International Journal of Advanced Technology & Engineering Research
(IJATER) www.ijater.com

Analyzing User Posts from online health communities for effective knowledge discovery

Ashwini Abhale[1], Vinod Mane[2], Shraddha Shelar[3], Ashish Saxena[4]
[1,2,3,4]Assistant Professor, DYPCOE Akurdi

***Abstract*—In this 21st century, with emerging digital media and globalization, internet has become the primary source of information. The main goal of our paper is to find the association among the drug-disease-symptoms from userposts. But while obtaining knowledge for health issue or medical related content, it is necessary to check the trustworthiness of users or the information posted by them. We have come up with a Credibility module that helps us to determine trustworthiness of users.This module assigns weights to users and filters out the less credible posts by considering the posts of highest weighted users for further processing of finding associations. We used UMLS (Unified Medical Language System) to find meaning of keywords in medical domain. From this we considered drug-disease-symptoms from posts. We used Apriori algorithm to find the association. As we are filtering the posts, results obtained are promising.**

***Keywords*—Apriori algorithm,Association Rule Mining, Trustworthiness, UMLS.**

I. INTRODUCTION

In this modern era, Internet is playing very crucial role in almost every aspect of life. It is becoming one stop solution for all problems. Social networking sites are important means of communication among people. Recently developed Facebook and Twitter is one of the information technology helping people across the world to share and communicate. Statistics shows that hundreds of millions of people are using social media. At the end of 2011, an average of 340 million tweets per day [1] and around 800 million facebook users were there [2]. These numbers are still increasing rapidly.Online health discussion forums are platform where patients share their experiences and also get important advice or suggestions from other patients or doctors [3].These platforms are becoming popular due to fact that generally doctor-patient communication is not user friendly. Also, medical terminologies used by doctors are not easily understood by patient. Health forums provide different ways of communication like patient-patient, patient-doctor, doctor-doctor, experts/researchers-patients, etc. [2]. Thus, these health forums help to reduce the communication gap.

Though these health forums provide useful information in easily understandable language the question arises of credibility of this information. As per Kevin R. Canini et.al [4], credibility is defined by trust as well as support from other professionals. Again, in this, relevance of person's discussion also matters. From this, we know that credibility is associated with users who frequently posts and who got replies for their posts.

This trustworthy information can be very useful. During disasters such as Hurricane Sandy and the tsunami in Japan people used social media to report injuries as well as send out their requests [1].

Medical keywords from social media needs to identified to obtain domain information.Unified Medical Language System (UMLS) is electronic Metathesaurus integrating different terminology systems. It is developed by National Library Of Medicine U.S. [5]. UMLS gives representation of biomedical knowledge providing relationship between different semantic types. This knowledge is useful in information retrieval, decision making, data minng, etc. To use this knowledge Metamap is one of the tool which find biomedical meaning of text from metathesaurus[6]. UMLS contains different categories like sign or symptoms, disease or syndrome, organic chemical, findings, body part, etc[*7]. We organized these categories in three groups drugs, disease and symptoms and done further processing.*

Health forum information is used to find association among disease-drug-symptoms [8]. It may be also useful to find sentiment of users about particular drugs [9].

II. RELATED WORK

How social media can be useful in situation like natural disasters or any social movement is shown by Mohammad-aliabbasi et.al in [1]. They also showed that not all the contents from social media are trustworthy. Hence they gave credrank algorithm to rank trustworthy users and their contents.

Ivan Ibnualim et al in [2] showed different types of communication in social media especially in health social media. They showed that how communication between doctor-doctor, doctor-patient have importance of their own. As per them to improve quality of information we need to focus on credibility and trustworthiness of information, security of information and distinct representation of health data from that of general social media data.

According to WojciechJaworski et al [10] Credibility is dependent not only on webpage but also on each statement on that page. They asked users to rate credibility of webpages as well as statements and observed the approach of users for giving credibility.

Jiang Yang et al [11] compared different factors like gender, naming style, profile image in microblog and their survey showed how these factors affect credibility of microblog in strong nations China and U.S.

III. MODULES

We are collecting data from healthboards.com [13]. At this forum, people share their experiences, ask questions, other user replies them. To do all these activities user need to register themselves. Our approach to find association includes two modules- Credibility and Association.

Details of these modules are as follows:

1) Credibility Module:

As this forum contains large amount of posts we need to filter these posts to obtain only trustworthy posts. For this purpose we collected all registered users data along with their posts. This sample user profiled data is shown in Fig.1.And the flow of module is shown in Fig.2.

We have given weights to users based on membership type of user, number of questions asked, number of replies got. Higher weighted users posts are considered for finding associations.

Our assumption of threshold value for filtering post of higher weighted users is based on following scenarios:

	A	B	C	D	E	F	G	H
1	Author_id	Gender	Location	Posts	Membership_type	Questions	Replies	Thanks_vote
2	130	female	Birmingham, AL	12921	Facilitator	114	8154	817
3	3243	female	Missouri, USA	11743	Senior Veteran	106	5751	639
4	342	female	null	10146	Senior Veteran	144	6148	38
5	1681	female	Chicago,IL	9872	Senior Veteran	116	4576	1012
6	27096	null	ma	9652	Inactive	80	1900	null
7	7	female	NJ, USA	9398	Senior Veteran	135	6111	1141
8	15155	male	null	8510	Senior Veteran	148	5161	6
9	1172	female	USA	7916	Facilitator	190	3404	986
10	6274	male	melbourne, vic, a	7384	Senior Veteran	291	5840	null
11	470	female	western us	7268	Senior Veteran	316	3434	null
12	14803	null	LI, NY	7096	Inactive	608	1150	null
13	1164	male	Kansas, USA	6852	Facilitator	96	3140	1812
14	991	female	charlotte, nc, usa	6837	Senior Veteran	133	2188	2035

Fig.1 Sample Author Profiles

We have given Membership_type"admin" the highest weight i.e. 0.9 and type Member as weight 0.2. Weights of questions and replies are based on average values of questions and replies of all users. In our case we have given weights as 0.3 if less than average and 0.7 if greater than average.

Case 1 : Now, suppose "Admin" user with less questions and replies than average, then total weight of such user would be= 0.9+0.3+0.3=1.5

Case 2: Now, suppose type of user is "member" and he/she has more number of questions and replies then weight would be= 0.2+0.7+0.7= 1.5

Case 3: Now, consider member type "Senior Member" having weightage 0.5, even if he has less question than average and more number of answers than average or vice-versa then weightage would be = 0.5+0.3+0.7= 1.5

Based on above cases we have considered threshold value for author_id 1.5. This threshold is kept simple and more improvement could be possible in this.

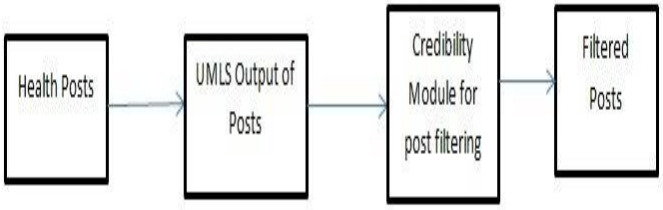

Fig 2. Module for Filtering posts

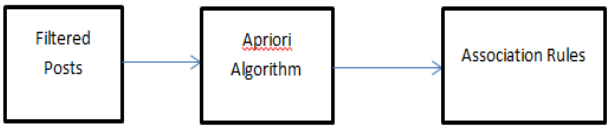

Fig 3. Association Rule Module

2) Association Module:

UMLS output posts of higher weighted users are taken into account and using Apriori algorithm associations among disease-drug-symptoms are obtained as shown in Fig. 3.

IV. EXPERIMENTAL RESULTS

We have carried out experimentation on the posts with credibility module and without credibility module. These two cases are discussed below.

1) Case 1: Association Rule with credibility module

We took 100 random posts of hydrocodone drug from healthboards.com. Fig.4 and Fig.5 shows output when this input file is processed with credibility module.

Fig. 4 Output with authors and their weights

This module removes 12 posts from input based on threshold and processes the remaining 88 posts which are found out to be more trustworthy than those 12 posts.

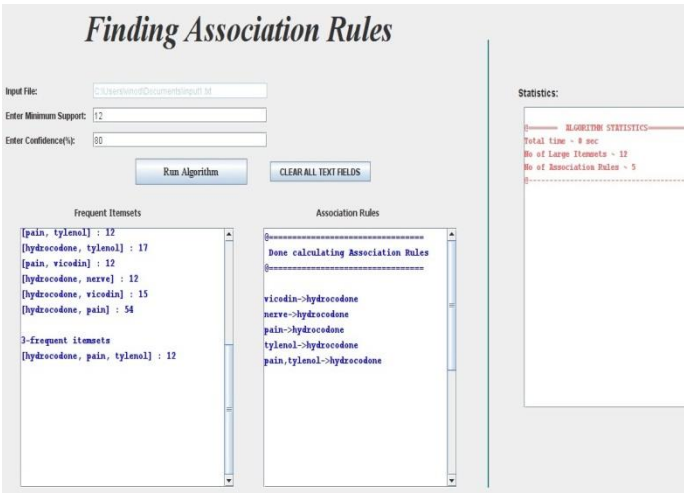

Fig 5. Association Rules

2) Case 2: Association Rule without Credibility modules

Fig. 6 shows output without credibility module in which all 100 posts are processed.

From Fig. 5 and Fig. 6, it is observed that the Association rules generated are also different in both cases. With credibility module numbers of frequent sets generated are 12 and association rules are 5. In other case, numbers of frequent sets are 32 and association rules are 7.

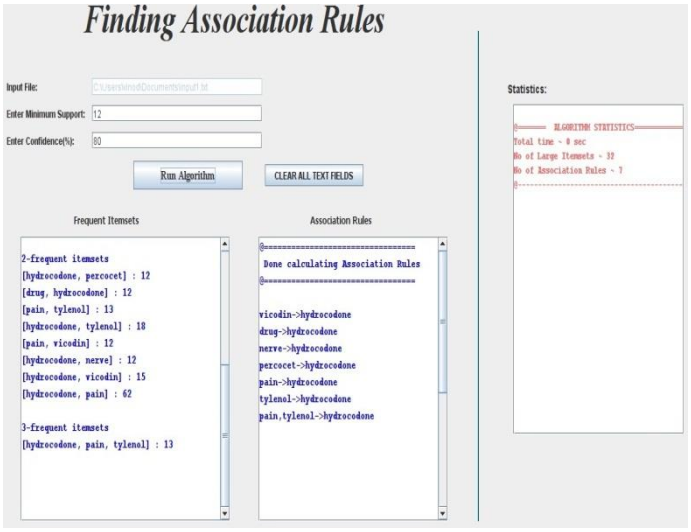

Fig. 6 Association Rules without filtering posts

We have done some analysis considering factors like number of posts, value of support, value of confidence on posts of two different drugs hydrocodone (most prescribed drug [12]) and Xanax (topmost discussed drug [13]).

We observed how number of rules generated varies with all these factors.

Table1. Analysis for posts of hydrocodone drug

Sr.No.	No.of posts	Support	Confidence	No.of Rules
1	25	3	100	12
2	25	2	100	38
3	50	6	100	8
4	50	3	50	105
5	75	8	70	9
6	75	4	80	96
7	100	12	80	5
8	100	8	70	16
9	125	15	80	10
10	150	20	90	12
11	175	30	80	12
12	200	30	80	20

Table 2. Analysis for posts of Xanax drug

Sr.No.	No.of posts	Support	Confidence	No.of Rules
1	25	4	80	1
2	25	3	60	3

3	50	5	40	7
4	50	3	80	24
5	75	8	70	2
6	75	4	80	23
7	100	12	80	1
8	100	6	40	16
9	125	10	80	11
10	150	10	70	15
11	175	10	80	20
12	200	30	80	2

As shown in Table 1 and Table 2, we have done experimentation by varying the number of posts in both the cases from 25 till 200.It is observed that lower value of support and higher value of confidence yields optimum number of rules.

V. FUTURE SCOPE

We just considered the user profiles for finding credible posts. But there is a lot of scope to find credible posts. Objectivity of language used, length of posts, geographic location of user, etc are factors needs to be considered. Again source of information plays very crucial role in getting quality information. Approach used for association rule mining is simple, it can be further advanced.

VI. CONCLUSION

False or incorrect information cannot be tolerable specially in health related contents. Thus User credibility and trustworthiness is much needed while considering large amount of such contents from internet. Our experimental results show that with credibility module rules generated are more trustworthy. Also only filtered posts are considered for processing thus reducing the overhead. One challenge is to verify the authenticity of membership type of user.

REFERENCES

[1] Mohammad-Ali Abbasi and Huan Liu, "Measuring User Credibility in Social Media", SBP'13 Proceedings of the 6th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction, pages 441-448, Springer, 2013.

[2] Ivan Ibnualim, SuhonoHarsoSupangkat, "Design Of Health Social Media To Improve The Quality Of Patient's Recovery", Cloud Computing and Social Networking (ICCCSN), 2012 International Conference on, pages 1-4, IEEE 2012.

[3]Annie T. Chen, "Patient Experience in Online Support Forums:

Modeling Interpersonal Interactions and Medication Use", Proceedings of the ACL Student Research Workshop, pages 16–22, Sofia, Bulgaria, August 4-9 2013.

[4]Kevin R. Canini, BongwonSuh, Peter L. Pirolli, "Finding Credible Information Sources in Social Networks Based on Content and Social Structure", IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 2011.

[5] Keith E. Campbell, Diane E. Oliver, Edward H. Shortliffe, "The Unified Medical Language System: Toward a Collaborative Approach for Solving Terminologic Problems", Journal of the American Medical Informatics Association ,Volume 5, Number 1, Jan / Feb 1998.

[6]Alan R. Aronson, "Effective Mapping of Biomedical Text to the UMLS

Metathesaurus: The MetaMap Program", Proceedings of AMIA, 2001.

[7]www.nlm.nih.gov/research/umls/

*[8]Vinod L. Mane, Suja S. Panicker, Vidya B. Patil, "*Knowledge discovery from user health posts", IEEE Sponsored 9th International Conference on Intelligent Systems and Control, 2014, in press.

[9]*Vinod L. Mane, Suja S. Panicker, Vidya B. Patil, "*Summarization and sentiment analysis from user health posts", International Conference on Pervasive Computing ICPC 2015 , in press.

[10]WojciechJaworski, Emilia Rejmundand Adam Wierzbicki, "Credibility Microscope: relating Web page credibility evaluations to their textual content", IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014.

[11] Jiang Yang, Scott Counts, Meredith Ringel Morris, Aaron Hoff, "

Microblog Credibility Perceptions: Comparing the United States and China", CSCW '13, February 23–27, ACM, 2013.

[12] http://preventdisease.com/news/13/021213_The-7-Most-Prescribed-Drugs-In-The-World-And-Their-Natural-Counterparts.shtml

[13] www.healthboards.com