International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Expert System for Healthseekers by Using Local Mining and Global Learning

Mrs.S.Suganya Devi[1], Kavitha.U[2], Krithika.K[3], Leelapriskila.R[4]
[1]Assistant Professor, [2,3,4]B.E.(Final Year) , Dept. Of Computer Science And Engineering
Alpha College of Engineering.

Abstract*—In this new era on the popularity of internet has made the people to examine their health status before they knock the door of the doctors. This motivated us to propose a system, which helps the healthcare seekers by bridging the vocabulary gap between health seekers and providers. We proposed a novel scheme for retrieving the answers forthwith by the system using following tactics namely Local Mining, Global Learning .In local mining the posted query undergoes Natural language processing which entail of three processes Noun phrase extractor, stop word remover and Spell checker. As corollary a key word is extracted then it is normalized into medical terminology as the result of local mining , a corpus aware terminology generated automatically normalized medical terms are indexed using Invert indexing to ensure the immediate retrieval of result. The approach of global learning is used to enhance the local mining through identifying the missing medical terms from resource PDF using lexical similarities and analysis it to derive the conclusion In case of lacking exact information in our system then the query is forward to experts thus making our system knowledge to answered all the queries posted by health seekers, which overcome delayed cross system operability and the inter usability.

Key Terms: Medical Terminologies Assignment, Porter-Stemmer, Question & Answering Blog, Natural Language Processing, Inverted Indexing.

I. INTRODUCTION

Data mining is knowledge discovered from data. The data mining processes include expressing a term, collecting data, performing preprocessing, estimating the model, and clarifying the model and draw the conclusions. It is the process of analyzing and summarizing data from different perspectives and converting it into useful information [1]. Thus Data mining holds great conceivable for the healthcare industry to enable health systems to systematically use data and analytics to identify inability and best practices that improve care and reduce costs. But due to the complexity of healthcare and a slower rate of technology adoption, our industry lags behind these others in implementing effective data mining and analytic strategies [2]. This motivated us to propose a system, which helps the health seekers by bridging the vocabulary gap between health seekers and providers.

We proposed a novel scheme of retrieving the answers instantly by the system with the help of two approaches called Local Mining and Global Learning. For health seekers, these systems provide nearly correct and trusted answers especially for complex and refined problems. An enormous number of medical reports have been assigned in their repositories, and in most consequences, users may directly locate better answers by surfing from these record archives, rather than waiting for the experts responses or browsing through a list of relevant documents from the Websites. NLP process is used because Users with different backgrounds do not use to share the same vocabulary. The requirements are written by seekers in narrative language. The same question may be described in different ways by other two individual health seekers. From the other side, the answers provided by the researchers and experts may contain acronyms with multiple possible meanings, and non-standardized terms.

The approaches of local mining carry the query posted by the health seekers undergoes Natural Language Processing(NLP) which entail of three processes POS tagger, Stemmer and WordNet. POS tagger(Parts-Of-Speech tagger)will extract the Noun phrase from the given query, Stemmer works from the POS tagger outcomes which will remove the stopping word like –es/ -ing/ -s/ and so on. Finally it checks the spelling using a external knowledge of dictionary of WordNet which results in identifying the medical words in the posted query. Then the medical word is normalized with the medical concepts using MediNet for example, if patient posted as “itching” as medical word then it is normalized into medical concept called “Tenia corpus”, thus automatically creating a corpus aware terminology. The approach of global learning is used to enhance the local mining through identifying the missing medical terms from resource PDF using lexical similarities. The plenty number of resource likes pdf, books, files are allowed to stuffed inside the database by creating an algorithm like inverted indexed which is used to index an items inside the database for easy retrieval of an output. In case of lacking exact information in our system then the query is forward to experts thus making our

system knowledge to answered all the queries posted by health seekers, which overcome delayed cross system operability and the inter usability. Thus the system will generate to produce the replies for the user's query instantly. The reminders and Structures as follows: Section 2 we discussed briefly about our related works. Section 3 contains Existing System. Section 4 we collaborated and explained each module. We conclude our work in Section 5.

II. RELATED WORKS

Over the year, with potential and inclined growth of web world to search information on internet plays a vital role in every day life. Due to the emerging information retrieval technologies transmute the ways of thinking nowadays the people are like to explore health oriented information via internet. A National survey conducted by the Pew Research Center in Jan 2013,where they reported that one in three American adults have gone online to figure out their medical conditions in the past 12 months the report time. As a outcome of many question and answering blogs and medical forums like haodf.com, healthtap.com, webMD4.But the knot in these are unstructured queries, sophisticated syntactic, and contextual processing induce information. To overcome this problem the existing approaches are Rule-based and Machine learning approaches.

In Rule-based techniques play major role in medical terminology assignment. By using morphological, semantic and pragmatic aspects of natural language. They discover and construct effective rule for medical terminology assignment [3]. Hersh & David develop a system, named [4] Semantic and Probabilistic Heuristic Information Retrieval Environment [SAPPHIRE], which addressed the problem of Indexing, Retrieval and Evaluation through Relevance Ranking, "Meta thesaurus" from National Library of Medicine's [NLM].A System named as Index Finder [5], which was posterior to sapphire .The key idea of index finder to implement the medical digital library by proposed new algorithm for generating all valid UMLS concepts. By giving a set of word as the input text then using syntactic and semantic filters they filtered out the irrelevant concepts. Generally this approach concentrated on hospital generated data or resources provide by doctors and medical practitioner. This method is suitable for small terminologies. Compare to this kind of data health forums generated data are more colloquial, in terms of inconsistency and ambiguity.

In Machine learning approaches construct the inference model from medical data with known annotations and then apply the trained models [6].Maria Taboada et al. proposed an automated approach for mapping external terminologies to the UMLS by using validation of lexical alignment, EMTREE and UMLS Meta thesaurus [7].Similar to this scheme, Pakhomov et al. [8] attempted to improve coding performance by combing machine learning approach and Auto coder. It uses Naive Bays integration is loosely coupled and learning model cannot incorporate heterogeneous cues, which is not a better choice for medical forums and Q&A blogs.

III. EXISTING WORKS

Most of the existing work focused on hospital generated health data or health provider released data by utilizing either isolated or loosely coupled rule-based and machine learning approach but it is worth notice that there already exist several efforts dedicated to automatically mapping medical datasets to terminologies using UMLS. Further most of the previous work simply utilizes the external medical dictionary to code the medical datasets rather than considering the corpus aware terminologies because of this external knowledge may regnant of missing key terms and inappropriate terminologies . It may cause morass situation among the health seekers constructing corpus aware terminologies vocabulary to prune the irrelevant terminologies of specific records. On the other hand data generated by healthcare forums and medical sites may contain forums and abbreviations which may contain multiple possible meaning and no standardized terms. Recently, some sites have encouraged experts to annotate the medical datasets with medical concepts. However, tags used often vary and medical concepts may not be medical terminology. For an example ,"Mental disorder "and "Brain fog" are employed by different doctors to refer to the same medical terms it shown inconsistency of community generated data. Due to the inconsistency between search term and accumulated medical records therefore automatically coding the medical datasets with standardized terminologies is highly desired.

IV. PROPOSED SYSTEM

To the best of our knowledge, this is the first work on automatically coding the medical forums and medical Q&A sites generated health data, which are more complex, inconsistent and ambiguous, compared to the hospital or clinical generated record. We proposed the novel approach to bridge the gap between the health seekers and providers by separate server implementation which utilizes local mining and global learning approaches. In local mining the query posted by the health seekers are supposed examine through Natural Language Processing. The term extracted from the NLP are normalized to get Medical term to bridge the vocabulary gap between the health seekers and providers. On the other hand Global learning helps in ameliorate local mining results

Volume 4 Issue IV, April 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

by graph-based approach, which combine missing key concepts and keeping off irrelevant terminologies.

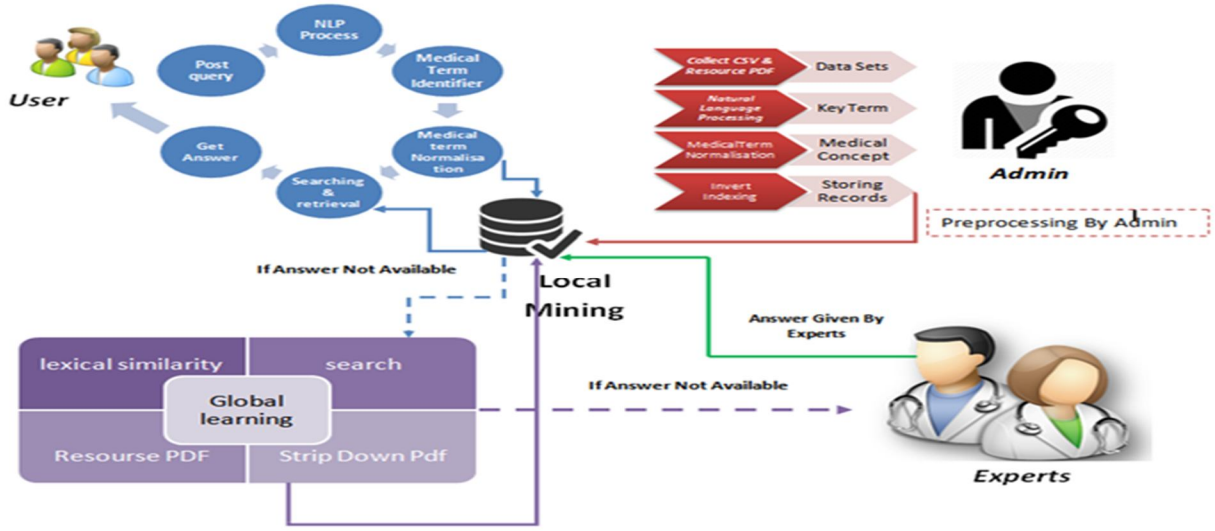


Fig. 1 System Architecture

There are 4 modules in the project:

Q and A Blogs.

Local Mining.

Global learning.

Experts Review and Answer.

A. Q & A Blog

In existing system we build effective question and answering website which could give instant answer to any query posted by the user. The extract answer retrieved by using inert indexing algorithm and the results are prioritize using bubble sort algorithm. The process behind this query posted by user are processed by NLP then extract key word taken into normalization to get medical concept we retrieve the extract information retrieved by using inverted indexing.

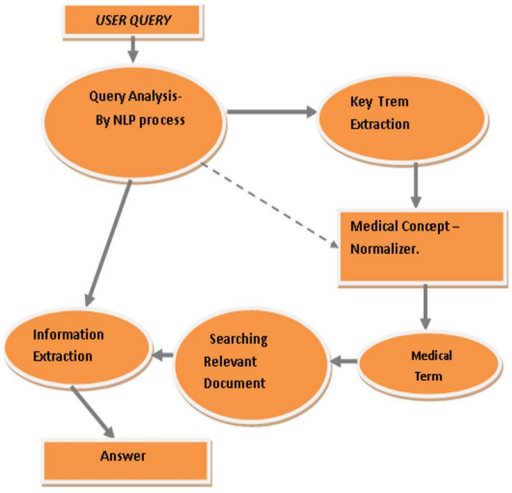


Fig. 2 Answer Retrieval System

Volume 4 Issue IV, April 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

B. Local Mining

In local mining we proposed a tri-stage framework, which are NLP processing, Medical Concept Identifier And Medical Term Normalization.

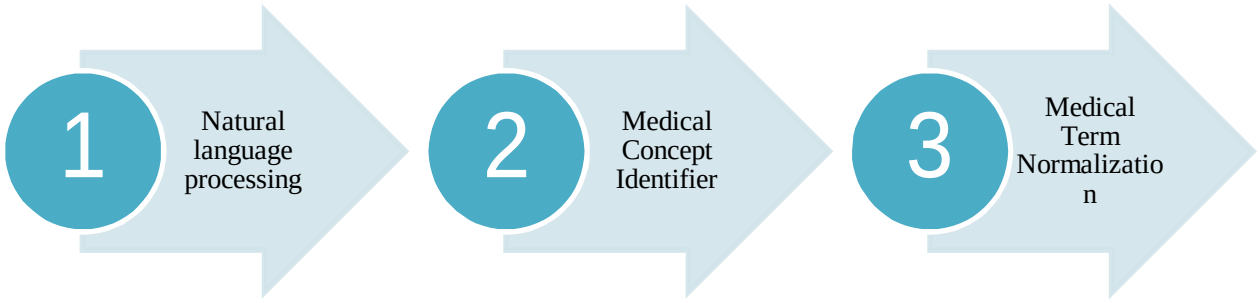


Fig. 3 Tri-Stage Local Mining

C. Natural Language Processing

The NLP process used to enhance the vocabulary gap by processing the query posted in natural language & making the system to understand the natural language. 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 then Spell Checker used to get the proper Spelling for the obtained keywords.

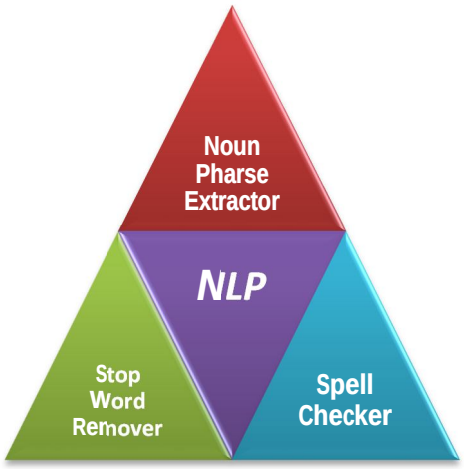


Fig. 4 Three Stage Process of NLP

1) *Noun – Phrase Extractor:* The raw queries posted by user are categorized into part of speech. The give query sentence examines and attaches each wording a sentence with a suitable tag from a given set of tags by Standard POS tagger. The pattern formulated as follows:

(Adjective / Noun)(Noun Preposition)?(Adjective/ Noun)* Noun.*

A sequence of tags matching this pattern ensure that corresponding word make up a noun phrase extractor.

2) *Stop Word Remover:* The stop words like -ing , -ed, – sses are eliminated from the the Noun which is extracted from Noun – Phrase Extractor. By using porter stemmer [9] algorithm. It utilizes suffix stripping. Porter stemmer algorithm, which help in the reduction of total number of terms, size and complexity of the documents.

Volume 4 Issue IV, April 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

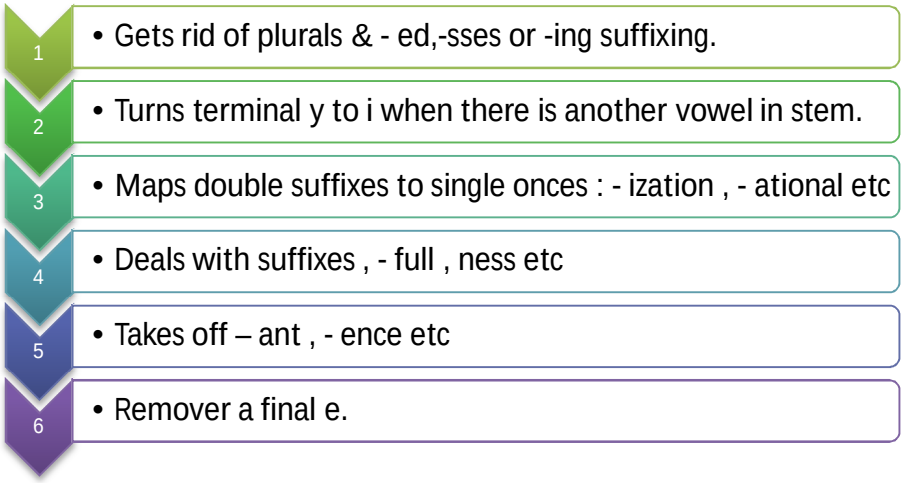


Fig. 5 Steps in Porter Stemmer

3) *Spell Checker:* Spell checking is done by using wordNet . It is a lexical knowledge based on conceptual look up . It organizes lexical information in terms of meaning of words rather its formation. It user lexical matrix ensure the synonyms of a word & there by checking the spelling of root word obtained from the stemmer process.

D. Medical Concept Detection

In this step we aim to separate medical concepts noun from other general noun phrases. The assumption we set to get medical concept from general noun is, the terms which are relevant to medical domain occur more in medical domain and general noun (i.e) non-medical ones. In order to get this we use the concept entropy impurity to measure the relevance in the domain. For a term t, its CEI is computed as:

$$CEI(t) = -\sum_{i=1}^{2} P(Di\backslash t) log P(Di\backslash t)$$

Where D_1 and D_2 represents medical term and general term respectively. $P(D_i \mid t)$ denotes the probability of term t is related to a domain D,

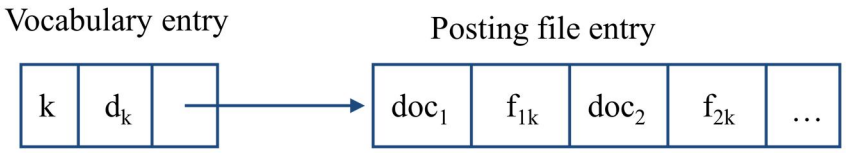
$$P(D_i \mid t) = count(t, D_i)/count(t)$$

E. Medical Concept Normalization

Still there is no assurance of Standardized terminologies even after medical terms are defined by domain specific noun phrases. In order to obtain standardized terminologies we need to normalize the term obtained from the medical term detection step. By using SNOMED CT (Systematic Nomenclature of Medicine Clinical Terms) which is an organized lists of a wide variety of clinical terminology defined with unique codes Perhaps the most comprehensive clinical terminology in the world. SNOMED CT – is better suited for capturing relevant data during an encounter Allows the user to capture the various aspects associated with a disorder (Post Coordination) This encourages the user to capture associated information like Severity, Body part affected, Cause (force or substance), laterality (viz., left or right), Morphology (form) in structured form. Usually, SNOMED CT is considered a good way to enter the medical information. The terminologies and their descriptions in SNOMED CT are indexed first then we search each medical terms against indexed SNOMED CT. For indexing the SNOMED CT we use Inverted Indexing Techniques.

1) *Invert Indexing:* In this algorithm, the user documents are indexed through some preprocessing steps. First they need the tokenize that will turn up the user documents into sequence of words, namely Tokens. By using tokens stream they will relate to their language through the linguistic modules and put them into some kind of Canonical form. Finally those modified tokens which will fed into the indexer then it will store into disk. Invert Indexing helps us to retrieve the information we needed efficiently.

Volume 4 Issue IV, April 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)



d_k : document frequency of term k
doc_i : i-th document that contains term k
f_{ik} : term frequency of term k in document i

F. Global Learning

The aim of global learning is to learn appropriate terminologies from the global terminology space T to annotate each medical terms q in Q [10]. Among existing machine learning methods, graph-based learning achieves promising performance. In this work, we also explore the graph-based learning model to accomplish our terminology selection task, and expect this model is able to simultaneously consider various heterogeneous cues, including the medical record content analysis, terminology-sharing networks, and the inter-expert as well as inter-terminology relationships. We will first introduce relationship identification and then we detail how to use our proposed model to link the underlying connected medical records. Next, we present the optimal solution for our learning model followed by the label bias estimation. And we discuss the scalability of our method.

1) *Relationship Identification:* The inter-terminology and inter-expert relationships are not intuitively seen or implied from medical records. We thus call them as implicit relationships. This subsection aims to introduce how to discover these kinds of relationships.
2) *Inter-Expert Relationship:* The inter-expert relationships will be viewed stronger if the experts are professionals in the same or related specific medical areas. This is reflected by their historical data, i.e., the number of questions they have co-answered.
3) *Experts Review And Answers:* In this final module, Experts are answering the query in case of unavailability of exact answer in both local mining and global learning approaches. The Answer given by the Experts are preprocessed and loaded into local mining datasets.

V. CONCLUSION & FUTURE WORKS

This paper presents a medical terminology assignment scheme to bridge the vocabulary gap between health seekers and healthcare knowledge. The scheme comprises of two components, local mining and global learning. The former establishes a tri-stage framework to locally code each medical record. However, the local mining approach may suffer from information loss and low precision, which are caused by the absence of key medical concepts and the presence of the irrelevant medical concepts. This motivates us to propose a global learning approach to compensate for the insufficiency of local coding approach. The second component collaboratively learns and propagates terminologies among underlying connected medical records. It enables the integration of heterogeneous information. Extensive evaluations on a real-world dataset demonstrate that 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. In the future, we will investigate how to flexibly organize the unstructured medical content into user needs-aware ontology by leveraging recommended medical terminologies.

REFERENCES

[1] Data Mining for Medical Systems: A Review Muhamad Hariz Muhamad Adnan,Wahidah Husain, Nur'Aini Abdul Rashid School of Computer Sciences University Sains Malaysia 11800 USM, Penang, Malaysia
[2] What is Data Mining in Healthcare? David Crockett Ph.D., Research & Predictive Analytics, Sr. Director, Brian Eliason, Vice President of Technical Operations.
[3] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X.Guo, “Fast tagging of medical terms in legal text,” in Proc. Int. Conf. Artif. Intell. Law, 2007, pp. 253–260.
[4] W. R. Hersh and H. David, “Information retrieval in medicine: The saphire experience,” J. Amer. Soc. Inf. Sci., vol. 46, no. 10,pp. 743–747, 1995.
[5] Q. Zhou, W. W. Chu, C. Morioka, G. H. Leazer, and H. Kangarloo, “Indexfinder: A method of extracting key concepts from clinical texts for indexing,” in Proc. AMIA Annu. Symp., 2003, pp. 763–767.
[6] E. J. M. Laur_ıa and A. D. March, “Combining Bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis,” J. Data Inf. Quart., vol. 2, no. 3, p. 13, 2011.

[7] An Automated Approach to Mapping External Terminologies to the UMLS Maria Taboada,Rosario Lal´ın, and Diego Mart´ınez.
[8] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," J. Amer.Med. Inf. Assoc., vol. 13, no. 5, pp. 516–525, 2006.
[9] The Porter Stemmer Algorithm, Daniel Waegel CISC889 / Fall 2011.
[10] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas, "Image retrieval via probabilistic hyper graph ranking," in Proc. IEEE Conf. Compute. Vis. Pattern Recognition., 2010, pp. 3376–3383.