INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Summarization and Sentiment Analysis from User Health Posts

Ramya J.
Students, Department of Computer Science & Engineering
The National Institute of Engineering, Mysore, India
Megha K. M.
Students, Department of Computer Science & Engineering
The National Institute of Engineering, Mysore, India
Nishkala Nayak M. N.
Students, Department of Computer Science & Engineering
The National Institute of Engineering, Mysore, India
Sheela G. V.
Students, Department of Computer Science & Engineering
The National Institute of Engineering, Mysore, India

Abstract:
Health communities offer huge variety of Information regarding Medical sector which is useful for Drug dealers, Doctors and patients. This work includes the collection of Real time health posts from trusted websites, these websites contain patients experiences and side effects on drugs used by them. By collecting these information from trusted websites, this paper perform summarization of user posts per drug and come out with useful conclusions for Drug dealers, Doctors, Patients.
Further, It classify the users based on their 'emotional state of mind' In this paper it perform the knowledge discovery from user post, which gives useful 'Patterns', 'Keyword extraction', by using Association rule Method.

1. 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.

2. Existing System and its Disadvantages

There are several Online health communities which provides information regarding Drugs, symptoms, Diseases. It just collects the data, stores in database and retrieves the same in future, but no Summarization and extraction of useful information which helps the Drug dealers, Doctors and patients. Hence in the existing system, it is difficult to analyse the data for the users.

3. Proposed System

In this system it includes the collection of Real time health posts from trusted websites, these websites contain patients experiences and side effects on drugs used by them. By collecting these information from trusted websites, this paper perform summarization of user posts per drug and come out with useful conclusions for Drug dealers, Doctors, Patients.

3.1. Advantages

- Proposed system is a medical sector application.
- Proposed system collects the posts from the users (Drug dealers, Doctors, patients) related to side effects on drug.
- Proposed system summarize all the user posts and come out with useful conclusions.
- Proposed system discovers useful patterns based on side effects per drug.
- Proposed system makes use of "Association Rules" for pattern discovery.
- Proposed system is a new online community where we huge variety of medical information useful for Drug dealers, Doctor, patients etc.

4. 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.

5. Design and Architecture

This system describes three tier architecture which consists of three layers, Data layer, Business layer, Presentation layer [Figure 1].

5.1. 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.

5.2. 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.

5.3. 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.

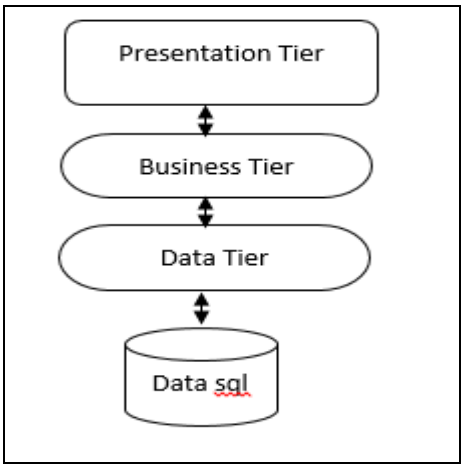

Figure 1: Three tier architecture

5.3.1. Data Flow Diagram:

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system. DFDs can also be used for the visualization of data processing (structured design).

5.3.2. Symbols used in DFD's

5.3.2.1. Processes

A process transforms data values. The lowest processes are our functions without side effects.

5.3.2.2. Data Flows

A data flow connects the output of an object or process to the input of another object or process. It represents the intermediate data values within the computation. It is draws as an arrow between the procedure and the consumer of the data value.

5.3.2.3. Actors

An actor is an active object that drives the data flow graph by producing or consuming values. Actors are attached to the inputs and the outputs of a dataflow graph.

5.3.2.4. Data Store

A data store is a passive object within a data flow diagram that stores data for later access. Unlike an actor, a data store does not generate any operations on its own but merely responds to requests to store and access data.

FIG 2 Shows the Data Structure, it consists of the following table, tblkeywords, tdlopinions, tdlLogin. In tblkeywords – keywords are the primarykey, in tblLogin – LoginId is primarykey, in tblOpinions – patientId is primarykey.

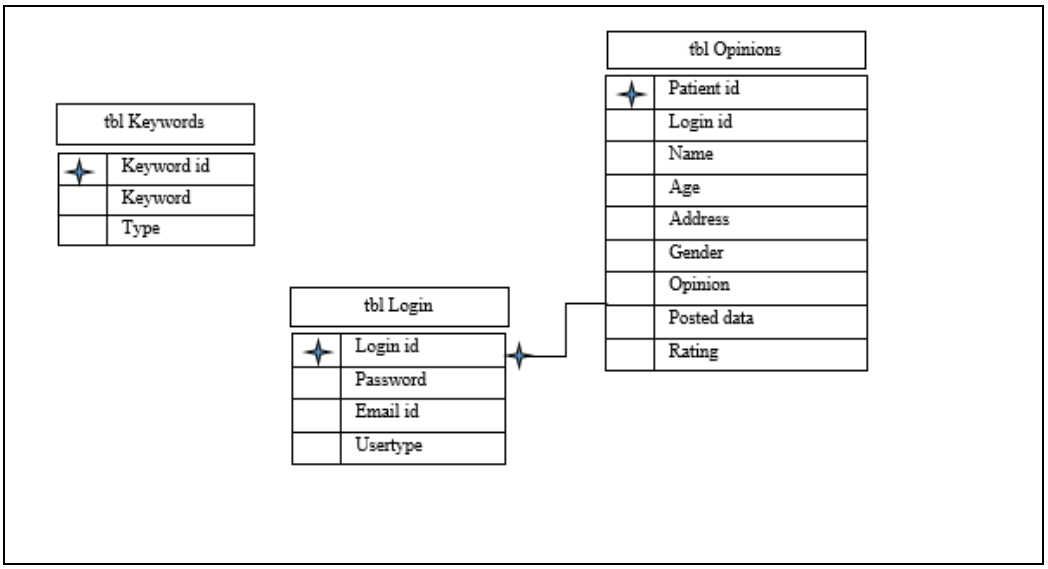

Figure 2: Data Structure

FIG-3 Shows the proposed system architecture; it describes the process of Association rule performed to extract the best pattern of symptom-disease-drug.

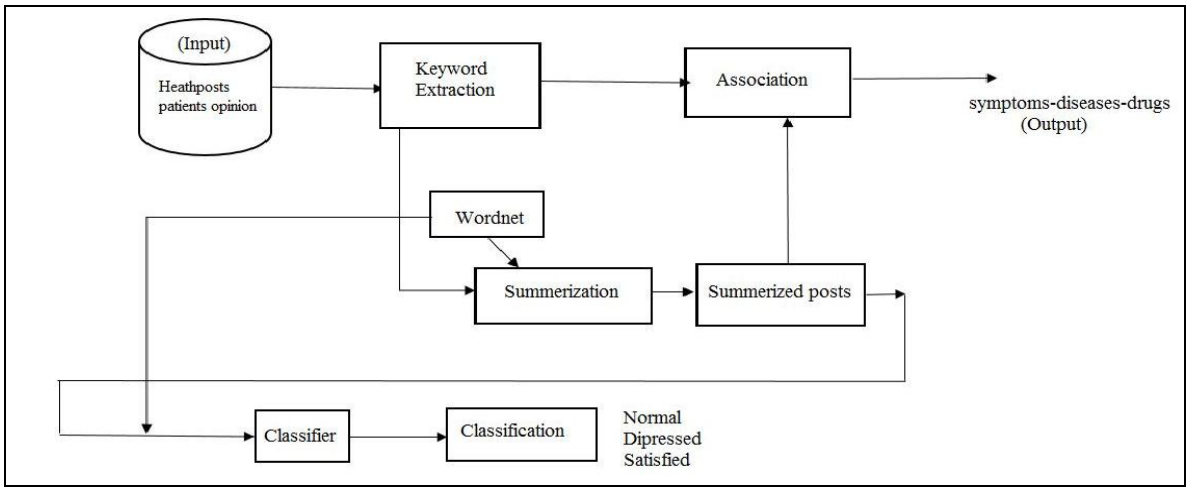

Figure 3: Proposed System Architecture

6. Conclusion and Future Enhancement

Health communities analyze user health posts for knowledge discovery is an interesting area in research. This work will help patients to find out association among different drugs, disease 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 also be benefited as we are classifying users of particular drug into different classes like 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.

- 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, which are not considered in this network. Not stopping just at looking at attitudes but also to help people at the moment of indecision.
- Advertising the best available medicine to the customers looking at his opinions and desires, to correlate between the sentiment and the behavior and come up with a pattern for each individual.

7. References

i. AlokRanjan Pal, DigantaSaha, "An Approach to Automatic Text Summarization using WordNet", IEEE International Advance Computing Conference (IACC), 2014.
ii. Lakshmi K.S, G. Santhosh Kumar, "Association Rule Extraction from Medical Transcripts of Diabetic Patients",IEEE,2014.
iii. JayashreeR,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.
iv. 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.
v. 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.
vi. 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.
vii. Yi Chen, Yunzhong Liu, "Connecting the Dots: Knowledge Discovery in Online Healthcare Forums", ICEC'14 August 05 - 06 2014, ACM.