



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 7, July 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.118



9940 572 462



6381 907 438



ijirset@gmail.com



www.ijirset.com

Naïve Bayes Classification Algorithm Based Improve the Performance of the Tracking and Clustering in Tweets

Mr. G. Dhanasekar¹, Mr. M.Dhamodhar², Mr. S. Sudharshan Reddy³ Ms.T.Yogakshi⁴

Assistant Professor, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India¹

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India²

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India³

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India⁴

ABSTRACT: The main aim of this project is finding the repeated activities in a social network. The different post that are continuously posted by a user in the twitter. The text information arriving continuously arriving over the time in the form of text stream are predicted and clustered for analysis. This project will track what type of information that continuously arriving in social media and the people interest. All the operations are going to be performed on the twitter dataset. The dataset which consisting of Twitter users and their tweets. This model is proposed to analyse the contagion dynamics that emerge in networks when repeated activation is allowed, that is, when actors can engage recurrently in a collective effort. In this project, the k-means clustering algorithm is used for clustering the users based on their interest and Naïve Bayes classification algorithm is used for classifying the users. And also in this project we can track the interest of the particular result. This project will increase the accuracy of the overall prediction results.

KEYWORDS: Clustering Naïve Bayes classification algorithm, Tweets, Clustering Etc...

I. INTRODUCTION

The goal is to identify the conditions under which coordination is more likely to arise from networks that are constantly pulsating with information. Our model aims to relax these assumptions and allow actors to repeatedly activate as a function of the dynamics unfolding in the rest of the network. We argue that this modification aligns our model of contagion more closely with what is observed in many empirical networks – in particular, with the communication dynamics observed in online networks and the temporal autocorrelation that results from those dynamics. Online campaigns are an important manifestation of this type of repeated activation, and they offer a good example of what we mean by “coordination”: a form of organizational effort to attract public attention or direct mobilization logistics on the ground.

We have theoretical and empirical reasons to allow repeated activation to be the driving force of contagion dynamics. The empirical reason is that most instances of diffusion do not involve a single activation but many activations building up momentum in time. Before a hashtag becomes a trending topic, a period of buzz is first required; prior to a protest day, calls announcing the mobilization are distributed in waves. Actors decide whether they want to engage in an online conversation or take part in a protest. This is what threshold models can capture. What threshold models are not devised to capture is the period of information exchange that follows the act of joining a collective effort. During this period, social influence trickles intermittently as a function of the context that actors inhabit – that is, as a function of activity in the local networks to which they are exposed; and this context is not stationary: it changes, sometimes drastically, over time. Our model aims to capture this temporal dimension.

Our model assumes that exposure to information is the driving force underlying contagion. What makes our model different from previous models is that failure to trigger a chain reaction depends not only on the distribution of thresholds or the impact of network structure on activation dynamics; it also depends on whether the network facilitates coordination, that is, an alignment of actions in time – which is an important organizational goal for social movements that want to gain public visibility in social media or use online networks to manage mobilization. By focusing on coordination dynamics, our model is in a better position to explain why, more often than not, large-scale contagion fails to take off. If the network is not conducive to coordination (i.e. if the timing of individual activations do not align over time), contagion ends up trapped in local activity clusters and, therefore, fails to synchronize the actions of the majority.

Our main assumption is that actors reach their activation zone at different speeds. The speed of activation is a function of two parameters: ω , which determines how quickly the actor reaches the threshold zone (i.e. it defines the concavity of the curve that maps progression towards activation); and ϵ , or the strength of the signal received from other actors – which, in our case, is restricted to actors one step removed in the network.

This paper is organized in five sections. After this introduction, in Section II, literature survey discussed of the paper, section III about the System Analysis, Section IV about System Design, as well as the novel feature of the proposed method. Finally, Sections V and VI provide the simulation results and the conclusions, respectively.

II. RELATED WORK

This section of Literature Survey eventually reveals some facts of Biometric Authentication based on the analysis of many authors work as follows:

Chandra Shekhar Vorugunti [1] has introduced a new concept of BioAaaS to maintain secure authentication. Based on SAAS model of Cloud it provides a light weight and secure authentication mechanism. It contains two steps for authentication. First is Enrolment and next is Verification. In Enrolment process the biometric data is converted into a binary form. The feature extractor then converts the binary string into a set of features. In verification process same process will be processed when the user logs in to the cloud. The matching module matches the features of the stored data and login data. Thus they have provided a service to do heavy weight cryptographic encryption and decryption operation on user's biometric data.

D J Craft [2] reports on fast hardware implementation of lossless data compression algorithms. It proposed Adaptive Lempel-Ziv Algorithm (LZ1 & LZ2). LZ algorithm are symbol based that is they operate one data one character at a time. They achieve compression by locating frequency occurring sequences of such symbol in input data stream. ALDC have two extensions as BLDC & CLDC. BLDC pre-processing works well on only bitmapped image data. CLDC is combination of ALDC & BLDC. The main difference between LZ1 & LZ2 is in the data structure employed & the way reference to sequence are coded. Cong Li et al. [3] proposed Burrows Wheeler Transformation based DNA sequence data multi-compression using Open MP & MPI. They proposed data compression (DNA sequence) using fewer bits rather than encoded data to represent information. BWT based DNA compression includes few steps. First DNA sequence data is encoded with 0/1 which has 4 characters. Then BWT transformation is performed over it. Again MTF transformation is performed. Then we compress data with classical algorithm.

Kiran Kumar K et al. [4] have described that there are two properties of fingerprint namely uniqueness and permanence that are used for identification and verification. These properties are judged by minutae and ridges. The method used in this paper has 8 stages. They are gray-level fingerprint image, binarization, thinning, minutae extraction, false minutae, matching scores, ridges extraction, minutae and ridge score fused using strength factor. The block filters preserve the outermost pixels along each ridge.

Jeff Collier [5] proposed a system for developing a software that would address the challenges such as in-memory map/reduce. It also deals with the node that has ability to leave and re-join the cloud by applying compression and image processing algorithms.

Hu Chun et al. [6] have proposed a situation where biometric data is kept encrypted in whole process of transmission and matching. It uses two approaches homomorphism encryption and garbled circuit. It provides highly computing capability.

Surender Sharma et al. [7] have introduced health care monitoring system application that provides the patients with necessary healthcare information yet it also gives a chance to threats of intervention that would make the critical data insecure. They have used Body Area wireless sensor network as monitoring component. The cloud based HMA was therefore developed using master slave like pattern, where the master could have generic functions while slave would have functionalities specific to the medical condition. Thus they have utilized biometric encryption for providing protection to the data. Here the user's biometric characteristics work as decryption key here fuzzy extractor scheme has been used to convert the scanned fingerprint data to some random string and an helper string to apply cryptographic techniques. This framework accomplishes both the goals, secure access and data protection.

Krishnaraj Madhavji Sunjiv Soyjaudah [8] discussed about eight points of vulnerabilities that can be hacked. In cloud data is moved dynamically so security is a major concern and there are the problems which arise in the management of biometric data. So a cancellable biometric authentication system is proposed by him. Cancellable Biometric Authentication is a concept in which the original image is first distorted then shared on

cloud. This distorted biometric image is used for authentication. This provides security and privacy as the original biometric are never revealed to authentication server. Data Hiding is also done using this technique to overcome the replay attack. This is done by secretly embedding the private information in biometric image.

Dr. Anandhakumar P et al. [9] has addressed the issues that arouse during storing different documents and files and photo contents on Cloud. Huge amount of photos are maintained by cloud providers. Huffman coding cannot achieve high levels of compression also all the binary strings or codes in the encoded data are of different lengths. So it is difficult to decode. The representative signal (RS) based approach is suitable only when images are highly correlated to each other so it fails badly in case of illumination changes. To overcome these drawbacks LZ-77 algorithm is proposed in this paper. LZ-77 replaces the repeatedly occurring data with reference to single copy which already exists in an uncompressed data stream. It uses length-distance pair to encode the match. As compression is in cloud environment, k-means algorithm is used to transfer up by using Map-Reduce concept. They also proposed an idea for effective compression of photo albums which also reduces the time complexity.

A. Inferring dynamic user interests in streams of short texts for user clustering S. Liang, Z. Ren, Y. Zhao, J. Ma, E. Yilmaz, and M. D. Rijke 2017

To propose a dynamic user clustering topic model (UCT). UCT adaptively tracks changes of each user's time-varying topic distributions based both on the short texts the user posts during a given time period. A Gibbs sampling algorithm where a set of word pairs from each user is constructed for sampling. UCT can be used in two ways: (1) as a short-term dependency model that infers a user's current topic distribution based on the user's topic distributions during the previous time period only, and (2) as a long-term dependency model that infers a user's current topic distributions based on the user's topic distributions during multiple time periods in the past. The clustering results are explainable and human-understandable, in contrast to many other clustering algorithms. For evaluation purposes, we work with a dataset consisting of users and tweets from each user. Experimental results demonstrate the effectiveness of our proposed short-term and long-term dependency user clustering models compared to state-of-the-art baselines.

B. Collaborative user clustering for short text streams S. Liang, Z. Ren, Y. Zhao, J. Ma, E. Yilmaz, and M. D. Rijke 2017

A UCIT model that integrates both users' and their collaborative interests for user clustering by short text streams. A user collaborative interest tracking model (UCIT) that aims at tracking changes of each user's dynamic topic distributions in collaboration with their followees', based both on the content of current short texts and the previously estimated distributions. We evaluate our proposed method via a benchmark dataset consisting of Twitter users and their tweets. Experimental results validate the effectiveness of our proposed UCIT model that integrates both users' and their collaborative interests for user clustering by short text streams.

C. Implicit Feature Identification via Co-occurrence Association Rule Mining Z. Hai, K. Chang, and J.-J. Kim 2011

A novel two-phase co-occurrence association rule mining approach to identifying implicit features. Specifically, in the first phase of rule generation, for each opinion word occurring in an explicit sentence in the corpus, we mine a significant set of association rules of the form [opinion-word, explicit-feature] from a co-occurrence matrix. In the second phase of rule application, we first cluster the rule consequents (explicit features) to generate more robust rules for each opinion word mentioned above. Given a new opinion word with no explicit feature, we then search a matched list of robust rules, among which the rule having the feature cluster with the highest frequency weight is fired, and accordingly, we assign the representative word of the cluster as the final identified implicit feature. Experimental results show considerable improvements of our approach over other related methods including baseline dictionary lookups, statistical semantic association models, and bi-bipartite reinforcement clustering.

D. Opinion word expansion and tar-get extraction through double propagation E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw

This paper aims to detect users generating spam reviews or review spammers. We identify several characteristic behaviors of review spammers and model these behaviors so as to detect the spammers. In particular, we seek to model the following behaviors. First, spammers may target specific products or product groups in order to maximize their impact. Second, they tend to deviate from the other reviewers in their ratings of products. We propose scoring methods to measure the degree of spam for each reviewer and apply them on an Amazon review dataset. We then select a sub-set of highly suspicious reviewers for further scrutiny by our user evaluators with the help of a web based spammer evaluation software specially developed for user evaluation experiments. Our results show that our proposed ranking and supervised methods are effective in discovering spammers and outperform other baseline method based

on helpfulness votes alone. We finally show that the detected spammers have more significant impact on ratings compared with the unhelpful reviewers.

E. Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site M. Trusov, R. E. Bucklin, and K. Pauwels 2009.

Social network sites record the electronic invitations from existing members, outbound WOM can be precisely tracked. Along with traditional marketing, WOM can then be linked to the number of new members subsequently joining the site (sign-ups). Because of the endogeneity among WOM, new sign-ups, and traditional marketing activity, the authors employ a vector autoregressive (VAR) modeling approach. Estimates from the VAR model show that WOM referrals have substantially longer carryover effects than traditional marketing actions and produce substantially higher response elasticities. Based on revenue from advertising impressions served to a new member, the monetary value of a WOM referral can be calculated; this yields an upper-bound estimate for the financial incentives the firm might offer to stimulate WOM.

F. Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site J. Qi, Z. Zhang, S. Jeon, and Y. Zhou 2015

Big data commerce has become an e-commerce trend. Learning how to extract valuable and real time insights from big data to drive smarter and more profitable business decisions is a main task of big data commerce. Using online reviews as an example, manufacturers have come to value how to select helpful online reviews and what can be learned from online reviews for new product development. In this research, we first proposed an automatic filtering model to predict the helpfulness of online reviews from the perspective of the product designer. The KANO method, which is based on the classical conjoint analysis model, is then innovatively applied to analyse online reviews to develop appropriate product improvement strategies. Moreover, an empirical case study using the new method is conducted with the data we acquired from JD.com, one of the largest electronic marketplaces in China. The case study indicates the effectiveness and robustness of the proposed approach. Our research suggests that the combination of big data and classical management models can bring success for big data commerce.

G. Predicting user behaviour through sessions using the web log mining G. Neelima and S. Rodda 2014.

It is the method to extract the user sessions from the given log files. Initially, each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Two types of logs i.e., server-side logs and client-side logs are commonly used for web usage and usability analysis. Server-side logs can be automatically generated by web servers, with each entry corresponding to a user request. Client-side logs can capture accurate, comprehensive usage data for usability analysis. Usability is defined as the satisfaction, efficiency and effectiveness with which specific users can complete specific tasks in a particular environment. This process includes 3 stages, namely Data cleaning, User identification, and Session identification. In this paper, we are implementing these three phases. Depending upon the frequency of users visiting each page mining is performed. By finding the session of the user we can analyse the user behaviour by the time spend on a particular page.

H. Pre-processing techniques in web usage mining: A survey M. Srivastava, R. Garg, and P. Mishra 2015.

Due to huge, unstructured and scattered amount of data available on web, it is very tough for users to get relevant information in less time. To achieve this, improvement in design of web site, personalization of contents, prefetching and caching activities are done according to user's behaviour analysis. User's activities can be captured into a special file called log file. There are various types of log: Server log, Proxy server log, Client/Browser log. These log files are used by web usage mining to analyse and discover useful patterns. The process of web usage mining involves three interdependent steps: Data pre-processing, Pattern discovery and Pattern analysis. Among these steps, Data pre-processing plays a vital role because of unstructured, redundant and noisy nature of log data. To improve later phases of web usage mining like Pattern discovery and Pattern analysis several data pre-processing techniques such as Data Cleaning, User Identification, Session Identification, Path Completion etc. have been used. In this paper all these techniques are discussed in detail. Moreover these techniques are also categorized and incorporated with their advantage and disadvantage that will help scientist, researchers and academicians working in this direction.

III. SYSTEM ANALYSIS

A. Existing System

The existing works ignore the semantic correlation among different reviews, causing the ineffectiveness for sentiment classification. Clustering users by reviews is more challenging than in the case of long documents associated with them as it is difficult to track users' reviews in streaming sparse data. Loss of information during the clustering. This may lead to the incorrect clustering result.

Dis-advantages

- Sparsity Problems
- Theoretical Limits.
- Loss of Information.
- Incorrect Clustering Results.

B. Proposed System

The proposed model is introduced to overcome all the disadvantages that arises in the existing system. It reduces the information loss and the bias of the inference due to the multiple estimates. To enhance the performance of our proposed system, we dynamically cluster reviews based on their reviews about the particular thing in social media.

Advantages

- High Performance.
- It avoids Sparsity problems.

IV. SYSTEM DESIGN

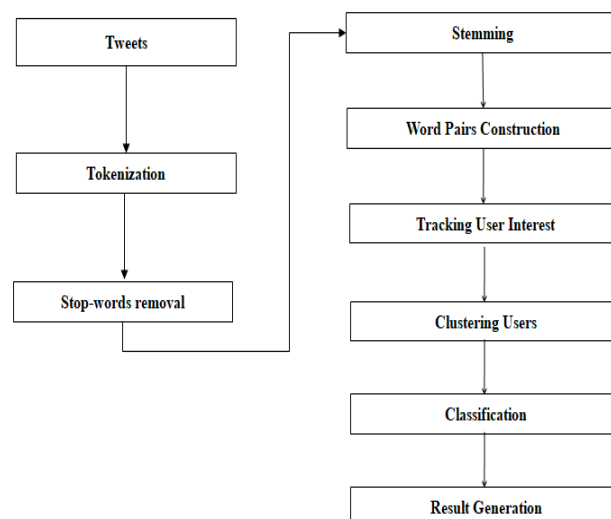


Figure 1: System Architecture

A. DATA SELECTION AND LOADING

Data selection is the process of selecting the appropriate data set for processing. The dataset which contains the fields of user id and their tweets. The tweet is a piece of message that is posted by the user. The selected dataset is going to be used for tracking and clustering user interest. The dynamically distributed topics are detected or tracked from the user's tweets.

B. DATA PREPROCESSING

The data is pre-processed to remove reviews from anonymous users, since we would like to associate each review with a unique user. The Data pre-processing is the process of detecting, correcting or removing, corrupt or inaccurate records from the dataset. The records which provide the incorrect clustering results are detected and removed from the dataset. In this process, we are going to delete the records that contain empty fields.

C. TOKENIZATION

Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences.

D. POSTAGGING

Part-of-Speech tagging is the process of marking up a word in a text as corresponding to a particular part of speech based on both its definition and its context. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction. The standard pos tagger tool is used for postagging process.

E. STOP WORDS REMOVAL

Stop words are natural language words which have very little meaning, such as "and", "the", "a", "an", and similar words. The stop words are detected from the reviews and it's removed.

F. STEMMING

Stemming is the process of converting the words of a sentence to its non-changing portions. The porter stemming algorithm is used for stemming the words. Term Frequency construction: After the stemming process, the term frequency are constructed from the stemming words.

G. BITERM CONSTRUCTION

It's the process of constructing the word pairs. Constructing word pairs rather than directly using each single word for topic inference are:

- Topics are groups of correlated words, and the correlations are revealed by words' co-occurrence patterns in documents.
- The underlying topic expressed by a single word is more ambiguous than that of a word pair.

H. TERM FREQUENCY CONSTRUCTION

Term frequency (TF) is used in connection with information retrieval and shows how frequently an expression (term, word) occurs in a document. Term frequency indicates the significance of a particular term within the overall document. This value is often mentioned in the context of inverse document frequency IDF.

I. CLUSTERING

It's the process of dynamically clustering the users based on their interest. The K-means Clustering algorithm is used for dynamically clustering the users based on the topic distributions.

J. CLASSIFICATION

Classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Classification is the process of classifying the user based on the data arrival in social media. Naïve-Bayes Classification algorithm is used for classifying the data based on their interest. Classification result displays the user's interesting and not-interesting topics.

K. PREDICTION

Classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

L. CHART GENERATION

The overall clustering report is generated based on their interest in twitter. The report which contain dynamically distributed topics in the tweeter and their level. The level that describe how many times the topics are distributed among the users. And also it will show the result of particular user interest.

V. SIMULATION RESULTS

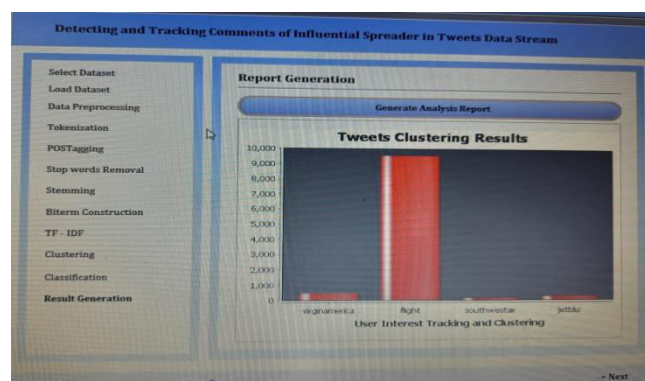


Figure 2: Tweets Clustering

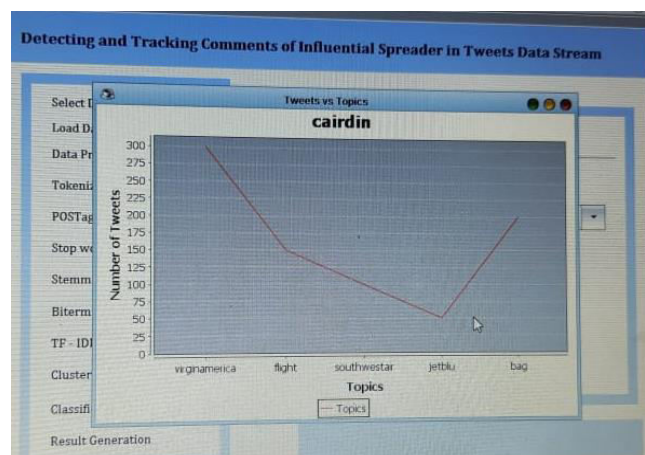


Figure 3: Tweets Vs Topics

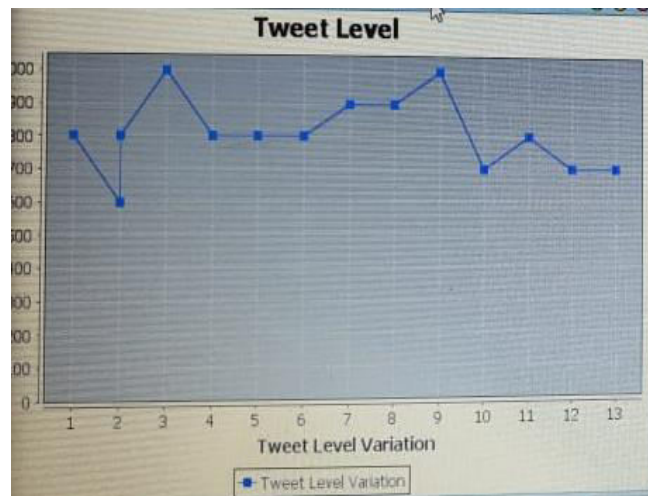


Figure 4: Tweets level variation

VI. CONCLUSION

This project enhances the performance of the overall clustering and prediction. It finds the different clusters and it is summary effectively and quickly. It alleviate the sparsity problem and reduces the information loss. The accuracy of the clustering and classification result is highly increased. It effectively track and cluster the user by avoiding the sparsity problems and also this project enhance the performance of the overall tracking and clustering. We evaluated the performance of clustering, topical representation and generalization effectiveness, and make comparisons with state-of-the-art models. We have also found that UCT produces higher quality topic representations than competing methods, and it comes with the benefit of offering explanations of the clustering.

Future Scope

In future we can store all the data in a hadoop storage for increasing the processing speed. This process will increase the effectiveness of the data storage, processing and classification results. We intent to incorporate other information such as the users' social network for user clustering. Like most previous work, it is challenging to obtain the ground-truth number of user clusters in our model. Another line of work is to develop a more efficient user clustering model to utilize previously captured topic distributions of users for inferring a user's current interests, and to improve efficiency.

REFERENCES

1. Aral, Sinan, Walker, Dylan, 2012. Identifying influential and susceptible members of social networks. Science 337, 337–341.
2. Aral, Sinan, Muchnik, Lev, Sundararajan, Arun, 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. PNAS 106 (51), 21544–21549.
3. Aral, Sinan, Muchnik, Lev, Sundararajan, Arun, 2013. Engineering social contagions: optimal network seeding in the presence of homophily. Netw. Sci. 1 (02), 125–153.
4. Backstrom, Lars, Boldi, Paolo, Rosa, Marco, Ugander, Johan, Vigna, Sebastiano, 2012. Four degrees of separation. In: Proceedings of the 3rd Annual ACM WebScience Conference, Evanston, Illinois: ACM, pp. 33–42.
5. Barabási, Albert-László., 2009. Scale-free networks: a decade and beyond. Science 325 (5939), 412–413.
6. Barberá, Pablo, Wang, Ning, Bonneau, Richard, Jost, John, Nagler, Jonathan, Tucker, Joshua, González-Bailón, Sandra, 2015. The critical periphery in the growth of social protests. PLoS One 10 (11).
7. Bond, Robert M., Fariss, Christopher J., Jones, Jason J., Kramer, Adam D.I., Marlow, Cameron A., Settle, Jaime E., Fowler, James H., 2012. A 61-million-person experiment in social influence and political mobilization. Nature 489, 295–298.



INNO  SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118



ISSN  INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details