ISSN:0975-9646

Knowledge Discovery from Various Algorithms: A Survey

VINOD L. MANE , PROF. S. S. PANICKER , PROF. V. B. PATIL

Maharashtra Institute of Technology,
Pune, India

***Abstract*-Text mining is nothing but extracting useful information from text. The information to be extracted is explicitly stated in the text. Text mining can be applied in various domains like medical, economical, etc. Medical text extraction can be done through medical patient reports, online medical journals, online health communities, etc. Text mining includes different tasks like classification, summarization, clustering, association, etc.Summarization is done to view important contents at a glance.Online health communities contain a lot of useful information. Hence, summarization is used to generate most important summary. Also, association between drugs, symptoms and diseases can be determined from contents of health communities.**

Keywords: **Summarization, classification, patterns, association rule mining.**

I. INTRODUCTION

Drug side effects are a threat to health, sometimes to the extent of causing death, as well as imposing a substantial time and financial cost from patients and healthcare providers. The most common of these negative side-effects are reported by their manufacturers; however, other side-effects may be discovered through different surveillance methods, including volunteer reporting by doctors and pharmacists, and recently, patient reporting themselves in online forums. The recent popularity of health related communities has enabled users to communicate about drugs, treatments and other health related issues over the Internet, making it a rich resource of information. This way of exchange of opinions and experiences has provided a rich source of information about drugs and their effectiveness and more importantly, their possible side-effects.However such user generated contents on health communities may be noisy, inaccurate or contain misinformation. Hence,online health communities,although a rich source of information, continue to have challenges like - trusting the users and their statements, as these shall adversely affect the performance of system.

With increase of web 2.0, online contents are increasing day by day. Online health communities contain a lot of valuable information. But, this information is unstructured, noisy, and written in formal way. So there is need to select important sentences from these communities i.e. proper summary should be generated out of this information.

II. RELATED WORK

Alok Pal et.al, provided an approach for summarization. They used WordNet dictionary. They used simplified lesk algorithm for giving weightage to each sentence. After weighting, sentences are arranged in descending order of weight. Summarization is performed on these sentences provided percentage of summarization as input. They used fifty texts from different categories like writer, soul, sports, etc. as input sentences and result is measured in terms of precision, recall, f-measure. When there are more numberof named entities, this approach may not give good results as less number of named entities means more number of meaningful words in the sentence. Hence, better weighting is done [1].

Rafael et. al gave extractive text summarization based on sentence scoring techniques. The main aim of their approach is to increase the quality of summarization. Their approach is based on context of the text. Word based, sentence based and graph based scoring methods are combined together to give weightage. Evaluation of summary is done by ROUGE as a quantitative measure and by counting number of sentences selected by system that match human gold standard as qualitative measure [2].

Jayashree R et al proposed a keyword based summarization approach. Their technique uses combination of GSScoefficient and IDF methods along with TF for extracting keywords from kannada webdunia, which is a portal offering news from politics, sports, cinema, etc. Using this approach, weight for each word is calculated. Summation

Table 1: Literature Reviewof weights of all keywords gives weight of each sentence. As per user input most weighted sentences are selected for summary generation [3].

C. Lakshmi Devasana et al developed a text analyzer. It uses rule reduction technique to derive the structure of input text. First tokens are created from input text then important features are identified and finally categorization and summarization is done. Rules are generated for noun phrase(NP), prepositional phrase (PP), possessives (POSS), verb phrase (VP). Text analyzer categorize tokens as per these rules and summarize them to formulate a sentence [4].

Lakshmi K.S. et al gave a new method to find out association rules from medical transcripts. These rules gave association between disease with other diseases, symptoms of particular disease, medications used for treating diseases and age group for developing particular disease. NLP tools along with apriori and FP-growth algorithms are used for extracting association rules. UMLS is used to identify medical terms. Correlation measure and lift are used as evaluation measure for selecting important rules [5].

Sara keretna et al used unstructured and informal medical text as input and extract drug named entities from it, using hybrid model of lexicon based and rule based techniques.

First, Lexicon rules are used to detect drug names. But if spelling mistakes are there then lexicon can't detect drug name as not exact match found.If lexicon fails to identify drug names then inference

rules are used to detect undiscovered drug names. This hybrid model is evaluated using F-measure [6].

SaeedMohajeri et al gave an approach to form an ontology of medical entities from medical discussion forums. This ontology provides an interface for navigating through discussions. They used MeSH to extract medical entities from discussions and automatically detect relationships between these entities using methods like pattern matching, co-occurrence matrix and distribution based method. The drawback of their approach is not considering the longer medical terms i.e. medical terms with one word are considered only [7].

Subhabrata Mukherjee et al identified interaction between trustworthy users, language objectivity and credibility of statements.Their model detect unknown side-effects of drug and removes false statements.To determine the objectivity and quality of posts, they used stylistic and affective features of language. Performance of this model is validated against expert knowledge source of drug and their side-effects[8].

Khairullah Khan et al reviewed opinion mining components. They gave in brief about opinion mining, their applications, different tasks in opinion mining. Opinion mining has problems like accuracy, quality, standard of data, etc. As per authors, ambiguity, semantic relatedness, context dependency are among major challenges in opinion mining. [9]

Sr.No.	Title Of Paper	Publisher And Year	Algorithms/Methods	Advantages	Disadvantages	Data Used
1.	People on Drugs: Credibility of User Statements in Health Communities	ACM-2014	Markov Random Field	Filter out false information, Identify Trustworthy Users and Find Side-effects of drugs from comments	Simple Information extraction method.	Healthboards.com, Mayoclininc.org
2.	Association Rule Extraction From Medical Transcripts of Diabetic Patients	IEEE-2014	Apriori Algorithm, FP-Growth Algorithm	Find out Disease, treatment from Symptoms, Find out age group for developing disease.	Dataset used is small.	Medical Transcripts
3.	An Approach to Automatic Text Summarization using WordNet	IEEE-2014	Simplified Lesk Algorithm, WordNet Dictionary	Summarize without depending on format of text.	Good for technical reports	Random text, WordNet
4.	A context based text summarization system	IEEE-2014	Combination of Sentence Scoring methods	Improves Quality of summarization	Different techniques to different documents	CNN,Cosmic Variance, Internet Explorer Blog
5.	A Hybrid Model For Named Entity Recognition Using Unstructured Medical Text	IEEE-2014	Lexicon based, Rule based techniques	Extract undetected drug names	Rule extraction algorithm needs to be improved	I2b2 discharge summary reports
6.	Categorized Text Document Summarization in the Kannada Language by Sentence Ranking	IEEE-2012	GSS coefficient, TF-IDF	Summary for different categories are generated	Coherence not considered	Kannada.webdunia.com
7.	Innovative Navigation Of Health Discussion Forums Based on Relationship Extraction and Medical Ontologies	IEEE-2013	Pattern Matching, Co-occurrence matrix	Effective interface for navigating through discussion	Longer medical terms not considered	Healthboards.com, ehealthforum.com

CONCLUSION:

In this work, we focus on different summarization methods for online communities along with their advantages and disadvantages. Techniques to identify relationship between medical entities are also studied. This survey will help to select most appropriate method for association as well as summarization of online health community content. Further, this association can be used to find out important information about disease, medicine, symptom that is not mentioned in the posts.

REFERENCES

1 AlokRanjan Pal, Diganta Saha,“An Approach to Automatic Text Summarizationusing WordNet”, 978-1-4799-2572-8/14, 2014 IEEE.

2 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”,978-1-4799-3243-6/14 $31.00 © 2014 IEEE.

3 Jayashree R, Srikanta Murthy K, Basavaraj S.Anami,“Categorized Text Document Summarization in the Kannada Language by Sentence Ranking”, 978-1-4673-5119-5/12/$31.00 ©2012 IEEE.

4 C. Lakshmi Devasenal and M. Hemalatha,“Automatic Text Categorization and Summarization using Rule Reduction”, ISBN: 978-81-909042-2-3 ©2012 IEEE.

5 Lakshmi K.S, G. Santhosh Kumar, “Association Rule Extraction from Medical Transcripts of Diabetic Patients”,978-1-4799-2259-14,IEEE,2014.

6 Sara Keretna, Chee Peng Lim, Doug Creighton, “A Hybrid Model for Named Entity Recognition Using Unstructured Medical Text”, 978-1-4799-5227-4 $31.00 © 2014 IEEE.

7 Saeed Mohajeri, Afsaneh Esteki, Osmar R. Zaiane, and Davood Rafiei, “Innovative Navigation of Health Discussion Forums based on Relationship Extraction and Medical Ontologies”, 978-1-4799-1310-7/13 $31.00 © 2013 IEEE.

8 ubhabrata Mukherjee, Gerhard Weikum, Cristian Danescu-Niculescu-Mizil,“People on Drugs:Credibility of User Statements in Health Communities”,KDD ’14, August 24 - 27 2014, New York, ACM 978-1-4503-2956-9/14/08.

9 Khairullah Khan, Baharum Baharudin, Aurnagzeb Khan, Ashraf Ullah, “Mining opinion componentsfrom unstructured reviews: A review”, Journal of King Saud University – Computer and Information Sciences (2014), Elsevier, 2014.