IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 10-16
www.iosrjournals.org

Generating Association Rules for Social Media Analysis

Deepika Jaiswal[1,] Shilpa Singh[2], Suhasini VijayKumar[3]
[1] *MCA Student, University of Mumbai*
[2] *MCA Student, University of Mumbai*

Abstract: *Association Rule Mining is a technique for identifying correlation between different data sets. It is also called as Market Basket Analysis, as this was original application area of association rule mining. Also the main aim of association rule is to identify association between items that occur together from a random sampling of all possibilities from data set. Social media mining has the process to represent, analyze, and extract patterns, trends from raw data of social media. These patterns and trends are useful to companies, governments and not-for-profit organizations, as these parties can usethose patterns and trends to design their strategies or to make new programs. This paper focus about generating association rules on these social media raw data which will help organization for analyzing trends very efficiently. It will discuss about various algorithm and techniques used in generating the rules with its procedure also will be compared from other algorithm to identify the best algorithm.*

Keywords: *Apriori Algorithm, Association Rule Mining, Data mining, Social Media Mining*

I. Introduction

A web based service where people can create a public/semipublic profile on some particular domain also which can connect and communicate within that particular network is known as Social media network [1]. A collection of people or a group of individuals which have similar way of contact or interaction like friendship is known as social network. This network can be represented in a graph format which has nodes indicating individual or group connected via a link which represented as line joining them according to their relationship. The graphs can either be depicted as directed or undirected graphs dependingupon the type of relation between those link nodes.

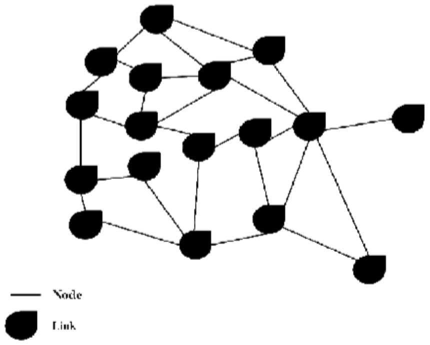


Fig. 1. Social Network showing nodes and links

Extraction of association rule which is said by [10]implies the importance of text mining field which involves identifying various association relationsfrom thewords found insome text data set. These association rules when applied on large amount of transaction database records will help in various decision making process.

There are three dominant issues in social network mining namely: size, noise and dynamism which can be handled by data mining techniques. The hugenature of social network arrangement datasets needs well programmedinformation dealing and analyzes it within a stipulated logical amount of time.Social network proves to be a good platform to mine significant patterns from large data set using data mining techniques [1]. This paper focus about generating association rules on these social media raw data which will help organization for analyzing trends very efficiently. It will discuss about various algorithm and techniques used in generating the rules with its procedure also will be compared from other algorithm to identify the best algorithm.

II. Literature Review

Social media analysis and Online social network are very popular in examine field. The majoreffort in social network research is put forth on social link [13],social connection prediction[14]. Various people across the globe also job on: Personality prediction for micro blog users [15], Using social media to expect real-world outcome[16], Predicting friendship concentration [17,18], Sentiment analysis and opinion mining [19].

Some more interesting areas of research focuses on esteem predicting social media depending on Comment mining[20], Predicting patterns of diffusion processes in social network[21].

Some research is also ongoing in identifying significant users using learning based approach or Page Rank Algorithm or adaptations of the same[21,22].

Social Media sites generate a large amount of data every minute which is shown in Fig 2.1 [11]. As per Technorati 1.2 million new post and approximate 75000 new blogs which gives opinion on a product or service is produced everyday. It's difficult for traditional methods to handle this huge data constantly generated from this site which makes it necessary to develop tools which are capable of analyzing these data. The data generated can be effectively mined by data mining techniques [3].

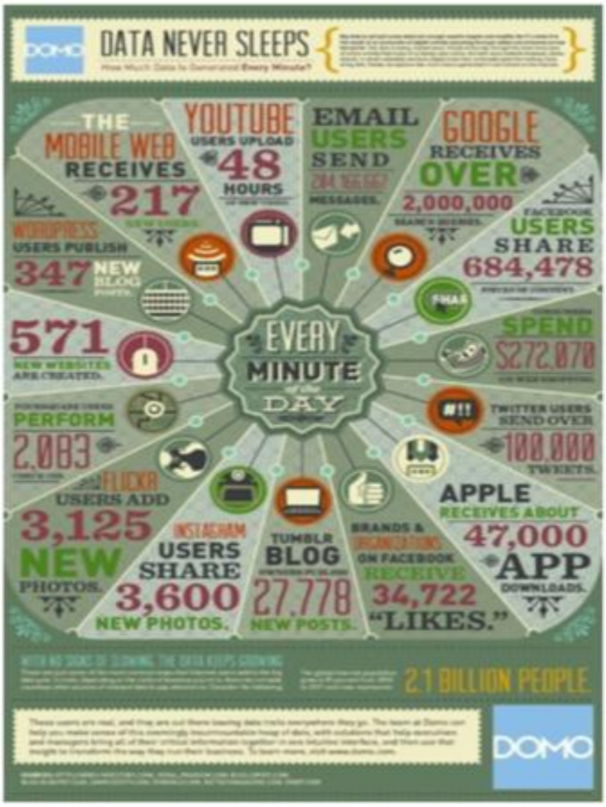
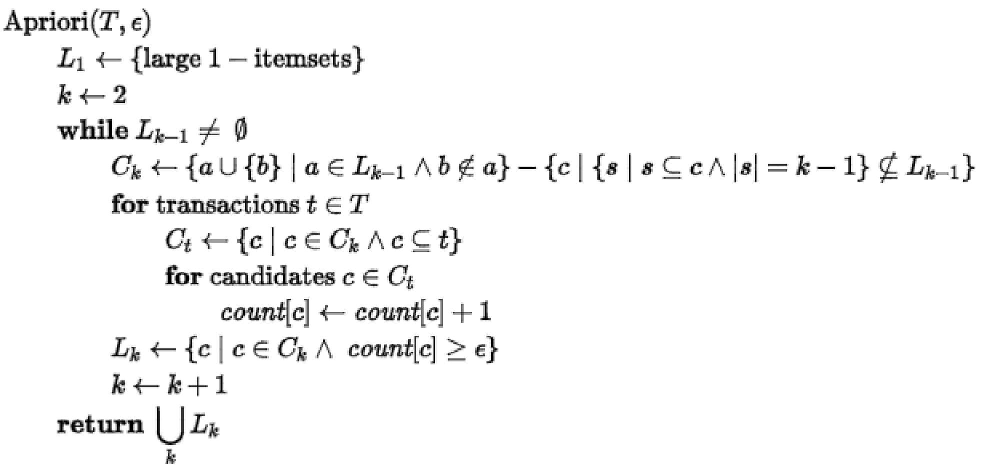

DOMO
DATA NEVER SLEEPS
How Much Data Is Generated Every Minute?
THE MOBILE WEB RECEIVES 217
YOUTUBE USERS UPLOAD 48 HOURS
EMAIL USERS SEND
GOOGLE RECEIVES OVER 2,000,000
FACEBOOK USERS SHARE 684,478
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS
571 NEW WEBSITES ARE CREATED
EVERY MINUTE of the DAY
SPEND $272,070
FOURSQUARE USERS PERFORM 2,083
TWITTER USERS SEND OVER 100,000 TWEETS
FLICKR USERS ADD 3,125 NEW PHOTOS
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS
TUMBLR BLOG 27,778 NEW POSTS
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS
2.1 BILLION PEOPLE
DOMO

Fig 2.1: Estimated data generated every minute

The method of mining association rulefocus on discovery huge item sets, which are group of items that are of same view together in a sufficient number of dealings. Usage number of association rule can be identifying if the database is large so for minimizing association rule minimum support and confident are consider, both are specifying by the user which helps us to create valuable rules from the database. The various association rule mining algorithm are

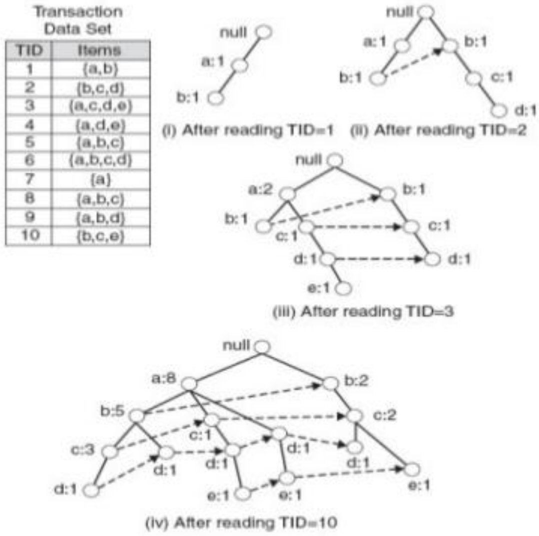
1. **AprioriAlgorithm:** It was developed by Agarwal and Srikant which was intended to operate on database which contain transactions. This algorithm use "bottom-up" technique, where all therecurrent subsetis extended one item at once and groups of suchcandidates are tested against the data. The algorithm terminates when no additional successful extensions are originated. The following Fig 2.2 shows apriori algorithm:



```
$\text{Apriori}(T, \epsilon)$
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\}$
        for transactions $t \in T$
            $C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$
            for candidates $c \in C_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

Fig 2.2: Apriori Algorithm

2. **FP (Frequency Pattern)-Growth Algorithm:** In this algorithm during the first pass, the algorithm counts the number of occurrence of objects in the dataset and saves it to 'header table'. During second pass, it constructs the FP-Tree by inserting those instance. Objects of each instance in the tree are arranged in reverse order of their lowest frequency in dataset, such that the fp-tree gets processed quickly.Objects in the tree are removed if they do not follow minimum threshold coverage. If there are many instance sharing most repeated items, FP-Tree provides high density close to the root of the tree. The only main difference in apriori and fp tree is that it does not generate separate candidate set rather it appends the header table from below by finding all instanceidenticalto the given condition. The following figure Fig 2.3shows the example of FP-Tree



Transaction Data Set

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

Fig 2.3: Example of FP Growth

The various association rules applications are market base data analysis, customer relationship management (CRM), web usage mining, bioinformatics and intrusion detection. There are two ways of computing usefulness in association rule mining that is subjective and objective. Objectives measures involve statistical analysis of the data such as support and confidence.

Support is given by the ratio of occurrence in the dataset which is denoted by Sup(X) and confidence is given by Confidence(Y=>X) = Support(YUX)/Support(Y).

III. Methodology

To generate Knowledge discovery, we are using raw data collected from social media sites. The proposed methodology is divided into three phases such as Text preprocessing, Association rule mining and Knowledge discovery phase shown in Fig3.1. In text preprocessing phase we first tokenize the input document in to tokens. Then the tokens are filtered by removing stop word as they do not carry any meaningful information. Normally token contains many suffixes and it is required to remove all the suffixes to achieve better result in knowledge discovery. Then the tokens are indexed using TFIDF (Term Frequency Inverse Document Frequency) values. The Association rule mining phase generates association rules based on weighting scheme TF-IDF that is depending on the users requirement the high frequency keywords are selected to generate association rules. The last phase is to generate Knowledge discovery using those generated association rules.

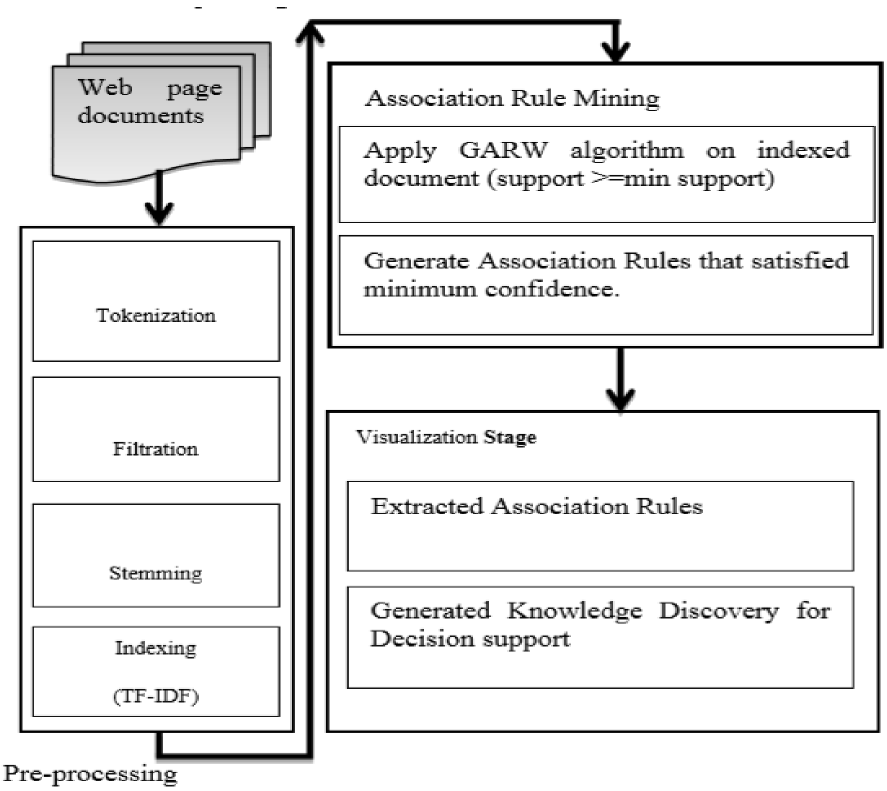

Web page documents
Tokenization
Filtration
Stemming
Indexing
(TF-IDF)
Pre-processing
Association Rule Mining
Apply GARW algorithm on indexed document (support >=min support)
Generate Association Rules that satisfied minimum confidence.
Visualization Stage
Extracted Association Rules
Generated Knowledge Discovery for Decision support

Fig 3.1: Block Diagram of Proposed System

3.1 Text Preprocessing Phase:

Since data on the web is in unstructured format also usage amount of data is generated everyday on social media, thus mining such large document it is necessary to preprocess the input document and store into structured format which can be further use for preprocessing and generating association in it. The text preprocessing phase involves the following phase shown below:

Tokenization:

Normally web page contain information in unstructured format. This create a problem for text mining and it also consume memory and time to process. So tokenization is the process of splitting the text into words. The main aim of this process is to convert unstructured document into structured document.

Filtration of keywords:

To generate accurate knowledge from the collection of raw data the user needs to find out relationship between all the keywords but this task is very tedious if we do not remove redundant and inefficient data. Finding relationship becomes very difficult if that document is not filter well, thus filtration of raw data can be done by removing stop words (does not have meaningful information) and suffixes (attached with the same word but with different form).

Stemming:

Stemming is a process in which variant form of same are reduced to common form. Stemming replace all the match suffix from the keywords with replacement character and words. The reasons for using stemming are it changes the meaning of term even main route word is same, ambiguous association rule are generated, it make data complex and occupy extra memory.

Indexing:

After all the above process the weight scheme Term Frequency, Inverse Document Frequency(TF-IDF)is use to allocate weight to distinguish expressions in the document. Frequency is the count that represent how many times of keywords has occurred in that document whereas inverse document frequency is the count that represent total number of document that contains the keywords atleast once. We have use this weighting scheme to select higher frequency keyword for generating association rule. With apriori algorithm the only disadvantage is that it consider all the keywords without knowing importance of those keywords for generating association rule.

3.2 Association rule mining stage:

Association rule is of the IF-THEN structure, but it can predict attribute combination, and they are not intended to be used together as a set. For each rule IF antecedent THEN consequent we count its support and confidence matches with user specified values. Support is the possibility that a randomly selected instance will fulfill both the predecessor and successor, and confidence is the conditional possibility that a randomly selected instance will fulfill the consequent given that the instance fulfill the antecedent. In our developed system we have generated only those association rules which satisfy criteria such as support, confidence, and TFIDF value of keywords. The algorithm called as GARW (Generating Association Rule using Weighting Scheme) is found to be better than that of conventional Apriori algorithm. The GARW algorithm works in similar way of Apriori but with some additional steps to resolve problem of Apriori and to generate relevant association rules.

GARW Algorithm

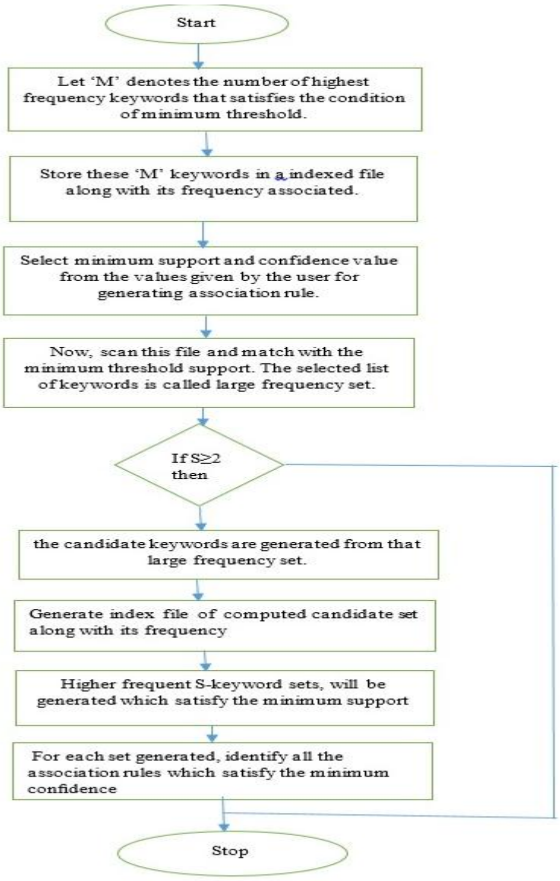

Start
Let 'M' denotes the number of highest frequency keywords that satisfies the condition of minimum threshold.
Store these 'M' keywords in a indexed file along with its frequency associated.
Select minimum support and confidence value from the values given by the user for generating association rule.
Now, scan this file and match with the minimum threshold support. The selected list of keywords is called large frequency set.
If S≥2 then
the candidate keywords are generated from that large frequency set.
Generate index file of computed candidate set along with its frequency
Higher frequent S-keyword sets, will be generated which satisfy the minimum support
For each set generated, identify all the association rules which satisfy the minimum confidence
Stop

Fig 3.2: GARW Algorithm

3.3 Visualization Stage:

Thus implementing all the above phases on raw data collected from social media generates a collection of association rules between the highest occurring keywords. Now from this generated association rules only the rules which are applicable on the area of interest depending upon the organization's programme is selected and analysis is done .This extracted association rule can be represented as graph, table or in a textual format helping in the decision making process.

IV. Result and Discussion:

We have implemented this system using R language on some dummy text files. The system involved phases such as text preprocessing where text file was filtered by removing all the stopwords and stemming the keywords that is reducing the same word to its common form. Later an indexed file was created containing list of keywords along with the frequency associated to it. Using this indexed file a word cloud was created by the system depicting the association between those keywords. The following Fig 4.1 shows the word cloud generated by the system showing keywords having higher frequency and Fig 4.2 shows the word dengue is having frequency greater than 12 which is generated by the system as a bar chart.

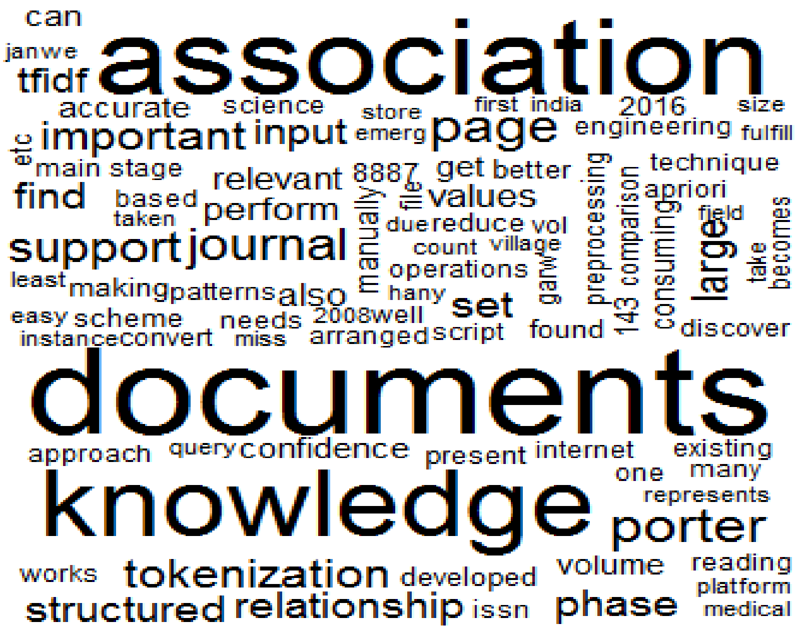


Fig 4.1: Word cloud created using highest frequency words

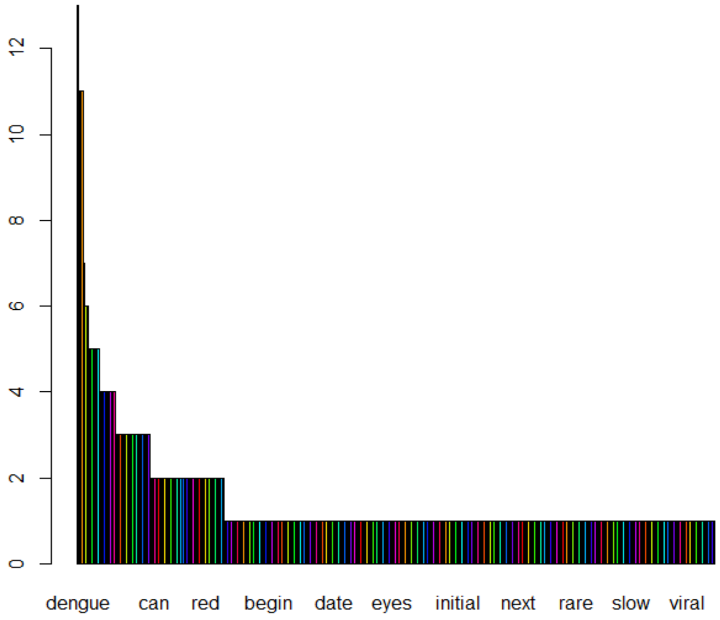


Fig 4.2: Bar plot showing highest frequency words

Hence this system can also implemented by collecting the user opinion or suggestion upon some idea or programme proposed by the organization over social media into a text file and applying the above methodology for the generation of association rule. This will help organization to identify the user support and for making strategic decisions.

V. Conclusion

As per our review there are many algorithms by which we can generate the association rule such as Apriori algorithm, FP-Growth Algorithm and GARW Algorithm. The best and effective algorithm among all three is GARW algorithm because it is based on weighting scheme having faster performance by reducing execution time whereas Apriori generates separate pair for each set, which consumes huge amount of memory and FP-growth algorithm appends the pair to the tree using frequency. Thus in this paper we proposed a system which uses GARW Algorithm having faster execution and better visualization of the extracted association rules.

References

[1]. Mariam Adedoyin-Olowe, Mohamed Medhat Gaber and Frederic Stahl A Survey of Data Mining Techniques for Social Network Analysis

[2]. Mariam Adedoyin-Olowe, Mohamed Medhat Gaber and Frederic Stahl A Survey of Data Mining Techniques for Social Media Analysis

[3]. Anu Sharma, Dr. M.K Sharma & Dr. R.K Dwivedi Literature Review and Challenges of Data Mining Techniques for Social Network Analysis

[4]. Frequent Pattern Mining[Book]

[5]. Said A. Salloum , Mostafa Al-Emran , and Khaled Shaalan Mining Social Media Text: Extracting Knowledge from Facebook . In Proceedings of the International Journal of Computing and Digital Systems ISSN (2210-142X) Int. J. Com. Dig. Sys. 6, No.2 (Mar-2017)

[6]. Chang Zhang , Yanfeng Jin, Wei Jin1, Yu Liu1 Study of Data Mining Algorithm in Social Network Analysis In 3rd International Conference on Mechatronics, Robotics and Automation (ICMRA 2015)

[7]. Anurag Agrahari , Prof D.T.V. Dharmaji Rao Association Rule Mining using RHadoop . In Proceedings of the International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 04 Issue: 10 | Oct -2017

[8]. Dr. R Nedunchezhian, K Geethanandhini Association Rule Mining on Big Data – A Survey . In Proceedings of the International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 5 Issue 05, May-2016

[9]. M. Vedanayaki A Study of Data Mining and Social Network Analysis In Proceedings of Indian Journal of Science and Technology, Vol 7(S7), 185–187, November 2014 ISSN (Print) : 0974-6846 ISSN (Online) : 0974-5645

[10]. Rakesh Agrawal ,Tomasz Imielinski, Arun Swami Mining Association Rules between Sets of Items in Large Databases

[11]. Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acmsigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.

[12]. Tepper, A.: How Much Data Is Created Every Minute?[INFOGRAPHIC]. 2012, http://mashable.com/2012/06/22/datacreated-every-minute/. Retrieved on 16/102013 at 19.00.

[13]. Liben-Nowell, D.; Kleinberg, J. The Link-prediction Problem for Social Networks. J. Am. Soc. Inf. Sci. Technol. 2007, 58, 1019–1031.

[14]. Utz,S.;Jankowski,J. Making"Friends"inaVirtualWorldTheRoleofPreferentialAttachment,Homophily, and Status. Soc. Sci. Comput. Rev. 2015, doi:10.1177/0894439315605476S.

[15]. Asur,S.;Huberman,B.A. PredictingtheFuturewithSocialMedia. In Proceedings of the 2010IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology–Volume 01; IEEE Computer Society: Washington, DC, USA, 2010; pp. 492–499.

[16]. Ahmad,W.;Riaz,A.;Johnson,H.;Lavesson,N. PredictingFriendshipIntensityinOnlineSocialNetworks. In Proceedings of the 21st Tyrrhenian Workshop on Digital Communications: Trustworthy Internet; Springer: Berlin/Heidelberg, Germany, 2010.

[17]. Nia, R.; Erlandsson, F.; Johnson, H.; Wu, S.F. Leveraging social interactions to suggest friends. In Proceedingsofthe2013IEEE33rdInternationalConferenceonDistributedComputingSystemsWorkshops (ICDCSW), Philadelphia, PA, USA, 8–11 July 2013; pp. 386–391.

[18]. Petz, G.; Karpowicz, M.; Fürschuß, H.; Auinger, A. Reprint of: Computational approaches for mining user's opinions on the Web 2.0. Inf. Process. 2015, 51, 510–519.

[19]. Jamali, S.; Rangwala, H. Digging Digg: Comment Mining, Popularity Prediction and Social Network Analysis. In Proceedings of the International Conference on Web Information Systems and Mining, (WISM 2009), Shanghai, China, 7–8 November 2009; pp. 32–38.

[20]. Jankowski, J.; Michalski, R.; Kazienko, P. The Multidimensional Study of Viral Campaigns as Branching Processes. In Social Informatics; Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7710, pp. 462–474.

[21]. Weng, J.; Lim, E.P.; Jiang, J.; He, Q. TwitterRank: Finding Topic-Sensitive Influential Twitterers. In Proceedings of the Third ACM International Conference on Web Search and Data Mining; ACM: New York, NY, USA, 2010; pp. 261–270.

[22]. Hotho, A.; Jäschke, R.; Schmitz, C.; Stumme, G. Information Retrieval in Folksonomies: Search and Ranking. In The Semantic Web: Research and Applications; Springer: Berlin/Heidelberg, Germany, 2006; pp. 411–426.