the neighborhood and worldwide data as the development of inquiry for information recovery errand. Be that as it may, both of the two methodologies on question extension are completely in light of the measurable information and the semantic data of terms are disregard. The likeness of terms were registered by the separation in the Word Net tree structure. Be that as it may, the low scope, work serious and non-opportune nature makes these semantic lexicons hard to adjust to data recovery on UGC, for example, address recovery in CQA administrations. The interpreted question terms in this way can be viewed as the extension terms for inquiry.

In spite of the achievement of past work, writing respecting the idea level question development via consequently investigating the semantic data of idea from UGC information is still inadequate. In this paper, we propose a turn language interpretation approach, which makes up for the current rewording research in a reasonable granularity, to endeavor idea level summarizes as extensions for question recovery.

Question Retrieval: Question and reply with a measurable model. Riezler used a monolingual interpretation based recovery show for answer recovery. They presented sentence level rewording system to catch lexical similarities amongst inquiries and answers. Duan initially recognized question point and center by utilizing a tree cut strategy. They then proposed another dialect model to topture the connection between question theme and center for question recovery. Jeon analyzed four diverse recovery models, i.e., VSM, BM25, LM and interpretation display for question recovery in CQA files. Trial comes about uncover that the interpretation show beats alternate models. Consolidated the dialect model and interpretation model to an interpretation based dialect display and get better execution being referred to recovery. Taking after that, A syntactic tree coordinating model to finding comparable inquiries, and evil presence started that the model is hearty against linguistic blunders. The monolingual standard corpora, which are gathered from the Wiki Answer site, the definitions and shines of a similar term in different lexical semantic assets, to prepare the interpretation demonstrate for question recovery.

3. RESULTS

Key Concept Detection Results: To survey the adequacy of our approach on key idea identification, we use the SVMrank14 apparatus for idea positioning. The SVM rank model is chosen for two reasons. To start with, key idea identification is basically a positioning assignment. As we have exhibited in Section 3.1, once we get the idea positioning rundown, we can acquire the key ideas. Second, positioning methods are more appropriate than characterization techniques practically speaking as it not just thinks about the contrasts between ideas in KC and NKC, additionally looks at the distinctions among the ideas in KC.

We can see that: First, the baseline1 can be upgraded by the elements proposed in our approach. The reason might be that we not just catch the measurable information, for example, the report recurrence and Google n-gram, yet we likewise get the upsides of semantic butt-centric for example, reliance parsing and named element recognition, and outside learning base, for example, Wikipedia.

Second, our proposed positioning based model to key concept location (RbKCD) beats the order based models. The reason might be that the RbKCD not exclusively can catch the contrasts between positive example (key idea) and negative case (non-key idea), however can likewise catch the distinctions among positive occasions. This is steady with the outcome in our experiments, the best execution is accomplished when just a single key idea was included into the question recovery display.

Third, the proposed approach outflanks the base-line2 at both p@1 and MRR. This is on the grounds that that the baseline2 approach just model the unigram, bigram and unordered window terms. Nonetheless, the unigram and bigram are typically vague in sense. Our proposed approach catches the weights of ideas in question by utilizing the measurement and phonetic data. Besides, the expression structure can better speak to the free semantic.

We additionally break down the utility of different elements utilized as a part of our key idea location assignment as portrayed. In every cycle, we expel one single element from list of capabilities and leave alternate elements for preparing and forecast. We expect that the elements are free with each other, and the diminishing precision in this manner shows the contribution of the expelled highlight to the general exactness.

We take note of that the greater part of the above elements contribute pretty much to key idea identification errand. This is on the grounds that for the ranking undertaking, record recurrence of idea generally mirrors its factual dispersion in general dataset, and thus the lower the report recurrence of idea, the more critical it is. In the interim, we can infer that the top rank ideas will probably be the subjects in the given inquiries. In future work, we plan to consider these distinctions in components for further enhancing the performance of key idea identification.

Concept Paraphrase Generation Results:

Evaluation on Paraphrase Generation: As the bilingual parallel corpora are utilized for reword era in our proposed approach, we call it "Biling Pivot" for short. In the mean time, summarize era should likewise be possible from monolingual parallel corpora by utilizing monolingual interpretation display For examination, we execute the cutting edge technique for summarize era from monolingual parallel corpora in as our pattern, which is dealt with as a statistical machine interpretation issue that used a monotone phrasal decoder to create rewords in same importance. We call it "Monoling Trans" for short. For preparing, we utilize two informational index as the monolingual parallel corpora. In the first place is the comparable question combines in which are gathered by the clients' clicking of the comparative inquiries of the hunt inquiries in Wiki Answer benefit. Here, the similar address sets which are picked by clients.

Pivot Languages Analysis: We found that the most rate of para-expressions in all the 10 turn dialects are NP (thing phrase), trailed by the VP (verb state) rewords. It demonstrates that a large portion of the interpretations are NP and VP.

Thing phrases in German are set apart with cases, which shows themselves as various word endings at things, determiners and so forth.

We really get 10 turn dialects. Be that as it may, distinctive rotate dialects might not have a similar execution. To check this, we configuration to expel one dialect at any given moment and utilize whatever remains of nine rotate dialects for reword generation. We can then recognize the distinctive capacities for para-state era among these rotate dialects. The exploratory consequences of turn dialect investigation. We arbitrarily select 110 ideas as contribution to acquire the summaries for manual assessment.

Abundance of German may clarify the most astounding commitments of it on the summarizing the performance by utilizing it as the rotate dialect. In addition, with Danish dialect is evacuated, we get the most modest number of created rewords. Albeit each of the dialect asset is about a similar scale as far as sentence number, the sparsity of the vocabularies on each rotate approach are different, which may lead to the different performance on paraphrasing. According to the statistics by Koehn the Finnish vocabulary is about five times as big as English, due to the morphology. By checking the number of unique words on each language resource, we find that the Danish and Swedish corpora have the largest and smallest numbers of unique words respectively. Hence, we can deduce that the differences on the quantities of generating paraphrases may be cause by the different scales of vocabularies of each corpus.

Overall, we can also see that when any of the 10 pivot languages is removed, the corresponding performance decreases. It suggests that all of the 10 pivot languages are contributing to paraphrase generation.

Comparison Systems: To evaluate the proposed key concept paraphrase based question retrieval model, we compare with the following question retrieval models.

TLM: The translation based language model proposed by Xue which is the state-of-the-art question retrieval model which combines the translation model and the language model to estimate the parameters in ranking function.

STM: The syntactic tree matching model which is mainly based on a syntactic tree kernel function to compute the structure similarity of the query and candidate questions.

REL: The improved pseudo relevance feedback (PRF) model with new optimized term selection scheme

KCM: The key concept based retrieval model pro-posed which is the state-of-the-art model for key concept detection in verbose queries (baseline 4). It uses the AdaBoostM1 model to classify the key concept from non-key ones with multiple features.

Mono KCM: The key concept paraphrase based question retrieval model, where the paraphrases are obtained by using the monolingual based paraphrase generation approach.

PBTM: The phrase based translation model for question retrieval in CQA archives which is the first work to use machine translation probabilities to estimation the term similarity for question retrieval.

ETLM: The entity based translation language model for CQA question retrieval which is an extension of TLM by replacing the word translation to entity translation for ranking.

WKM: The world knowledge (WK) based question retrieval model which used the Wikipedia as an external resource to add the estimation of the term weights from Wikipedia space into the ranking function.

M-NET: The M-NET which is a state-of-the-art approach to CQA question retrieval using continuous word embedding, which added the meta-data (category information) of the questions to obtain the updated word embedding and Fish Vector is utilized to regularize the question length.

Para KCM: The proposed key concept paraphrase based question retrieval model in CQA archives.

Question Retrieval Results: We can conclude from KCM model outperforms TLM model. It indicates that the key concept based query refinement scheme is effective in question retrieval task. The reason is that TLM model employs IBM translation model 1 to capture the word translation probabilities. How-ever, as we described in Section 1, questions in CQA.

Performance Variation by Integrating Different IR Models: We also check the variation of the performance of question retrieval over different IR models that are integrated into the proposed question retrieval framework.

We can see that the performance of all the four models are boosted by being integrated into the proposed question retrieval framework. It again reveals that the paraphrase model is compatible with the existing IR models and contributes effective semantic connection among the key concepts in the query and the retrieved questions.

4. DISCUSSION AND CONCLUSION

In this paper, we proposed a key idea summarizing based way to deal with successfully handle the real issues of word verboseness and word confuse being referred to recovery by investigating the interpretations of rotate dialects. Promote, we extended inquiries with the produced summarizes for question recovery. The exploratory outcomes demonstrated that the key idea summarize based question recovery display outflanked the cutting edge models in the question recovery assignment.

Later on, we plan to create the idea para expressions to together assessing their probabilities on the multiple semantic assets. In the mean time, we will consider to receive the word or express implanting way to deal with investigate the phrasal summarizes because of its energy on measuring words or expressions similitudes utilizing the setting of monolingual asset. Furthermore, we plan to recognize the contrast of the POS on the idea rewords era by utilizing the assorted blends of turn dialects and genuine find their weights for various rotate dialects.

Enhancement:

- Separate Server Implementations for Local Mining and Global Learning
- Medical Net Library
- NLP Techniques
- PDF Searching using Indexing
- Artificial Intelligence
- Multiple language Translation

5. ACKNOWLEDGEMENT

This paper is inspired from “Creation and Amalgamation for Health Information Exchange using Cloud Computing System based on Clinical Document Architecture.” We would like to express our deepest gratitude and appreciation to our guide assistant Prof Mrs. Merry Ida A. We are thankful to her for the excellent guidance and encouragement she has provided us during the course of this project.

REFERENCES

Allan J, Callan J.P, Croft W.B, Ballesteros L, Broglio J, Xu J and Shu H, “INQUERY at TREC-5, in Proc. TREC, 1996, pp. 119– 132.

Callan J.P, Croft W.B and Broglio J, TREC and tipster experi-ments with INQUERY,” in Proc. Inf. Process. Manage, 1995, 327–343.

Collins K, Thompson and Callan J, Query expansion using ran-dom walk models, in Proc. 14th ACM Int. Conf. Inf. Knowl. Man-age., 2005, 704–711.

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

Park J.H and Croft W.B, Query term ranking based on dependency parsing of verbose queries, in Proc. ACL, 2010, 829–830.

Resnik P and Smith N.A, The web as a parallel corpus, Com-put. Linguist, 29 (3), 2003, 349–380.

Turney P.D, Learning algorithms for key phrase extraction, Inf. Retr., 2, 2000, 303–336.

Xu J and Croft W.B, Query expansion using local and global document analysis, in Proc. 19th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1996, 4–11.

Xue X, Jeon J and Croft W.B, Retrieval models for question and answer archives, in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, 475–482.

Zhou G, Cai L, Zhao J and Liu K, Phrase-based translation model for question retrieval in community question answer archives," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Lang. Technol, 1, 2011, 653–662.