EFFECTIVE CONSUMERS VOCABULARY MAPPING TO CLINICAL TERMINOLOGY

Priyanka R , PG Scholar,
Dept of Computer Science and Engineering,
Government College of Technology,
Coimbatore.
Mail id : priyaprgood@gmail.com

Dr. J.C.Miraclin Joyce Pamila,
Assistant Professor(Sr.Grade),
Dept of Computer Science and Engineering,
Government College of Technology,
Coimbatore.

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. Proposed novel scheme retrieves the answers forthwith by the system using following tactics namely Local Mining and 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 keyword 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 analyses it to derive the conclusion. In case of lagging in exact information, the query is forwarded to experts thus making our system knowledge to answer all the queries posted by health seekers, which overcome delayed cross system operability and the inter usability.

Keywords – Medical Terminologies Assignment, Porter-Stemmer, Question & Answering Blog, Natural Language Processing, Local and global mining.

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

Nie,M. Akbari, T. Li, and T.-S. Chua [1] have proposed a joint local-global approach for medical terminology assignment .In community-based health services, vocabulary gap between Healthseekers and community generated knowledge has hindered data access. To bridge this gap, this paper presents a scheme to label question answer (QA) Pairs by jointly utilizing local mining and global learning approaches. Local mining attempts to label individual QA pair by independently extracting medical concepts from the QA pair itself and mapping them to authenticated terminologies. However, it may suffer from information loss and lower precision, which are caused by the absence of key medical concepts and presence of irrelevant medical concepts. Global learning, on the other hand, works towards enhancing the local mining via collaboratively discovering missing key terminologies and keeping off the irrelevant terminologies by analyzing the social neighbors. Practically, this unsupervised scheme holds potential to large-scale data.

G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt [2], have proposed “Exploiting medical hierarchies for concept-based information retrieval”. Search technologies are critical to enable clinical staff to rapidly and effectively access patient information contained in free-text medical records. Medical search is challenging as terms in the query are often general but those in relevant documents are very specific, leading to granularity mismatch. In this paper we propose to tackle granularity mismatch by exploiting subsumption relationships defined in formal medical domain knowledge resources. In symbolic reasoning, a subsumption (or `is-a') relationship is a parent-child relationship where one concept is a subset of another concept. Subsumed concepts are included in the retrieval function. In addition, we investigate a number of initial methods for combining weights of query concepts and those of subsumed concepts. Subsumption relationships were found to provide strong indication of relevant information; their inclusion in retrieval functions yields performance improvements.

Hina, E. Atwell, and O. Johnson [7] have proposed “Semantic tagging of medical narratives with top level concepts from SNOMED CT healthcare data standard”. Medical narratives written by clinicians constitute critical information in healthcare domain and are required to be correct with respect to contextual meaning. SNOMED CT (Systematized Nomenclature of Medicine -- Clinical Terms) is a standardized reference terminology that consists of 390023 SNOMED CT concepts with SNOMED CT codes. This paper describes the extraction of SNOMED CT concepts from free text discharge summary reports. For the evaluation of the medical concepts, we used 300 discharge summaries corpus provided by University of Pittsburgh Medical Centre, and compared it with the SNOMED CT concept file which is a preprocessed and cleaned file listing SNOMED CT concepts. In this paper, we present ongoing research on SNOMED CT concept extraction from discharge summaries using natural language processing and introducing SNOMED CT core concepts as a single gazetteer list for concept extraction. Out of 390023 concepts, 21563 concepts were found in the test set of discharge summaries from the SNOMED CT core concepts gazetteer list.

Patrick, Y. Wang, and P. Budd [4], have proposed “An automated system for conversion of clinical notes into snomed clinical terminology”, The automatic conversion of free text into a medical ontology can allow computational access to important information currently locked within clinical notes and patient reports. This system introduces a new method for automatically identifying medical concepts from the SNOMED Clinical Terminology in free text in near real time. The system presented consists of 3 modules; an Augmented Lexicon, term compositor and negation detector. The Augmented Lexicon indexes the SNOMED-CT terms, the term compositor finds qualification relationships between concepts and the negation detector identifies negative concepts. The system delivers the services through a variety of interface including direct programming access and web-based access. It is currently in use in a hospital environment to capture patient data response with SNOMED-CT codes in real time at the point of care.

R. L. Cilibrasi and P. M. B. Vitanyi[9], have proposed “The google similarity distance. Words and phrases acquire meaning from the way they are used in society, from their relative semantics to other words and phrases. For computers the equivalent of ‘society’ is ‘database,’ and the equivalent of ‘use’ is ‘way to search the Database.’ We present a new theory of similarity between words and phrases based on information distance and Kolmogorov complexity. To fix thoughts we use the world-wide-web as database, and Google as search engine. The method is also applicable to other search engines and databases. This theory is then applied to construct a method to automatically extract similarity, the Google similarity distance, of words and phrases from the world-wide-web using Google page counts. The world-wide-web is the largest database on earth, and the context information entered by millions of independent users averages out to provide automatic semantics of useful quality.

III. EXISTING SYSTEM

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.

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

In this work, we define medical concepts as medical domain-specific noun phrases, and medical terminologies as authenticated phrases by well-known organizations that are used to accurately describe the human body and associated components, conditions and processes in a science-based manner.

For example, "heart attack" and "myocardial disorder" are employed by different doctors to refer to the same medical diagnosis. It was shown that the inconsistency of community generated health data greatly hindered the cross-resource data exchange, management and integrity. Therefore, automatic coding of the QA pairs with standardized terminologies is highly desired. It leads to a consistent interoperable way of indexing, storing and aggregating across specialties and sites. In addition, it facilitates QA pair retrieval via bridging the vocabulary gap between the queries and archives by coding the new queries with the standardized terminologies.

Constructing a corpus-aware terminology vocabulary to prune the irrelevant terminologies of specific dataset and narrow down the candidates is the tough issue we are facing. In addition, the varieties of heterogeneous cues were often not adequately exploited simultaneously. Therefore, a robust integrated framework to draw the strengths from various resources and models is still expected.

It is noteworthy that most previous efforts, including our local approach, attempted to map the QA pairs directly to the entries in external dictionaries without any pruning. This approach often presents problems since the external dictionaries usually cover relatively comprehensive terminologies and are far beyond the vocabulary scope of the given corpus.

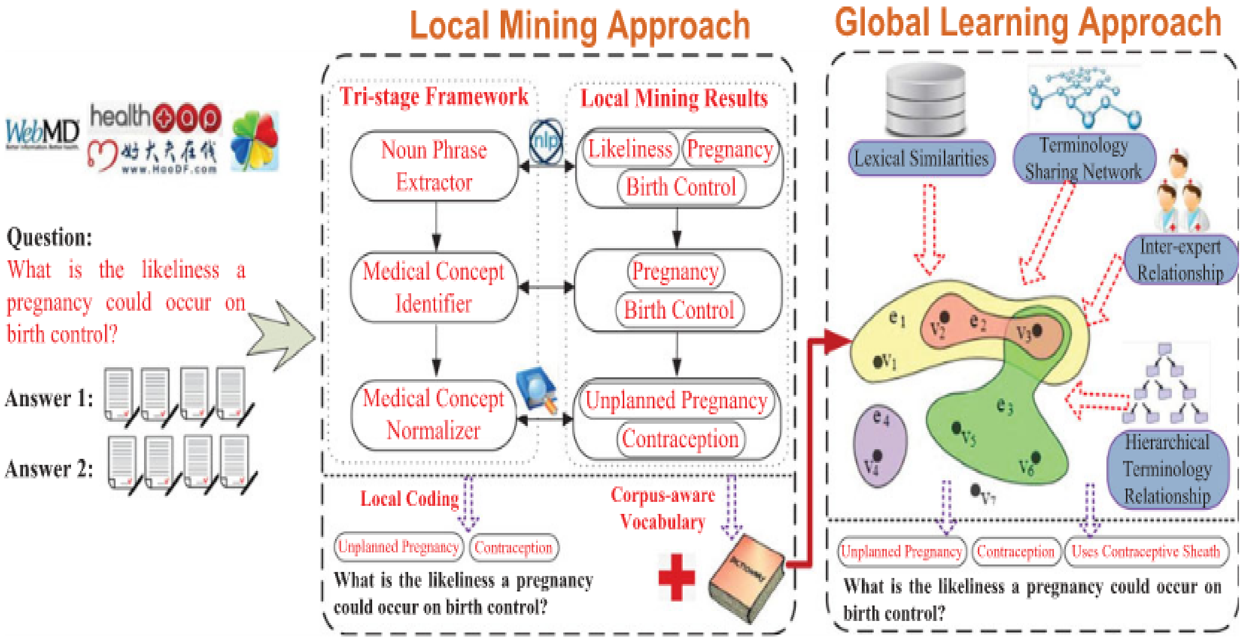


Figure 1. Proposed System Diagram

There are 4 modules in the project:

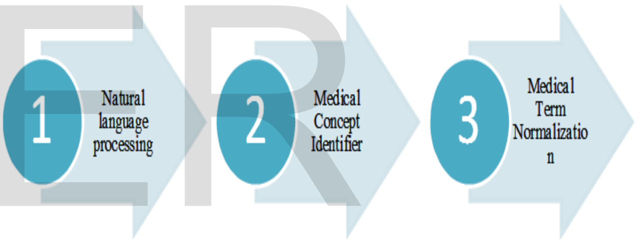
- Q and A Blogs.
- Local Mining.
- Global learning.
- Experts Review and Answer

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 process behind this query posted by user are processed by NLP then extract key word taken into normalization to get medical concept.

B. LOCAL MINING

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


1
Natural language processing
2
Medical Concept Identifier
3
Medical Term Normalizatio n

. Figure 2. Local Mining Steps

C. NATURAL LANGUAGE PROCESSING

The NLP process is 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 several steps of which Parts Of Speech Tagging (POST) results in Phrases and Nouns Extraction. The keywords thus extracted is subjected 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.

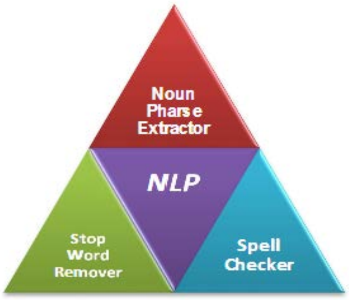


Figure 3. NLP Process

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)(NounPreposition)?*
(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. Porter stemmer [9] algorithm utilizes suffix stripping. Porter stemmer algorithm helps in the reduction of total number of terms, size and complexity of the documents.

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 uses 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 D1 and D2 represents medical term and general term respectively. P(Di│t) denotes the probability of term t is related to a domain D,

$$P(D_i | t) = count(t,D_i)/count(t)$$

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. It 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.

F. GLOBAL MINING

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.

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.

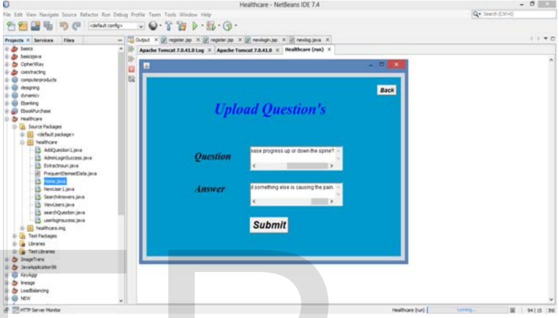
3)EXPERTS REVIEW AND ANSWERS:

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. IMPLEMENTATION

A. QUESTION ANSWER BLOG :

User interface page is created using Java Apache Tomcat Server. This page is used to ask questions for patients to doctors about their health. So, patients can seeks information and get instant answers for their query.

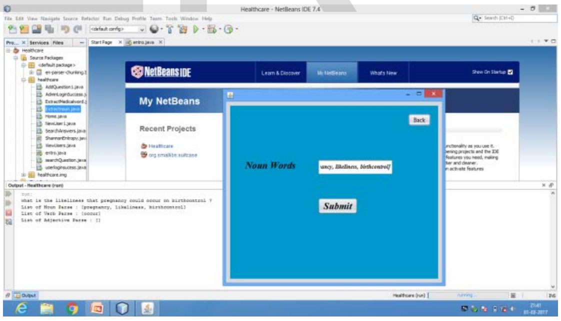


Question: ease progress up or down the spine?

Answer: it something else is causing the pain.

Figure 4. User Interface Page

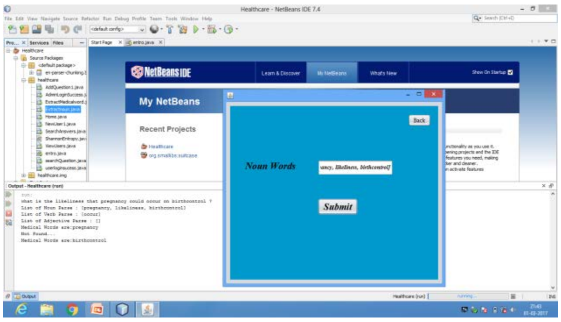
B. NATURAL LANGUAGE PROCESSING:



Noun Words: ancy, likeliness, birthcontrol]

```
run:
What is the likeliness that pregnancy could occur on birthcontrol ?
List of Noun Parse : [pregnancy, likeliness, birthcontrol]
List of Verb Parse : [occur]
List of Adjective Parse : []
```

Figure 5. Noun Phrase Extraction



Noun Words: ancy, likeliness, birthcontrol]

```
run:
What is the likeliness that pregnancy could occur on birthcontrol ?
List of Noun Parse : [pregnancy, likeliness, birthcontrol]
List of Verb Parse : [occur]
List of Adjective Parse : []
Medical Words are:pregnancy
Not Found...
Medical Words are:birthcontrol
```

Figure 6. Medical Words identification Using entropy

VI. CONCLUSION

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.

REFERENCES

[1] L. Nie, M. Akbari, T. Li, and T.-S. Chua, "A joint local-global approach for medical terminology assignment," in Proc. Int. ACM.

[2] Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt, "Exploiting medical hierarchies for concept-based information retrieval," in Proc. Australasian Document Comput. Sym. 2012, pp. 111–114

[3] Yan, G. Fung, J. G. Dy, and R. Rosales, "Medical coding classification by leveraging inter-code relationships," in Proc. ACMSIGKDD Int. Conf. Knowl. Discov. Data Mining, 2012, pp. 193–202.

[4] S. Hina, E. Atwell, and O. Johnson, "Semantic tagging of medical narratives with top level concepts from SNOMED CT healthcaredata standard,"Int. J. Intell. Comput. Res., vol. 2, pp. 204–210, 2010.

[5] H. Suominen, F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S.Salanter, and T. Salakoski, "Machine learning to automate theassignment of diagnosis codes to free-text radiology reports: Amethod description," inProc. ICMLWorkshop Mach. Learn.Health-Care Appl.,2008.

[6] Sigurbjornsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in Proc. 17th Int. Conf. World Wide Web, 2008, pp. 327–336

[7] S. Hina, E. Atwell, and O. Johnson, "Semantic tagging of medical narratives with top level concepts from SNOMED CT healthcaredata standard,"Int. J. Intell. Comput. Res., vol. 2, pp. 204–210, 2010.

[8] Patrick, Y. Wang, and P. Budd, "An automated system for conversion of clinical notes into snomed clinical terminology," in Proc. 5th Australasian Symp. ACSW Frontiers , 2007, pp. 219–226.

[9] R. L. Cilibrasi and P. M. B. Vitanyi, "The google similarity distance,"IEEE Trans. Knowl. Data Eng. , vol. 19, no. 3, pp. 370–383, Mar. 2007.