Hybrid Clustering for Identification of Distinct Topics of a Domain using User Influence Pattern

DwarapuSuneetha, MogallaShashi

Abstract: ***Content based tweet clustering is extensively used for automatic topic identification of tweets in social media analytics. However due to restrictions on the length of the content in social media platforms like twitter mere content is not enough to provide sufficient information for clustering. In this paper the authors proposed to enhance the clustering quality by adding tweeting behavior of influential users. Spearmen correlation is appropriately adapted for identifying mergeable clusters. A new methodology for hybrid clustering is proposed and tested using entropy on real data related to three domains namely sports, politics, and health. The proposed method achieved distinct cluster formation which is reflected by reduced entropy after applying merging based on user influence patterns.***

Key words: Tweet clustering, entropy, influence patterns, hybrid clustering

I. INTRODUCTION

The internet is connecting more and more people across the world through social networking. Social media sites like "Twitter" provides a platform where in people share information related to events happening all over the world. On twitter every user is identified with a unique account number, with this unique ID a user can become a "follower" with the liberty to choose whom he/she can follow, in similar lines he/she becomes "followee" also when followed by other users. Twitter users post "tweets" to share their views which are displayed in authors profile page and becomes visible to his/her followers. Some of the followers may "retweet" the tweets they liked. The twitter world has become an avenue to many well-known brands, athletes, teams, celebrities and many big organizations to stay connected and propagate their opinions and impressions. Twitter platform has 300 million monthly active users. Among the active users, lies a category called "Influential users", whose tweets are responded by a huge number of retweets from their followers. The followers may in turn retweet that tweet which creates a snowball effect; by more and more people reading their tweets. These influencers spread news and express their opinions on current events and/or happenings of their interest which then pave the way for a wider reach and a faster spread of breaking news, which, more often than not, is quicker than traditional news and media outlets.

In a wide network such as Twitter, a large number of tweets, encompassing various domains are posted by these social influencers. However, a specific influential user is prone to confine to tweet on a subset of distinct sub-topics of the domain. For example in a broader domain like sports the events related to cricket may receive tweets from a separate set of influential users who are different from those users posted on events related to tennis. Though some of the influential users post tweets on multiple sub-topics their intensity/level of influence in different sub-topics varies unless the sub-topics are not clearly distinct. This observation referred to as the "tweeting behavior of influential users" has lead to the development of a hybrid methodology for automatic sub-topic identification of tweets belonging to a domain based on the content of the tweet as well as tweeting behavior of influential user.

The rest of the paper is organized as follows. Section 2 describes related work section 3 discusses the frame work section 4 and 5 about the data sets we have used and the results observed during the framework and finally section 6 concludes the paper.

II. RELATED WORK:

2.1 Measuring user influence in Twitter: The million Follower Fallacy[1] 2010: As with evolving information environment in social networking sites like Twitter, etc , the dynamics of influence have begin to emerge quickly. Studying influence patterns is useful for understanding how business operates and how society functions and to get knowledge about certain trends and innovations and also helps advertisers and marketers to build effective campaigns. Me young Cha et.al., measured the influence of a user in 3 ways namely in degree, retweets and mentions based on tweets posted in twitter data set.

2.2 On summarization and timeline generation for evolutionary tweet streams [2] : The authors in the paper proposed a novel continuous summarization frame work called SUMBLR to deal with dynamic and fast arriving and large scale tweet streams. It contains three components; In the first component they designed an algorithm namely online Tweet Stream Clustering algorithm to maintain distilled statistics and store it in a data structure called as Tweet Cluster Vector (TCV) and in the second step they designed TCV rank summarization technique for generating summaries. In the last step they effectively designed topic evolution detection method to detect distinct topics of a domain.

*2.3 Topic discovery and future trend forecasting for texts [7] :*The authors in this paper proposed a framework to

Revised Manuscript Received on December 08, 2018.

DwarapuSuneetha, Assistant Professor, Department of CSE, GITAM institute of technology , Visakhapatnam , India

MogallaShashi, Professor, Department of CS & SE, ANDHRA UNIVERSITY, Visakhapatnam , India

automatically discover topics from a set of documents and forecast their evolving trend by considering data mining and machine learning as data domains. An association analysis process is applied followed by temporal correlation analysis and ensemble forecasting approach in order to identify the set of topics, discover correlation between the topics and analyze the popularity of the topics in future. Their frame work yields better performance and it is helpful to express large scale text collections in concise form. The frame work is also beneficial for many applications such as modelling the evolution of the direction of research for forecasting future trends of IT industry.

III. METHODOLOGY:

Tweet clustering is widely used for automatic sub-topic identification [2][4][7] [18][19]. Most of the method for tweet clustering relies on the content of the tweets. However, the content of the tweet may not accurately represent the topic as it may contain only a small number of words with some of them in short forms (when the meaning is indicative) due to length restrictions. More accurate topic identification is possible if additional information like tweeting behavior of users is also used in addition to the content of the tweets for clustering. This paper explores the usage of tweeting behavior of users for merging the clusters originally formed based on the content of the tweets and a framework is developed to investigate the effectiveness of the proposed hybrid clustering.

As influential users vary from topic to topic, sub-topic based influence scores are estimated for influential users based on the proportion of the number of retweets obtained for their tweets on a topic to the number of tweets they posted on the topic. The influential users of each sub-topic are ordered in the descending order of their influence scores and pairs of sub-topics sharing considerable number of common users with their influence order agreeing with each other are considered similar. Similarity between pairs of subtopics are measured to understand whether users with a high influence score in one sub topic tend to have a high score in another subtopic. The similarity of a pair of sub-topics can be measured with a rank correlation coefficient.

Spearmen rank correlation method was developed by a British psychologist namely Charles Edward Spearmen in the year 1904. It was originally proposed to find the correlation between a pair of quantitative measures to determine whether they assess distinct aspects of an entity or not in the context of selecting non-redundant set of measures for evaluation of entities. In this paper, Spearman correlation is used to check the consistency of tweeting behavior of influential users on specific sub-topics represented by the clusters and accordingly estimate the similarity of sub-topics to determine whether they are mergeable clusters.

The proposed framework for hybrid clustering to identify distinct topics of a domain is shown in Figure 1. It contains two modules in cascade as detailed below

Module 1: Identification of influential users in a domain and cluster their tweets based on content.

Module 2: Identification of mergeable clusters based on tweeting behavior of influential users to form distinct/merged clusters corresponding to the topics of the domain.

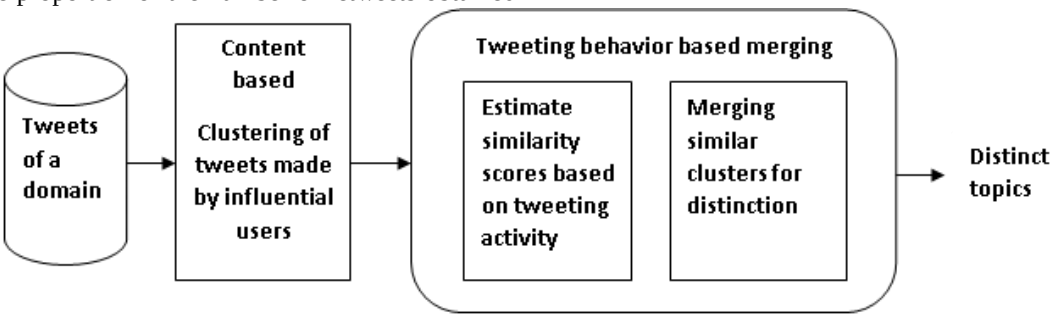

Figure 1: The framework of hybrid clustering to identify distinct topics of a domain

3.1: Identification of influential users in a domain and content based clustering of their tweets

The first module of the framework identifies "influential users" as they play an important role in spreading the information by posting tweets that impress others and their tweets are frequently reposted by others. This research considers only retweeted tweets to identify influential users. The retweeted tweets and the users whose tweets are retweeted on a particular domain are identified by a twitter search API. The users thus obtained are known as active users. Among them influential users are identified by applying log transformation on the number of retweets obtained by them and comparing it with a threshold value fixed at 1.5 .

Tweets made by Influential users are collected and preprocessed for content analysis; each tweet is represented as a tweet term vector specifying the prominence of a word in a tweet in terms of TF*IDF scores, where TF is the term frequency i.e. number of times a word occurs in the tweet and IDF is inverse document frequency which measures how specific a word is in the corpus. K-means clustering algorithm is applied on the tweet term vectors to form k-clusters which may be interpreted as k sub-topics. However due to length limitations on tweets and their context sensitiveness, content based clustering has limited capability to identify topics when applied on tweets rather than documents though it was widely accepted in document retrieval.

3.2: Identification of mergeable clusters based on tweeting behavior of influential users and topic identification:

The hybrid clustering of tweets based on content as well as tweeting behavior of influential users is proposed by the authors for automatic identification of topics from the tweets of a domain. Once k-clusters are formed based on the content, each cluster is interpreted as a topic and the local influence score of each influential user is estimated. The author proposes that the tweeting behavior of influential users has a pattern whose manifestation can be observed in terms of correlation in influence score ordering of common users contributing to similar topics. In other words, a close correspondence can be observed when the common influential users of two similar clusters are ordered in accordance with their decreasing local influence scores in the clusters separately. Hence such clusters are mergeable in the process of identifying distinct topics of a domain and the effect of merging should be propagated. In order to capture such correspondence between mergeable clusters the authors devised a new metric for similarity estimation named Similarity Score (SS) for a pair of clusters based on tweeting behavior. The second module of the framework implements the whole process in the following steps:

a)Estimation of local influence score for each influential user on k topics:

For each cluster obtained in module 1, the number of tweets made by each user is identified and the number of retweets obtained from their followers is counted. Influence scores for each influential user is estimated as the ratio of the number of retweets obtained to the total number of tweets posted by the influential user on the sub-topic in the specified time duration. The formula for score estimation is given below.

$$\textit{Local Influencescore}\ (u,t) = \frac{\#\ \text{Retweetsobtained}}{\#\ \textit{Tweetsposted by u on the topic t}} \text{---------} \quad (1)$$

Some influential users will tweet on multiple topics and their local influence scores may differ topic-wise.

b) Estimation of Similarity Score of pair of topics based on their tweeting behavior:

K sub-topics and the influential users who posted tweets on these topics are identified and their influence scores were calculated on each sub-topic. The second step is to measure the Similarity Score (SS) between the identified k clusters in pairs. The product of Jaccard Similarity and Spearmen Rank Correlation between the pairs of clusters gives the Similarity Score.

$$\mathbf{SS\ (C_i,C_j) = JC\ (C_i,C_j) * SR(C_i,C_j)} \text{ ----} \quad (2)$$

"Jaccard Similarity Coefficient" denoted by JC (C_i,C_j) measures the relative overlap between the clusters/topics CiandC_j i.e. the number of common users over the total number of users.

$$\mathbf{JC\ (C_i,C_j)} = \frac{|\ Ci \cap Cj\ |}{|\ Ci \cup Cj|} \text{------} \quad (3)$$

Where C_i and C_j are the set of users in i^{th} and j^{th} cluster.

The percentages of Jaccard Similarity Coefficient are calculated for each pair and are compared with the threshold value (0.1) to obtain clusters with considerable overlap. If the percentage of Jaccard Similarity between the two sub-topics C_i,and $C_j$is less than threshold value, implies that the sub-topics have no sufficient number of common users to merge and hence such pairs are eliminated from merging process. If percentage of JC(C_i,C_j) is more than the threshold value, such pairs will be promoted for further analysis to estimate the SS.

The next step is to check the correspondence between the tweeting behaviors of influential users within the promoted pairs of topics which is measured by "Spearmen Rank Correlation Coefficient". Once the common users of the pair of topics are ranked in descending order of their local influence scores separately, assuming ranks in one cluster as Xi and ranks in another cluster as Yi, Spearmen rank correlation coefficient for a pair of topics with N common users is estimated using the formula

$$\mathbf{SR\ (C_i,C_j) = 1 - 6 \sum (X_i - Y_i)^2 / (N^3 - N)} \text{---} \quad (4)$$

The estimation of Similarity Score (SS) for each pair is depicted in algorithm 1.

Algorithm 1: Estimate SS(C_i , C_j)

Input: A pair of clusters C_i and $C_j$from K*K clusters

Output: Estimation of Similarity score for each pair C_i and C_j

Begin:

Assign a variable L to store number of clusters i.e. L=K

For each pair of clusters C_i and C_j

Identifying common users

Calculate Jaccard similarity coefficient as percentage JC (C_i , C_j)

If JC (C_i , C_j) >0.1 (Threshold value)

Calculate Spearmen rank correlation SR (C_i , C_j)

Estimate similarity score SS(C_i , C_j)

Else

Display "No sufficient number of common users for merging"

"

c) Identification of distinct topics by hybrid clustering

The pair-wise similarity scores are estimated for each pair of clusters out of which only the pairs whose SS (Ci ,Cj) value is greater than 0.5 are considered for merging by storing them in a descending priority queue based on SS values. Variable L is used to keep track of new clusters being generated by merging pairs of similar clusters, which is initialized to the total number of clusters K and incremented whenever a new cluster is formed. The next step is to delete a pair of most similar clusters Ci and Cj from the

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

queue to merge them as a single cluster named CL. The local influence scores of common users in the new cluster CL are updated using equation 5 whereas the scores of distinct users remain the same.

$$Influencescore(commonusers) = \frac{(\#\text{Retweets in first cluster}) + (\#\text{Retweets in second cluster})}{(\#\text{Tweets in first cluster}) + (\#\text{Tweets in second cluster})} \quad \text{----- (5)}$$

The next step is to propagate the result of merging along the other nodes of the priority queue by finding the pairs of clusters involving either $C_i$or C_j paired with the other cluster Cm to form a merged clusters (C_L ,C_m) and estimate the influence score of its user again as described above. In other words every (C_i ,C_m) or (C_j ,C_m) are to be replaced by (C_L,C_m) and reinserted in appropriate place of the priority queue based on similarity score estimation. The merging process described above is applied on successive nodes of the priority queue until the next pair to be deleted involves a newly generated cluster which is indicated based on its index greater than k. This framework generates 2k-L clusters which are interpreted as distinct topics of the domain. The process of identification and merging the clusters is presented as algorithm 2 which calls Algorithm1 for estimating Similarity Score of a pair of clusters. Figure 2 depicts the pictorial representation of algorithm 2.

Numeric Illustration: Once the content based clustering generates k clusters/topics say14 for sports domain, the influence score of each influencer in each sub-topic is estimated. In order to assess similarity among sub-topics pair-wise, Jaccard Similarity Coefficient is calculated for every pair of clusters to form 14*14 matrix to obtain relative overlap of users. Those pairs whose percentage of Jaccard similarity is greater than the threshold value (0.1) are qualified for similarity estimation. Accordingly Spearmen Rank Correlation coefficient and Similarity Score are calculated for 29 pairs. Similarity Scores are stored in a descending priority queue. In our experimentation the pair of topics 3 and 9 are found to be the most similar. So combining the two topics as single topic and named the newly formed merged cluster as $(k+1)^{th}$i.e 15^{th} cluster. Influence scores of common users are updated in 15^{th} cluster. If the pair of subtopics involves newly merged cluster then repeat the process of calculating Jaccard Similarity percentages, Spearmen Rank Correlation, and similarity scores for those pairs. Then the highest score in descending priority queue is observed for the pair 2 and 11. So 2 and 11 sub-topics are merged together and stored in newly created cluster 16. In the 3^{rd} iteration the highest similarity score is obtained for 1 and 13 and they are merged to form a new cluster 17. This process is terminated since next node in the priority queue involves newly created node. In our experimentation we found that 3 pairs of clusters are deleted so 6 clusters are deleted from the table and 3 clusters are created. So after merging we have 14-6+3 or 2K-L (2*14-17)

i.e 11 clusters.

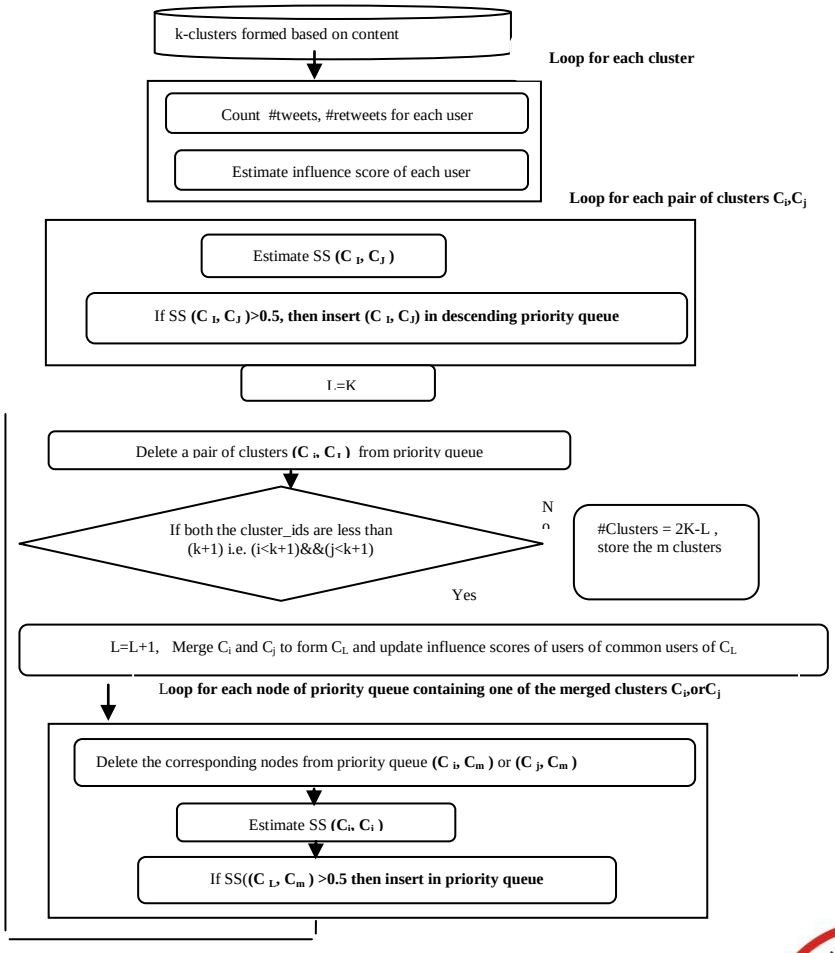

Figure 2: Flowchart for tweet behavior based merging

Algorithm 2: For identification of mergeable clusters to form distinct clusters

Input: K clusters formed based on content

Output: Distinct/merged clusters

Begin: Collecttweets of a domain in a specified time duration

Divide tweets into K clusters using any partition algorithm (k-means)

For each cluster in k

Count the **#tweets** posted by each user and **#retweets** obtained from their followers. Calculate **influence score** for each user

For every pair of clusters C_i, C_j

Estimate similarity score **$SS(C_i , C_j)$** using **Algorithm 1**

If **$SS(C_i , C_j) > 0.5$** then insert the pair $<C_i , C_j>$into descending priority queue

Initialize the variable **L** with **K**

Label:

Delete a pair of clusters **(C_i , C_j)** from priority queue

Increment L value and merge C_i, C_j to form C_L

Update the influence scores for its users

If **i<k+1 && j<k+1** (both the cluster_ids in priority queue are less than (k+1))

If each node in priority queue containing one of the merged clusters **C_i, C_j**

Delete the corresponding node from priority queue i.e. **(C_i, C_m) (C_j, C_m)**

Estimate similarity score for **(C_L,C_m)** using **algorithm 1**

If **$SS(C_L,C_m) > 0.5$** then insert it into priority queue

Goto Label

Else

Store **2K-L** clusters and stop the process

IV. DATA SET:

A data set of one lakh retweeted tweets related to each of the selected domains namely sports, politics and health based on the corresponding hashtag were collected by continuously querying twitter search API. The influential users in the selected domains are identified based on number of retweets obtained for the tweets posted by each active user domain-wise separately. For example in sports domain the number of tweets made by 75350 active users were counted along with the number of retweets obtained from their followers. The weight of each user is calculated by applying logarithmic transformation on the number of retweets obtained and those active users with weight greater than 1.5 are considered as influential users. In our experimentation 63312 social influencers are identified in sports domain. #tweets made by the social influencers are 87536. The tweets are converted from text format to numeric format by pre processing. K-means clustering algorithm is applied on different values of k. To fix the k-value Mean Squared Error(MSE) is calculated for each value of k to observe the elbow point; elbow point is the value of minimum k value corresponding to near zero rate of decrease in MSE with respect to k. $\boldsymbol{MSE} = \frac{\mathbf{SSE}}{\mathbf{n}}$where SSE is the sum of squared error and n is the number of tweets. Accordingly k value is fixed at 14 for sports and formed 14-clusters which imply that the social influencers expressed their opinions on 14 sub topics of sports.

V. EXPERIMENTAL RESULTS:

Mean entropy is used to check the quality of the clusters before and after merging the clusters.

Higher values of entropy indicates the affinity of users to post tweets on multiple sub-topics covered by more number of clusters which indicates that the separation of clusters is not clear enough. Since the aim of clustering the tweets is to partition them into distinct sub-topics, better quality of clustering is indicated by lower values of entropy.

Entropy for the clustering solution is calculated as the sum of entropies of individual users using the formula given below:

Entropy for clustering solution

$$H(U) = \sum_{u \in U} \mathrm{H(u)}$$

Entropy is calculated for each influential user before and after merging using the formula.

$$\mathrm{H(u)} = -\sum \mathrm{Pi} * \log \mathrm{P_i}$$

$$\mathrm{P_i} = \frac{\#\ \text{Tweets made by user u in i th cluster}}{\#\ \text{Total number of tweets made by user u}}$$

Mean entropy is estimated by dividing the entropy of cluster solution H(U) by the number of clusters. It may be noted that the number of clusters before merging is K and after merging is 2K-L where L is the index of the last merged cluster. The value of mean entropy of cluster solution thus obtained before merging is 877 and after merging is 373 which indicates reduced entropy due to merging of clusters appropriately. The mean entropy is tested for another two domains namely "politics" and "health" before and after merging. The value of mean entropy of cluster solution for political domain is reduced from 547 to 314 and the entropy for health domain is reduced from 636 to 419 as a result of merging. Figure 4 shows the mean entropy values for three domains namely "sports", "politics" and "health" before and after merging. Hence the proposed hybrid methodology could generate better quality clusters to capture distinct topics of a given domain using the content of the tweets as well as the users tweeting behavior.

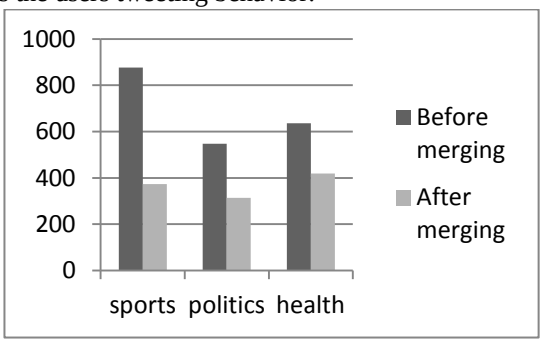

Figure 3. Mean entropy values of different domains

VI. CONCLUSION:

Twitter is a micro blogging service that has quickly emerged as a dominant social networking service for sharing facts, opinions, and ideas in real time. Tweeter's popularity and the huge amount of data being generated every hour attracted the attention of many data science researchers. Unlike document clustering due to the limitation on the length of tweets and their context sensitivity, the results of content based clustering of tweets may not be used for Automatic topic detection of tweets. In this paper the authors proposed a hybrid clustering method to enhance the quality of content-based clustering by integrating tweeting behavior of influential users. The authors empirically analyzed the tweeting behavior of social influencers of a domain by appropriately adopting established metrics like Jaccard Similarity, Spearman Correlation Coefficient and evaluated the results of topic identification using information theoretic Entropy.

REFERENCES:

1. Meeyoung Cha., HamedHaddadi.,FabricioBenevenuto., Krishna P..Gummadi. "Measuring User Influence in Twitter: The Million Follower Fallacy". Proceedings of the Fourth international AAAI Conference on Weblogs and Social media
2. Zhenhua Wang., LidanShou., Ke Chen., Gang chen., and sharadmehrotra, "On summarization and timeline generation for evolutionary tweet streams", IEEE Transactions on Knowledge and data engineering, Vol 27, No.5 May 2015.
3. Hungyuncai., Zi Huang., Diveshsrivatava., and Qing zhang, "Indexing evolving events from tweet streams",
4. VasiliiA.Gromov., and Anton .S.Konev, "Precious identification of popular topics on twitter with the employement of predictive clustering", Neural Comput&Applic ,DOI 10.1007/s00521-016-2256-1.
5. Lei Tang.,andHuan Liu, " Leveraging social media networks for classification", Data Mining Knoweledge discovery, DOI 10.1007/s10618-010-0210-x
6. Yi-chen Lo., Jhao-Yin-Li., Mi-YenYeh.,Shou-de Lin., and Jianpei," What distinguishes one from its peers in social networks " Data mining and knowledge discovery(2013) 27:396-420 DOI 10.1007/S 10618-013-0330-1.
7. Jose L.Hurtado., AnkurAgarwal., and Xingquan Zhu, " Topic discovery and future trend forecasting for texts", Journal of big data(2016)3:7,DOI 10.1186/s40537-016-0039-2.
8. Ryosuke Nishi., Taro Takaguchi., Keiguoka.,TakanoriMachara., and NaokiMasuda, " Reply trees in twitter:data analysis and branching process models, Social Network Analysis Mining (2016) 6:26, DOI 10.1007/s 13278-016-0334-0.
9. Wayne Xin Zhao., Sui Li Yulan He., Edward Y.Chang., Ji-rong wen., and Xiaoming li, "Connecting social media to e-commerce: cold-start product recommendation using micro blogging information", IEEE Transactions on knowledge and data engineering,Vol 10,No.10 XXX 2016.
10. Eva Garcia Martin., Nikalas., and Mina Doroud, " Hashtags and Followers", Social Networks Analysis Mining(2016) 6:12, DOI 10.1007/s13278-016-0320-6
11. Wenjun Wang., and W.Nick Street., "Modelling Influence diffusion to uncover influence centrality and community structure in social networks, Social Network Analysis Mining (2015) 5:15 DOI 10.1007/s13278-015-0254-4.
12. XufeiWang.,Lei Tang.,HuanLiu.,AND Lei Wang," Learning with multi resolution overlapping communities"Knoweledge and information systems(2013) 36;517-535 DOI 10.1007/s0115-012-0555-0
13. FabinRiqueine and Pablo Gonzalez- cantegiani, "Measuring User Influence on twitter: A Survey" ar XIV:1508.07951V2[CS.S1]

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication