Topic categorization on social network using Latent Dirichlet Allocation

S.S.Ramyadharshni  
PG Student, *Department of computer technology*   
anna university, MIT campus  chennai, india  
ramyasakthi05@gmail.com

**Abstract:**

**Topic modelling is a powerful technique for analysis of large document collection. Topic modelling is used for finding hidden topic from the collection of document. In the twitter api, it is essential all the tweet documents are properly categorized. For automatically categorizing the twitter document topics The efficient detection is modelled by an LDA method for probabilistic model and for separation of words from the document. LDA is widely used to estimate the multinomial observation and each topic is categorized by a probabilistic distribution over the words. The multinomial distribution of the topics is regarded as the feature of the document. The proposed system resulted in an increase in accuracy for detection of the topic categorization.**

***Keywords: LDA, topic model, multinomial distribution, probabilistic distribution.***

I INTRODUCTION

Social networks is an online platform that allow user to communicate with each other and also share their information through the internet using a computer, tablet, mobile phone etc. which is also used to posting information, comment the post, message to friends etc. Images, video also shared through the social network which is very useful for spreading the information quickly. Social network and micro-blogging site have become dynamic and widely used media for communication purpose. By using these sites user can share their information on various topics.

Social network analysis has emerged as a key technique in modern sociology. It has also gained a significant role in the following fields anthropology, biology, demography, computer science, history, geography, communication studies economics, political science, developmental studies, and social psychology.

An important process in natural language processing is to determine the category of documents in a process of text classification[1]. The techniques of text classification has the wide

Dr.P.Pabitha

Assistant professor, *Department of computer technology*   
anna university, MIT campus chennai, india  
[pabithap@gmail.com](mailto:pabithap@gmail.com)

range of applications in areas such as web, teaching, classifying mails whether spam or not.The process of extracting interesting information and the knowledge from unstructured text from a large text files and sentences is called as text mining[2]. Usually, the occurrence of a term in a document is measure by IDF and Term Frequency-IDF. The topic similarity measure defines the number of keywords that are similar present in a document.Information diffusion prediction aims at predicting the users who will spread information. The information diffusion probability calculation has its own areas of applications such as crowdsourcing, rumor diffusion, army, government and many. As there has been an enormous increase in network size and the interaction frequency among the users for an effective generalization and efficient inference[3]. Most of the information diffusion models that have been proposed has observed the structure of the network and interactions between the users. In accordance with the time the models have not been analyzed.[4] It is clear from the data that there is a need for detecting the most influential user in a group of network. A proposed mechanism which is designed to detect the influential user detection based on the attributes of the user and the network and the evaluation of the proposed system is done by compiling metric valuation.

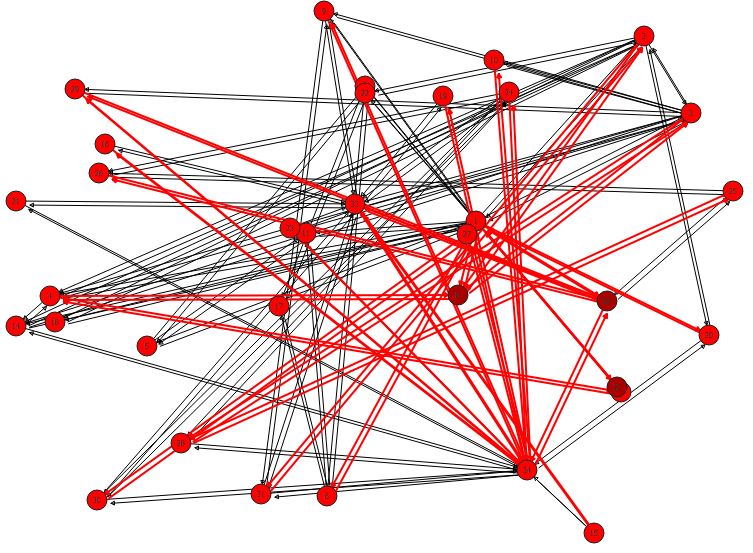


Fig 1.1 Social network

II RELATED WORKS:

K. Zhang et. al proposed to Predict retweeting using probabilistic matrix factorization technique. convert retweeting behaviour problem to a matrix factorization problem. Message semantic embedding information is employed in designing a semantic regularization term to constrain the matrix factorization objective function the technique used are Clustering algorithm Gradient descent Algorithm Probabilistic Matrix Factorization the advantage of this methods introduced above fail to discover these intrinsic geometric structure of the message embedding space. To deal with this limitation to introduce message semantic embedding and assume that messages can be divided into a number of semantic group [5]. Wei Chen et al. proposed that the inﬂuence maximization is the problem of ﬁnding a small set of seed nodes in a social network that maximizes the spread of inﬂuence under certain inﬂuence cascade models. The scalability of inﬂuence maximization is a key factor for enable-ng prevalent viral marketing in large scale online social networks. Prior solutions, such as the greedy algorithm and its improvements are slow and not scalable, while other heuristic algorithms do not provide consistently good performance on inﬂuence spreads. A new heuristic algorithm that is easily scalable to millions of nodes and edges in the experiments. The algorithm has a simple tuneable parameter for users to control the balance between the running time and the inﬂuence spread of the algorithm[6].

Bhagyasree vyankatro barde et. al proposed an overview of topic modelling methods and tools for Topic modelling. It is a powerful technique for analysing of collection of document. It is used for discovering hidden structure from the collection of document. The topic modelling include VSM, LSI, and LDA. Tools available for topic modelling are gensim, standford topic modelling toolbox, mallet and bigARTM. Topic model has wide range of application like tag recommendation, text categorization, keyword extraction, information filtering and similarity measures [7]

Zhang et. al described that a multidimensional latent semantic analysis (MALSA) enables us to mine local information from a document with the term association and spatial distribution. This technique works by first partitioning each document into a paragraph and then built a term affinity graph, this graph gives the frequency of term co occurrence in a paragraph. A 2D principle component analysis (PCA).is used for semantic mapping. It is used to find the leading eigen vector of the covariance matrix of the training set to characterize the lower dimensional space[8].

In 2015, Yanguang et. al compared the four text classifiers. These classifiers were tested on movie reviews whether they were able to classify the movie reviews as either positive or negative. This involves the analysis of the feature in the reviews and sometimes this could be result in curse of dimensionality which means the analysis of both the useful and useless features and hence it necessary to carefully select the feature for the correct classification[9]. In the same year, Lianjing et al. stated that the text classification is the base of text mining. Naïve Bayes is an effective method for text classification and improve the accuracy of Naïve Bayes classification using information gain is one the method of future extraction. This can be done by reducing the impact of low frequency word. A corpus of NLTK is used the accuracy of the classification was improved significantly[10]. Devesh Varshney et al. designed a model for predicting the probabilities of diffusion of a message through the social networks the machine learning based Bayesian approach utilize user interest and content similarity modelled using the latent topic information. The main aim is to finding the time by which user is expected to perform an action to propagate the information further and the techniques used are EM algorithm, Greedy approximation algorithm, Inﬂuence maximization algorithm advantage is to improve the performace analysis by using bayesian network approach disadvantage is information diffusion through social networks has focused on the problem of inﬂuence maximization[11].

Lan et. al described that in the vector space model (VSM), document can be transformed into vector in the term space, this text representation can be recognized and can be classified. To improve the text classification, it assigns different weights to the term using the term weighted method. A supervised and unsupervised term weighting method is used along with the SVM and KNN to do the classification. In supervised term weighting method called tf-irf is used and it seems to perform better when compared to the tf-idf weighting whether it is by using the linear or the nonlinear SVM classification algorithm[12].

Chanhyun Kang et. al proposed that emergence of online semantic social networks where vertices have properties and edges are labelled with relationships and weights . the technique used here are Hypergraph fixed point algorithm HyperDC algorithm Hyper LEP Advantage is increasingly important problem in social networks is that of assigning a “Centrality” value to vertices reflecting their importance within the social networks. and disadvantage is diffusion centrality is also often faster to compute that betweeness, closeness and stress centrality, but slower than degree and eigenvector centrality[13]

Lianjing et. al that the text classification is the base of text mining. Naïve Bayes is an effective method for text classification and improve the accuracy of Naïve Bayes classification using information gain is one the method of future extraction [14]. This can be done by reducing the impact of low frequency word. A corpus of NLTK is used the accuracy of the classification was improved significantly.

Shahin Mahdizadehaghdam, Han Wang et. al proposed a diffusion equation for the three layer inter connected network, moreover we consider the external effect on each node by assuming that the whole system is a multiple Brownian system. The prediction for an interconnected network achieves a lower error the technique used are K-nearest neighbour algorithm Epsilon-Neighbourhood algorithm Single-layer prediction method and advantage is toexplicitly accounting for the data of the network as it evolves among the agents and also showed that increasing the size of the network yields an improvement in the error and disadvantage an interconnected network has connected intra layer networks, but interlayer networks and smallest eigen value of the supra laplacian matrix [15].

Kazumi Saito et. al proposed a graph construction approach to overcome the problem of discovering the influential nodes in a social network under the Suspectible Infected Suspecitble(SIS) model by means of final-time and integral-time. Pruning and burnout are the strategies used here for finding a single and multiple influential nodes effectively. A greedy algorithm needs a large amount of computation as it estimates the marginal gains for the expected number of nodes influenced a set of nodes which is considered as a drawback[16].

Masahiro Kimura et. al proposed a combinatorial optimization method for extracting influential nodes for diffusing information [18] to overcome the disadvantage of greedy algorithm. Bond percolation and graph theory is used. The marginal gains are efficiently estimated. This method overcome the conventional methods efficiently[17].

Sheng Wen et. al proposed that the dissemination speed, large amount of users can swiftly distribute information to the masses, but they are not highly connected users for dissemination scale, many powerful forwards in OSNs cannot be identified by the degree measures. To control Dissemination, popular users cannot capture most bridges of social communities the technique used are Identify influential users Heuristic algorithm, disadvantage is highly necessary to develop an accurate mathematical platform that can be used to evaluate all these measures together and advantage is communication medium, organisers rapidly spread messages of the riots to others. These platforms have also been utilised to commit cybercrimes, such as distributing rumours, malicious URLs and spams[18]

Yanhua Li et. al proposed that inﬂuence diffusion and inﬂuence maximization in large-scale online social networks (OSNs) have been extensively studied because of their impacts on enabling effective online viral marketing. Existing studies focus on social networks with only friendship relations, whereas the foe or enemy relations that commonly exist in many OSNs, e.g., Epinions and Slashdot, are completely ignored. The ﬁrst attempt to investigate the inﬂuence diffusion and inﬂuence maximization in OSNs with both friend and foe relations, which are modelled using positive and negative edges on signed networks[19].

III PROPOSED SYSTEM

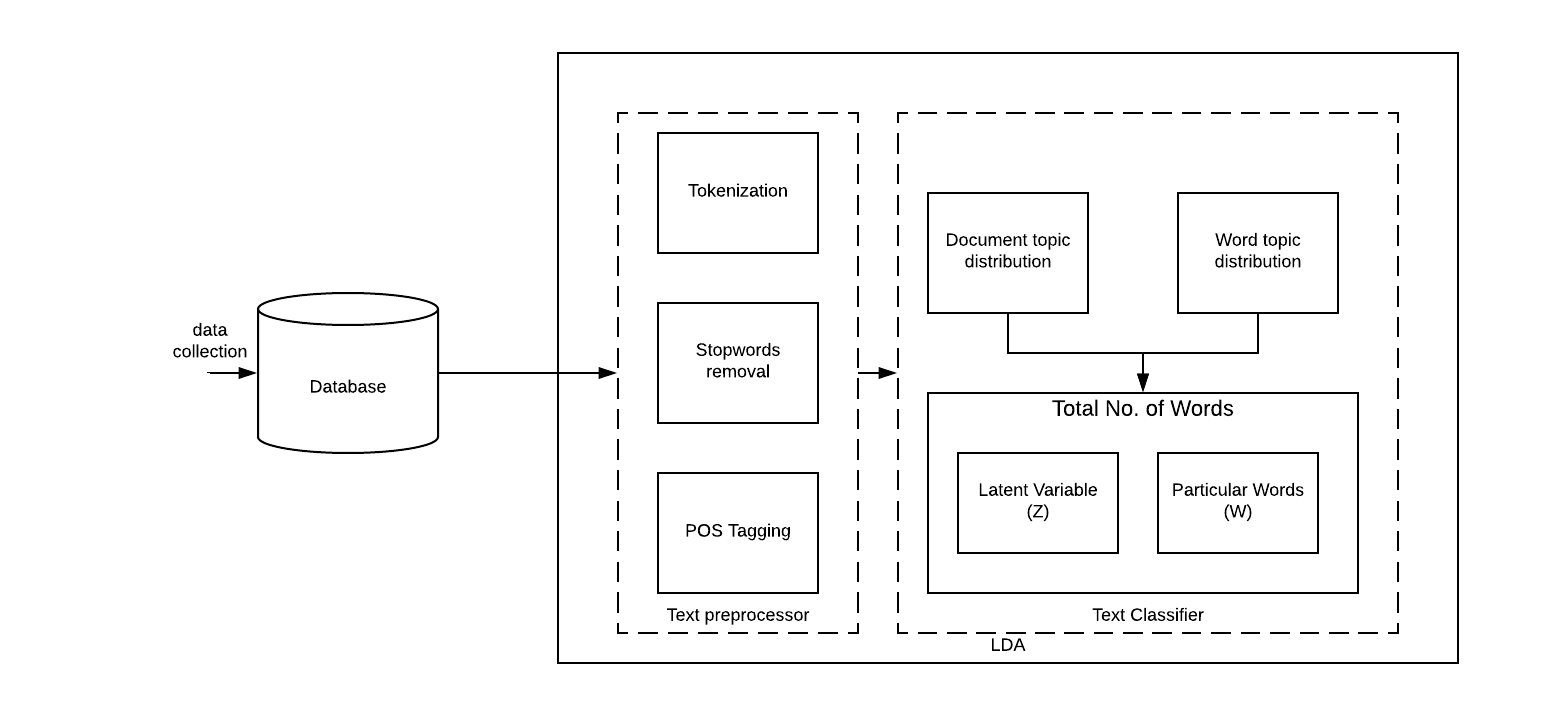
The proposed system, data is collected from the RESTapi(Representational State Transfer) or restful . by collecting the dataset topic are categorized using machine learning LDA algorithm Finally, the evaluation of the proposed model is done by using metrics such as precision, recall and f1 score which delivers the accuracy of the detection system. Fig. 1 shows the architecture of the proposed model.

1. Dataset collection:

The dataset is collected from a streaming twitter application REST(Representational State Transfer) RESTful web service is a way of providing interoperability between computer system on the internet and after passing the secured authentication mechanism achieved by generation of an aunthetication key. The Oauth tool tab in the twitter application is accessed to get the dataset containing the recent tweets and followers. For accessing the Oauth tool tab the user should have a consumer key, consumer secret, access token, access token secret.

1. Text preprocessing

Classification begin with preprocessing which is taken as the training dataset . preprocessing involves tokenization, stemming, stopwords removal case folding and capitalization and par of speech tagging which can be done efficiently using natural language processing. tokenization such as removing of numbers, punctuation marks, special characters ect it also include the word and sentence tokenizer. Stemming is used to remove the suffixes. Part of speech tagging identifies the part of speech(verb, noun, adjectives,ect)

****

**Fig. 2: Architecture of the proposed topic categorization**

1. Topic Classifier

The tweets contain some data that should be removed before further processing which is done by means of preprocessing the data. Text preprocessing is done by removing the stop words and punctuations. Then splitting the file into documents.

The topic categorization is done through Latent Dirchlet Allocation (LDA) algorithm which is a probabilistic model for separation of collective data. For topic modelling, each of the latent topic is created for the observed words from a group of words. The identification is done through probalisitic semantic analysis which automatically discovers the observed words which are taken as the topic. A tweet document-t usually consists of a series of topics, which can be described by a specific frequency or probability distribution. It is assumed that a text document consists of several hidden topics. The topic model is considered as a generative process for tweet documents [20,21,22]. Each topic obey a probability distribution over the feature words. LDA is a probabilistic generative model, which is first introduced by Blei et al. [22]. LDA has been widely used to estimate the multinomial observations and adopt LDA model to generate the probabilistic topic.

In LDA topic model, documents are described as random mixtures over latent topics. Each topic is characterized by a probabilistic distribution over words. LDA assumes the parameter and variables are the generative process for a corpus consisting of documents each of length.

is an particular word or an observed variable and another are the latent variable. The probability model is

Integrating out of and the likelihood of the document is obtained

**function** topic\_classifier (tw, α) **returns** categorized topics

**Inputs:** tw, a set of user tweets

α,β parameter of the dirichlet

**Outputs:** Zij, returns topic for jth word in document i

Wij, returns a particular word

**local variables:** i,j defines the position of the word

M, number of documents

N, number of words in a document

K, number of topics

**while** sample is not empty do

Sample

return α

Sample

return β

**return** Zij,Wij

Fig 3: Topic categorization algorithm

IV EXPERIMENTAL RESULT

To demonstrate the performance enhancement of our proposed approach and to evaluate our method with different parameters. The experimental setup is done using python in an anaconda navigator environment.

Topic are categorized using LDA algorithm using count vector, tf-idf. The highest probability value is calculated and ploted as the graph

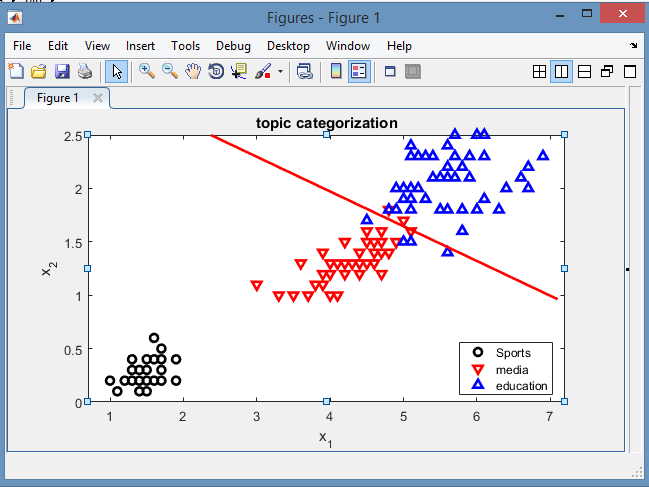


Fig. 4: Topic classification

The graph in Fig. 4 denotes the topic that are classified by LDA which denotes each of the topics are plotted according to their frequency values and class of classification.

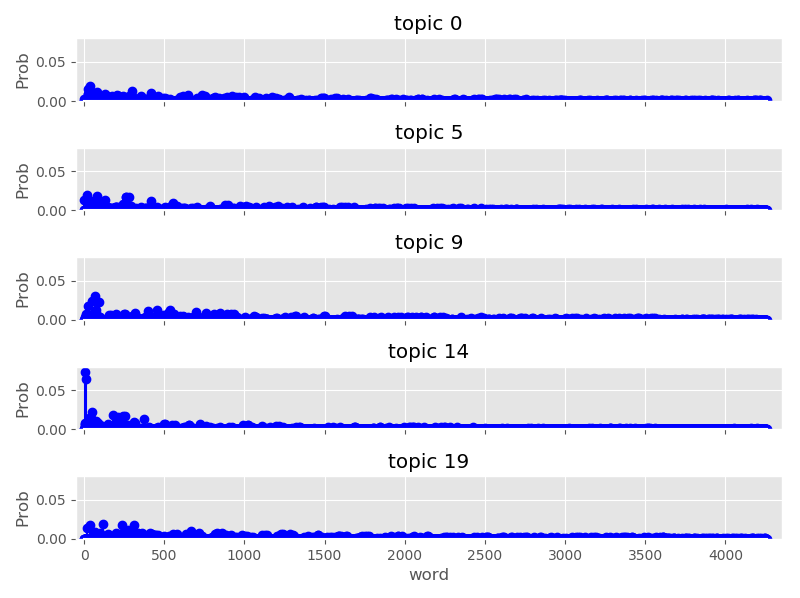


Fig. 5: Rate of Topic categorization

The graph shows that the probability of the vector count is increased through the implementation. For each of the topic iterations, an increase in the probability of finding the word is observed which symbolizes a maximum efficiency in the topic categorization.

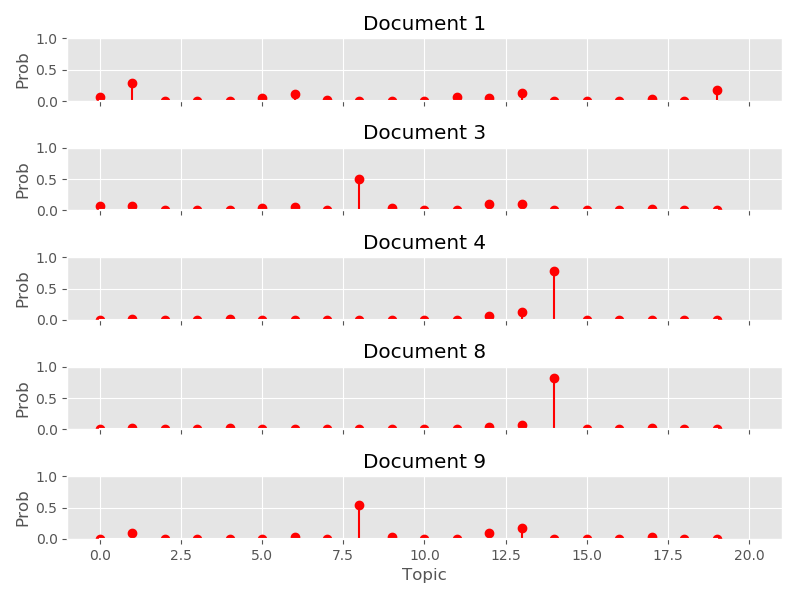


Fig 6 topic categorization based on documents

For each iterations, an increase in the probability of finding the topic in the document is observed which symbolizes a maximum efficiency in the topic categorization.

V CONCLUSION

The proposed system clearly states that there is an efficient increase in the detection of topic categorization using Latent Dirichlet allocation (LDA). The latent topics are generated by LDA With the LDA , the multinomial distribution of the twitter topics as the feature of the document. Experimental validation on a twitter dataset collected from the REST API , demonstrate the significant performance of proposed model.

REFERENCES:

1. Anupriya, P. and Karpagavalli, S., 2015, January. LDA based topic modeling of journal abstracts. In Advanced Computing and Communication Systems, 2015 International Conference on (pp. 1-5). IEEE.
2. Chen, Y.H. and Li, S.F., 2016, July. Using latent Dirichlet allocation to improve text classification performance of support vector machine. In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 1280-1286). IEEE.
3. Yang, Y., Lu, Z., Li, V.O. and Xu, K., 2017. Noncooperative Information Diffusion in Online Social Networks Under the Independent Cascade Model. IEEE Transactions on Computational Social Systems, *4*(3), pp.150-162.
4. Wang, Z., Yang, Y., Pei, J., Chu, L. and Chen, E., 2017. Activity Maximization by Effective Information Diffusion in Social Networks. IEEE Transactions on Knowledge and Data Engineering, *29*(11), pp.2374-2387.
5. Zhang K , Yun .K , Liang.J , Zhang.X , Li.C , Tian.B,(2016) Retweeting behavior prediction using probabilistic matrix factorization, in: Computers and Communication(ISCC), IEEE Symposium on, IEEE, 2016, pp. 1185–1192.
6. Chen .W, Wang .Y, and Yang .S, “Efﬁcient inﬂuence maximization in social networks,” in Proc. Int. Conf. Knowl. Discov. Data Min., Paris, France, pp. 199–208, 2009.
7. Bhagyashree Vyankatrao Barde, and Anant Madhavrao Bainwad”an overview of topic modelling methods and tools” international conference on intelligent computing and control system ICICCS,pp 978-1-5386-2745-7,2017
8. Zhang H., Ho J. K., Wu Q.M., & Ye Y.( 2013), ‘Multidimensional latent semantic analysis using term spatial information ‘,IEEE Transaction on Cybernetics,vol.43,No.6,pp.1625-1640.
9. Yuncheng Jiang, Xiaopei Zhang,Yong Tang and Ruihua Nie,( 2015)’ Feature based approaches to semantic similarity assessment of concepts using Wikipedia’, Information Processing and Management, vol.51,No.3,pp.215-234.
10. Lianjing jin, Wei Gong, Wenlong Fu, Hongbin Wu,(2015) a Text Classifier of English Movie Based on Information Gain;, International Conference on Computational Science and Intelligence(ACIT-CSI),vol.1,pp263-268.
11. Varshney.D., Kumar, S. and Gupta, V (2017). Predicting information diffusion probabilities in social networks: A Bayesian networks based approach. Knowledge-Based Systems, 133, pp.66-76.
12. Lan M and Tan C.L,(2009)’supervised and traditional term weighting methods for automatic text categorization’ , IEEE Transactions on Knowledge and Data Engineering, vol.31,No.4,pp.721-735.
13. Kang.C ,Molinaro.C , Kraus.S , Shavitt.Y ,Subrahmanian.V ,(2012) Diffusion centrality in social networks, in: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 558–564
14. Lianjing jin, Wei Gong, Wenlong Fu, Hongbin Wu,’a Text Classifier of English Movie Based on Information Gain;, International Conference on Computational Science and Intelligence (ACITCSI),vol.1,pp263-268, 2015.
15. Shahin Mahdizadehaghdam, Han Wang, Hamid Krim, and Liyi Dai,”Information Diffusion of Topic Propagation in Social Media” , IEEE transactions on signal and information processing over networks, vol. 2, no. 4, december 2016.
16. M. Kimura, K. Saito, and R. Nakano, “Extracting influential nodes for information diffusion on a social network,” in Proc. AAAI, vol. 2. Vancouver, BC, Canada, pp. 1371–1376, 2007.
17. M. Kimura and K. Saito. Tractable models for information diﬀusion in social networks. In Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006), pages 259–271. LNAI 4213, 2006
18. Sheng Wen, Jiaojiao Jiang, Yang Xiang., (2014)” Are the Popular Users Always Important for Information Dissemination in Online Social Networks?” IEEE Network,0890-8044/14.
19. Li .Y, Chen .W, Wang .Y, and Zhang .Z-L, “Inﬂuence diffusion dynamics and inﬂuence maximization in social networks with friend and foe relationships,” in Proc. Int. Conf. Web Search Data Min., Rome, Italy, pp. 657–666, 2013.
20. David M Blei, “Probabilistic topic models,” Communications of the ACM, vol. 55, no. 4, pp. 77–84, 2012.
21. Thomas Hofmann, “Probabilistic latent semantic indexing,” in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999, pp. 50–57.
22. David M Blei, Andrew Y Ng, and Michael I Jordan, “Latent dirichlet allocation,” the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.