Survey on Data Mining Approach for Instant Medical Assistance

Prittha Tikariha[1], Prashant Ricchariya[2]
[1, 2] Department of Computer Science and Engineering
[1, 2] Chhatrapati Shivaji Institute of Technology,Durg,Chhatisgarh(India)

Abstract- *Data Mining helps to analyze the large amount of unused data from different prospect and summarize it into useful information. Data mining provides a great potential for the healthcare services to enable health systems to systematically use data. There is a relentless development in the measure of electronic health records (EHRs) being gathered by healthcare offices and using health web portal. In Healthcare space, there is a wide gap common between health seekers and medical experts. This vocabulary gap is because of the presence of ambiguity in the community language in which the clients post their questions in online medicinal applications which is informal, as far as inconsistency, complexity and ambiguity is considered. This brings big challenges for data access and analytics. To overcome this, there is a plan which has two components. First is local mining which extract the medical concept form the queries posted by the health seekers and then normalize it to authenticated terminologies. Second is global learning which works towards enhancing the local medical coding by discovering missing key terminologies and eliminating the irrelevant terminologies by analyzing the social neighbors.*

Keywords- local minning; global minng; community based health care

I. INTRODUCTION

According to a survey conducted by the Pew Research Center1 in Jan 2013, one in three people in America opt for online health advice rather than consulting the doctor to know about their medical condition. There are many online health care web portal are available like HealthTap, HaoDF and WebMD, where the health seeker can post there query regarding any medical issue and the health experts can analyses the query and respond accordingly. These applications proved to be very interesting and beneficial to both health seekers as well as for health experts. For health seeker, they can get an instant answer from authenticated and renowned doctors, where the doctors can improve their knowledge in different fields by connecting to the many different expert worldwide .Over a time a large amount of medical data has been accumulated in the repository of such web portal. In many cases, health seeker can directly look for a satisfying answer by searching in those repositories or just by browsing internet for the related document.

But the unavoidable issue regarding these web portals is that they are community based, the queries submitted by the health seekers are in the community language and it is not understandable and is not useful to extract information from that data. Users from many different backgrounds do share the same vocabulary. It is possible that the same question is asked in two different ways by two individuals. Similarly the same query can be answered by the doctors in different ways. It is possible that the answers provided by the doctors contain acronyms with different meanings or may contain some non-standardized word which may not be understandable by the common people. Thus it results in inefficient management of data and also it will be a great hindered to reuse the existing data due to the incompatibility between the existing medical data and the keyword provided for searching. This is making the health services data rich but information poor. This requires a good mechanism to code the medical record accumulated into a standard medical concept. This survey paper summarize the various method already proposed to code the medical record into standard medical terminology. Most of the work has been done in the hospital generated data only. Compared to the hospital generated data the community generated data are more informal and complex to code into standard medical terminology.

II. LITERATURE SURVEY

J. Patrick et al., [1] have introduced a system which can automatically identify the medical concept from the free text using SNOMED Clinical Terminology. This will provide a very good platform to access the hidden information from the clinical notes and the patients report. SNOMED is the most thorough reference for medical terminology in the world which is helpful for analysis and retrieval of clinical data. The authors have applied Token Matcher algorithm to map the clinical notes to the SNOMED CT terminology. Initially the clinical text tokens are matched with the SNOMED CT token and then broken up into chunks. Then the algorithm is applied on these chunks for the identification of negations and

qualification. They have implemented a qualification identifier and negation detector to recognize the composite term and the negative concept, which provides an effective way for information retrieval.

Mi-Young Kim et al., [2] developed an information retrieval system which finds the medical name from the UMLS meta-thesaurus. The system is based on domain specific term frequency and adaptive ranking. Domain specific term frequency helps to eliminate the negative terms and adaptive ranking determine the appropriate documents for each sentence. They consider the medical sentence as the input query and UMLS ontology entry as the document and apply the language model based method for information retrieval. To identify the term frequency they have used the document specific term frequency as the current query cannot determine the specificity of each term so it was better to extend it to medical domain specific corpus. As medical domain specific document is used it is possible that some term does not appear in the query but is identified in the term frequency. To omit such words they have used document frequency for ellipsibility. By using Adaptive ranking they are determining the relevant document for each query.

L.V Lita et al., [3] have put their efforts for coding the patient records to standard medical codes using ICD9 and CPT. They have compared two algorithms for assigning codes to patient visits as well as to a real world data set. The first algorithm is a support vector machines. Super vector machine is a supervised learning algorithm which is generally used for classification problem. The second approach they followed for the problem is the probabilistic approach based on Gaussian Process. It is views as machine learning algorithm. In this process an observation occurs in continuous domain of space and time. From the experiment they concluded that both support vector machine and Gaussian process are fast to train and archive comparable results.

K. Carmmer et al., [4] also presents a system for assigning ICD9 CM clinical codes to radiology reports. They have developed three automated system for mapping the ICD9 codes to free text reports. The first system uses the natural language system from the ICD9 description and text from the patients report. The second system a rule based system that assign the code based on the overlap between the free text and the code description. The third system is automatic code assignment system that mimics the guideline provided by the medical staff who will assign these codes. And at last they combined all the three system to improve the performance if the system. They cascaded the automatic policy and the rule based system and then the result of the cascaded system is applied to the learning system.

Yefeng Wnag et al., [5] have created an algorithm which maps the medical expression in the patients report into the medical terminology. They have run the system on 470,000 reports from ICS. The system is used to create index of the medical terms into augmented lexicon. The augmented lexicon is used to keep the track of the words that appear in the concept of medical terminology. The algorithm includes a pattern search mechanism, which align the text from the patient report into target medical terminology to find the possible match. The algorithm iteratively performs the alignment to find the best match.

L. Nie et al., [6] have presented their work in eliminating the vocabulary gap between the health seeker and health provider, which is exists due to the community generated content in the web health portals. The queries posted by the health seeker are in the narrative language which cannot be used in an efficient manner to extract medical information. A similar question might be portrayed in generously unique courses by two individual health seekers. On the opposite side, the answers gave by the doctors may contain acronyms with various conceivable meaning, and no standardized terms. To overcome this issue the authors have proposed two approach; local mining and global learning. First is local mining which extract the medical concept form the queries posted by the health seekers and then normalize it to authenticated terminologies. Second is global learning which works towards enhancing the local medical coding by discovering missing key terminologies and eliminating the irrelevant terminologies by analyzing the social neighbors.

III. METHODOLOGY

A. Local Mining

This area point of interest is the local mining approach. To achieve this, they build up a tri-stage frame work. In particular, given a medical record, first concentrate the embedded noun phrases. They then recognize the medical concepts from these noun phrases by measuring their specificity. At last, they standardize the detected medical concepts to medical terminologies.

1. Noun Phrase Extraction: To concentrate all the noun phrases, first allot part of speech tags to every word in the given medical record by Stanford POS tagger. Next haul out sequences that match a settled example as noun phrases. For instance, the accompanying complex sequence can be separated as a noun expression: "incapable treatment of terminal lung cancer ". In addition to essentially hauling out the phrases, they have performed some basic post preparing to interface the

variations together, for example, singularizing plural variations.

2. Medical Concept Detection: This stage plans to separate the medical concepts from other general noun phrases. The authors assume that ideas that are important to medical domain happen often in medical domain and once in a while in non-medical ones.
3. Medical Concept Normalization: It is essential to standardize the identified medical concepts to some suitable standard medical dictionary. There exist various validated vocabularies, including ICD, 7 UMLS, and SNOMED CT.8 these medical and clinical terminologies were made in various times by various associations for various purposes. Take ICD for instance: it is regularly utilized for external reporting prerequisites or different uses where information accumulation is worthwhile. In this work, SNOMED CT is utilized since it gives the center general wordings to the electronic health record and formal rationale based various leveled structure. Local mining wordings may experience the various issues. The primary issue is incompleteness. This is on account of some key medical concepts not expressly introduce in the medical records are rejected. The second one is the lower precision. This is because of some medical concept ideas expressly inserted in the medicinal records, and is erroneously distinguished and standardized by the local approach. Another issue, which merits facilitate talk here, is the terminology space. It might bring about the crumbling in coding execution as far as productivity and Effectiveness.

B. Global Learning

The objective of this area is to take in suitable terminologies from the global terminology space to clarify every medical record. L. Nie et al., [6] additionally investigate the graph based learning model to achieve better performance, and expect this model can at the same time consider different heterogeneous prompts, including the medical record content examination, terminology-sharing systems, and the inter-expert as well as inter-terminology relationships.

1. Relationship Identification: The inter-terminology and inter-expert relationships are not instinctively observed or inferred from medical records. In this manner it is called as implicit relationships. This subsection means to acquaint how with find these sorts of relationships.
2. Inter-Terminology Relationship: The medicinal terminologies in SNOMED CT are composed into non-cyclic ordered (is-a) hierarchies systems. For instance, "viral pneumonia" is-an "infectious pneumonia" is-a "pneumonia" is-a "lung ailment". Terminologies may likewise have many parents. For instance, "infectious pneumonia" is likewise a child of "infectious ailment". The medical terminology hierarchy of command will upgrade the plan in two ways. To begin with, it handles the granularity mismatch issue, where the terminologies found in the medicinal records are extremely detailed and particular, while those in the inquiry might be more broad and high-level. This is accomplished by remunerating the ancestral nodes with appropriate weights. Second, the hierarchical relationships help the coding accuracy through filtering through the sibling terminologies.
3. Inter-Expert Relationship: The inter-expert relationships will be seen more grounded if the specialists are experts in the same or related particular medical areas. This is reflected by their chronicled information, i.e., the quantity of inquiries they have co replied.

IV. CONCLUSION

This paper presents a survey on different system proposed to code the medical concept into standard medical terminology. There are various web health portal which have community generated data in their repositories, which cannot be utilized in an efficient manner to extract the medical knowledge from them. For the same reason a system is proposed which comprises of two components, local mining and global learning. Local mining is a tri-stage framework which locally codes the medical concept form the queries posted by the health seekers and then normalize it to authenticated terminologies. Second is global learning which works towards enhancing the local medical coding by discovering missing key terminologies and eliminating the irrelevant terminologies by analyzing the social neighbors.

REFERENCES

[1] J. Patrick, Y. Wang, and P. Budd, “An automated system for conversion of clinical notes into snomed clinical terminology,” in Proceedings of the fifth Australasian symposium on ACSW frontiers, 2007.

[2] Mi-Young Kim and Randy Goebel,” Detection and Normalization of Medical Terms Using Domain-specific Term Frequency and Adaptive Ranking,” in Proceedings of the 10th IEEE International Conference, 2011.

[3] L. V. Lita, S. Yu, S. Niculescu, and J. Bi, “Large scale diagnostic code classification for medical patient records,” in Proceedings of the Conference on Artificial Intelligence in Medicine, 1995.

[4] K. Crammer, M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll, "Automatic code assignment to medical text," in Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing, 2007.

[5] Y. Wang, J. Patrick, " Mapping clinical notes to medical terminology at point of care," in Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, 2008.

[6] L. Nie, Y. Zhao, M. Akbari, J. Shen, and T. Chua. "Bridging the vocabulary gap between health seekers and healthcare knowledge", In IEEE Transactions, 2014.