International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

Summarization and Sentiment Analysis from User Health Posts

Sharath M P, Prof. H. P. Ramyashree

Department of Master of Computer Application, PES Engineering of College, Mandya, India

ABSTRACT

Online health communities continue to offer huge variety of medical information useful for medical practitioners, system administrators and patients alike. In this work we collect real time health posts from reputed websites, where patients express their views, including their experiences and side-effects on drugs used by them. We propose to perform Summarization of user posts per drug, and come out with useful conclusions for medical fraternity as well as patient community at a glance.With the enormous increase in web, electronic information is also increasing in huge amount which, although good with respect to Information Age, creates overhead of time and space. Also understandability of information and consequent knowledge continue to be big challenges.

Keywords : Data mining, Apriori Algorithm, patterns, drugs-symptoms-medicine' is done by Association Rule Mining.

I. INTRODUCTION

Summarization is defined as taking information from the source, extracting content from it, and presenting the most useful content to the user in a condensed form and in a manner suitable to the user's application needs. There are two types of summaries, first one is Extract in which contents from text that is words and sentences are reused. Second one is Abstract which includes regeneration of extracted contents.

Association rule generation is used, were rules are extracted and post processed. The extracted rules from the health boards dataset could take one or more of the following form-

1. symptoms->disease
2. disease->disease
3. medicine->disease
4. disease->medicines

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.

II. EXISTING SYSTEM

A. Existing System

- Online health communities continue to offer huge variety of medical information useful for medical practitioners, system administrators and patients.
- Health communities just collects real time health posts, where patients express their views, including their experiences and side-effects on drugs used by them.
- These systems just collect the data, stores in database and retrieves the same in future, but no extraction of useful information which helps the medical practitioners.
- But no Summarization and extraction of useful information .
- Hence in the existing system, it is difficult to analyze the data for the users.

B. Drawbacks of Existing System

- Just stores the health posts
- No Summarization
- No extraction of useful information
- Less user satisfaction
- Stores huge amount of data

III. PROPOSED SYSTEM

Proposed system is a web-enabled application which helps the medical sector to know the popularity of a particular drug before manufacturing. Here for the prediction of product demand we make use of data mining technique called as classification rules. Proposed system is an online medical-sector application. This system collects posts from the users related to side effects on drug and discovers useful patterns based on side effects per drug with the help of Association Rule.

IV. LITERATURE SURVEY

In this system, it describes three tier architecture which consists of three layers, Data layer, Business layer, Presentation layer. Following work has been reviewed with respect to summarization task. A summarization approach using simplified Lesk algorithm was used in [1]. After weighting, the sentences are arranged in descending order and summarization is performed by taking percentage of summarization as input. Result is measured in terms of precision, recall, f-measure. This algorithm is simple and each sentence is considered separately for evaluation hence useful in summarization of user posts.

Following work has been reviewed with respect to Association task.A new method to find out association rules from medical transcripts, Apriori and FP-growth algorithms is used in [2]. They used small dataset therefore rules generated are less as well as already known. To get new rules dataset must be large.

A keyword based summarization approach is proposed in [3]. A combination of GSS coefficient and IDF methods along with TF was done for extracting keywords. The most weighted sentences as per user input are selected for summary generation. The results given by them are not promising.

Jesmin Nahar et al applied association rule mining on UCIDataset of heart disease. After getting rules they analyzed rules in association with gender and significant risk factors. They used apriori, predictive apriori and tertiusalgorithms for rule generation. Their research shows that how computational intelligence can be used to identify important factors responsible for disease [4].

Hybrid model of lexicon based and rule based techniques was used on unstructured and informal medical text in [5].

Sentence level information is not considered otherwise result would be better. A medical ontology that provides an interface for navigating through discussions using MeSH was proposed in [6]. Its major disadvantage is that medical terms with only one word are considered.

V. DESIGN AND ARCHITECTURE

This system describes three tier architecture which consists of three layers, Data layer, Business layer, Presentation layer

The Data Layer

The key component to most applications is the data. The data has to be served to the presentation layer somehow. The data layer is a separate component (often setup as a separate single or group of projects in a .NET solution), whose sole purpose is to serve up the data from the database and return it to the caller.

Business Layer

Though a web site could talk to the data access layer directly, it usually goes through another layer called the business layer. The business layer is vital in that it validates the input conditions before calling a method from the data layer. This ensures the data input is correct before proceeding, and can often ensure that the outputs are correct as well. This validation of input is called business rules, meaning the rules that the business layer uses to make “judgments” about the data.

Presentation Layer

The ASP.NET web site or windows forms application (the UI for the project) is called the presentation layer. The presentation layer is the most important layer simply because it’s the one that everyone sees and uses. Even with a well-structured business and data layer, if the presentation layer is designed poorly, this gives the users a poor view of the system.

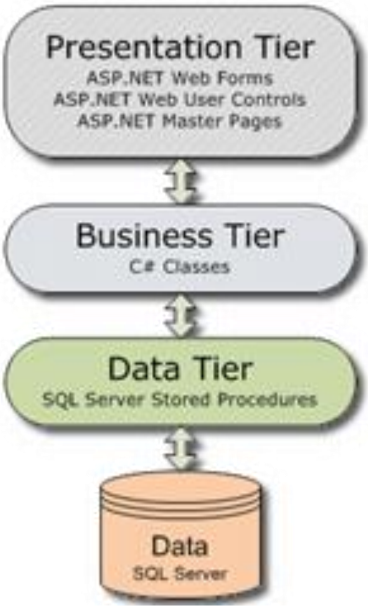

Figure 1. Three Tier Architecture

Shows the proposed system architecture; it describes the process of Association rule performed to extract the best pattern of symptom-disease-drug

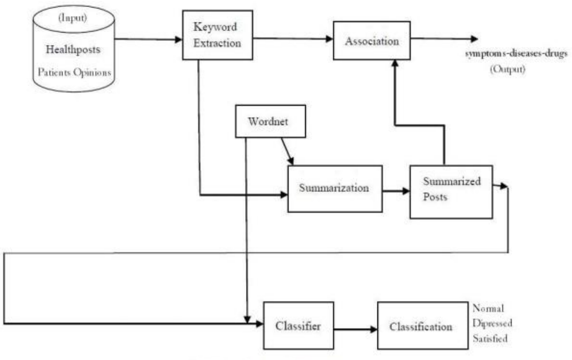

Fig.3 Proposed System Architecture

Figure 2. Proposed System Architecture

VI. VI.METHODOLOGY

Apriori Algorithm

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.

VII. CONCLUSION

Analyzing user posts from health communities for knowledge discovery is an interesting area in research. This work will help patients to find out association among different drugs, diseases and symptoms. It will help doctors to find out side-effects of different drugs so they can prescribe better drugs to other patients with similar disease. Pharmaceutical companies will be also benefited as we are classifying users of particular drug into different classes like normal, depressed and satisfied. This will be indirect input to companies to decide which drug is popular, whether to produce alternate drug to this etc. Thus our work shall equally benefit all three parties–medical fraternity, patient community and pharmaceutical companies.

VIII. FUTURE ENHANCEMENT

We can add live chat feature with the medical practitioner where the visitor can chat to clarify about the drug and other side effects. Social media posts contain a lot of errors or spelling mistakes. We are not considering spelling mistakes and their correction. So this could be further improvement.

Posts in social networking may also contain symbolic expressions So this could be further improvement.

IX. REFERENCES

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