Opinion based Data mining Model in Medical Sector

Sagar B. S[1] Arpita Kulkarni[2] Chaithra[3] Jyothi[4] Natesh[5]
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

Abstract— Online health communities keep on offering enormous assortment of clinical data helpful for clinical professionals, framework executives and patients the same. In this work we gather ongoing health posts from reputed sites, where patients express their perspectives, remembering their encounters and symptoms for drugs utilised by them. We propose to perform Summarization of client posts per medication, and come out with helpful resolutions for clinical club just as patient network initially. Further, we propose to arrange the clients dependent on their 'enthusiastic perspective'. Additionally, we will perform information revelation from client posts, whereby helpful 'designs' about the triad 'drugs-symptoms-medicine' is done by Association Rule Mining.

Keywords: Association rule, pattern prediction, Summarization.

I. INTRODUCTION

With the colossal increment in web, electronic data is additionally expanding in tremendous sum which, albeit great as for Information Age, makes overhead of reality. Additionally understand ability of data and subsequent information keep on being huge difficulties.

For information mining of the wellbeing posts, we propose to apply diverse significant activities like - Association Rule Mining, Summarization and Sentiment Analysis on information acquired from the wellbeing gathering site healthboards.com.Summarization is characterized as taking data from the source, removing content from it, and introducing the most valuable substance to the client in a consolidated structure and in a way reasonable to the client's application needs.

Summarization is significant in various NLP applications like Information Retrieval, Quality Analysis, and Text Comprehension and so on. Generally there are two kinds of synopses. Initial one is Extract in which substance from text for example words and sentences are reused. Second one is Abstract which incorporates recovery of removed substance [Association rule mining is a well known and broadly realized information mining task. It is utilized to discover fascinating relations between factors with regards to huge database. Rules produced by affiliation have two disjoint arrangements of things having structure LHS (Left Hand Side) => RHS (Right Hand Side). The standard says that RHS is probably going to happen at whatever point the LHS set happens.

.Extraction of association rules includes two steps:

- Association Rule generation
- Interesting Rule Selection

After the rules have been obtained, they are extracted and post processed. The extracted rules from the health board dataset could take one or more of the following form-

1. Symptoms->disease
2. Disease->disease
3. Medicine->disease
4. Disease->medicines
5. Age group->disease.

2.Implementation

Sentiment Analysis (SA) or Opinion Mining (OM) is task of finding sentiments from text. These sentiments may take different forms like – opinions from people, attitudes and emotions toward an entity. The entity can represent individuals, events or topics. These topics are most likely to be covered by reviews. Sentiment Analysis is a classification process. Classification levels considered were- Document level, sentence level and aspect level. While doing SA first the important features are selected from text then classification is done using appropriate classifier. We are considering reviews from health posts and in our case represented entity is drug. So our classification falls in aspect level.

Summarization module: In health service related sites they take the input from the users in two ways i.e. structured way and the other way of taking input is unstructured way. The structured way takes input from the users in a drop down format or a labeled format example they may give the labels like age, gender, dropdown to select the disease or the text box to mention the disease, priority for the symptoms that the user is suffering from. Unstructured way of taking input involves free form of users descriptions where he can freely write everything about the disease , symptoms ,tests he had taken , the drug he/she using In our project we used Unstructured way for the opinion by the use but this unstructured form of taking inputs may be lengthy for the doctors to read and it is problematic to extract the relevant contents from the lengthy text and it is time consuming too so there are chances of skipping the lengthy descriptions by the doctors so we need to summarize the data i.e. we need to extract the only important information from the lengthy text and the summarized data is been given to the doctors/predicators.

A) Summarization model

In health service related sites they take the input from the users in two ways i.e. structured way and the other way of taking input is unstructured way. The structured way takes input from the users in a drop down format or a labelled format example they may give the labels like age, gender, dropdown to select the disease or the text box to mention the disease, priority for the symptoms that the user issuffering from. Unstructured way of taking input involves free form of users descriptions where he can freely write everything about the disease , symptoms ,tests he had taken , the drug he/she using In our project we used Unstructured way for the opinion by the use but this unstructured form of taking inputs may be lengthy for the doctors to read and it is problematic to extract the relevant contents from the lengthy text and it is time consuming too so there are chances of skipping the lengthy descriptions by the doctors so we need to summarize the data i.e. we need to extract the only important information from the lengthy text and the summarized data is been given to the doctors/predicators.

Summarization is defined as taking the information from the input by the user, extracting the important content from the source input, and presenting only the most important content to the doctors. Now the source input is in condensed form [1].

Summarization of data is very important in natural language processing applications like Retrieval of important information, Analysis of quality, condensed text etc.

For the summarizing of the input we use Lesk based algorithm[2]

Lesk based algorithm is to compare the words in the ambiguous word dictionary with the terms contained in its neighborhood.

Lesk Based Algorithm steps:

1. Scan the Opinion Database (Retrieval of all the Patient'sOpinion)
2. Scan the WordNet (Collection of all Symptoms, Disease, Drugs and Sentiment Analysis words like satisfied and depressed)
3. For each entry of Opinion in stored in buffer do
4. Trace all the keywords using the flowing step:
    a) Tokenization [keywords extraction method- removing stop words and retrieving the keywords]
    b) Remove punctuation, special characters, number etc.
    c) Clustering the keyword (group of similar objects)
        1. By comparing with the predefined data sets (created by the admin)
        2. .String Comparison and Identify the symptoms, Disease, Drugs and Positive and Negative Words

Output – Summarized Results

Algorithm implementation:

1st step: scan the opinion

2nd step: scan the WordNet

3rd step: store the all opinions in a buffer

4th step: tracing of Keywords:

1. Tokenization step- removing the stops words

Output after the tokenization- Tramadol, cough, headache, relief good

2. Removing punctuation,special characters,numbers

Output –Tramadol cough headache relief good

3. Clustering the Keywords

Drug: tramadol

Symptoms: cough headache

Satisfied: relief good

Disease: - [No disease is mentioned in the opinion]

Depressed: - [No depressed word in mentioned in the opinion]

B) Pattern prediction module:

From removed catchphrases got in above module we propose to decide the various sorts of affiliation. These affiliations could be among illness, medication and side effects. We will utilize Apriori calculation for this reason.

Apriori algorithm steps:

STEP 1: Scan the opinion data set and determine the support(s) of each item.

STEP 2: Generate L1 (Frequent one item set).

STEP 3: Use Lk-1, join Lk-1 to generate the set of candidate k - item set.

STEP 4: Scan the candidate k item set and generate the support of each candidate k – item set.

STEP 5: Add to frequent item set, until C=Null Set.

STEP 6: For each item in the frequent item set generate all non empty subsets.

STEP 7: For each non empty subset determine the confidence. If confidence is greater than or equal to this specified confidence .Then add to Strong Association Rule.

If opinion is related to paraacetomol drug and the it is used more than other drugs, then this module will create a pattern that if fever is the disease then some percentage of people is using paracetomol as a drug.

C) Sentiment analysis:

Sentiment analysis deals with the diagnosis of health care related problems identified by the patients themselves. It collects the patient's opinions into perspective to make policies and modifications that could directly address their problems. Sentiment analysis is used in different product for great growth and also in other application areas. We use Machine learning techniques to analyze the review documents and conclude them towards an efficient and accurate decision. Where the structured way of data techniques has high accuracy but is not extendable to unknown domains while unstructured data techniques have low accuracy [3].

We use Lesk based algorithm for sentiment analysis. The purpose of applying sentiment analysis techniques is to process drugs related opinions of millions of users and giving them useful information. Sentiment analysis result can be in the form of binary classes which representing the percentages of positive and negative sentiments. If finer level categorization is required the aggregated result can be in different categories like satisfied, excellent, good, mediocre, unsatisfied, bad, worse etc.

Example: I took the Tramadol, basically I was suffering from cough, headache, after taking the drug, I got relief and the drug is good.

Sentiment analysis Results:

Opinion	Result
Example: I took the Tramadol, I was suffering from cough, headache After taking the drug, I got relief and the	satisfied

drug is good.

Example: I took Aspirin, basically I had unsatisfied/bad Body pain even though after taking tablet I didn't get relief.

Conclusion:

In this work, we gather original comments from rumored sites, and perform data mining to decide the different potential relationship from these posts and perform information revelation from client posts and identify valuable 'designs' about gatherings like: sickness to ailment, illness to medication and medication to side effect. This is finished utilizing Association rules calculation. This will assist the specialists with finding symptoms of various medications and with this they can recommend better.

medications to different patients with comparative sickness. Pharmaceutical organizations can the reaction of a few medications on individuals and will get a thought regarding which medication is well known and ought to be created. This will likewise assist the patients with knowing about the assessment of past clients, hence will be in a superior situation to choose which medication ought to be taken for a specific infection and furthermore improve mindfulness on different symptoms of medications looked by others.

Future work:

We can include live messaging feature with the clinical specialist where the guest can visit to explain about the medication and symptoms which are caused as side effects. Web-based social networking posts contain a great deal of blunders or spelling botches. We are not considering spelling botches and their remedy. So this could be further improvement. Posts in person to person communication may likewise contain expression which contains symbols, which are not considered in this work.

REFERENCES

[1] JayashreeR,SrikantaMurthy K,Basavaraj .S.Anami, "Categorized Text Document Summarization in the Kannada Language by Sentence Ranking", 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp 776-781, 2012.

[2] AlokRanjan Pal, DigantaSaha, "An Approach to Automatic Text Summarization using WordNet", IEEE International Advance Computing Conference (IACC), 2014.

[3] JesminNahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, "Association rule mining to detect factors which contribute to heart disease in males and females", J. Nahar et al. / Expert Systems with Applications 40 (2013) 1086–1093, Elsevier, 2012.

[4] Lakshmi K.S, G. Santhosh Kumar, "Association Rule Extraction from Medical Transcripts of Diabetic Patients",IEEE,2014.

[5] WalaaMedhat, Ahmed Hassan, HodaKorashy, "Sentiment analysis algorithms and applications: A survey", In press, Elsevier, 2014.

[6] Rafael Ferreira, FredericoFreitas, Luciano de Souza Cabral, Rafael DueireLins, Rinaldo Lima, Gabriel Franca, Steven J. Simske, and Luciano Favaro, "A Context Based Text Summarization System",11th IAPR International Workshop on Document Analysis Systems,pp 66-70, 2014.

[7] C. Lakshmi Devasenal and M. Hemalatha, "Automatic Text Categorization and Summarization using Rule Reduction", IEEE- International Conference On Advances In Engineering, Science And Management (ICAESM - 2012), pp 594-598, 2012.

[8] Sara Keretna, CheePeng Lim, Doug Creighton, "A Hybrid Model for Named Entity Recognition Using Unstructured Medical Text", Proc.Of the 2014 9th International Conference on System ofSystems Engineering (SOSE), Adelaide, Australia- June 9-13, pp 85-90, 2014.

[24]SaeedMohajeri,AfsanehEsteki, Osmar R. Zaiane and

DavoodRafiei, "Innovative Navigation of Health Discussion Forums based on Relationship Extraction and Medical Ontologies",IEEE International Conference on Bioinformatics and Biomedicine, pp 13-14, 2013.

[9] Yi Chen, Yunzhong Liu, "Connecting the Dots: Knowledge Discovery in Online Healthcare Forums", ICEC'14 August 05 - 06 2014, ACM.

[10] Subhabrata Mukherjee, Gerhard Weikum, CristianDanescu-Niculescu-Mizil, "People on Drugs: Credibility of User Statements in Health Communities", KDD '14, August 24 - 27 2014, New York, ACM, 2014.

[11] Khairullah Khan, BaharumBaharudin, Aurnagzeb Khan, Ashraf Ullah, "Mining opinion components from unstructured reviews: A review", Journal of King Saud University – Computer and Information Sciences (2014), Elsevier, 2014.

Miller G, Beckwith R, Fellbaum C, Gross D, Miller K., "WordNet: an on-line lexical database", OxfordUniversityPress.\ DavoodRafiei, "Innovative Navigation of Health Discussion Forums based on Relationship Extraction and Medical Ontologies",IEEE International Conference on Bioinformatics and Biomedicine, pp 13-14, 2013.

[12] Yi Chen, Yunzhong Liu, "Connecting the Dots: Knowledge Discovery in Online Healthcare Forums", ICEC'14 August 05 - 06 2014, ACM.

[13] Subhabrata Mukherjee, Gerhard Weikum, CristianDanescu-Niculescu-Mizil, "People on Drugs: Credibility of User Statements in Health Communities", KDD '14, August 24 - 27 2014, New York, ACM, 2014.

[14] Khairullah Khan, BaharumBaharudin, Aurnagzeb Khan, Ashraf Ullah, "Mining opinion components from unstructured reviews: A review", Journal of King Saud University – Computer and Information Sciences (2014), Elsevier, 2014.