Recommendation System for Criminal Behavioral Analysis on Social Network using Genetic Weighted K-Means Clustering

V. Soundarya[1*], U. Kanimozhi[1], D. Manjula[2]

[1] Research Scholar, Department of Computer Science and Engineering, CEG, Anna University, Chennai-25, Tamil Nadu, India.
[2] Department of Computer Science and Engineering, CEG, Anna University, Chennai-25, Tamil Nadu, India.

* Corresponding author. Tel.: +91-9176675656; email: soundar_riya@yahoo.co.in
Manuscript submitted August 10, 2015; accepted March 23, 2016.
doi: 10.17706/jcp.12.3.212-220

Abstract: The accessibility and usage of social networking sites constructs both prospects and menaces for the users. In this research article, we propose a new recommendation system for predicting and recommending the criminal behavioral users on social network based upon the activities of the users. Our recommender system uses the proposed nine factor analysis method, clustering technique called Genetic Weighted K-Means clustering (GWKMC) and the existing classification algorithm namely Negative Selection Algorithm (NSA). The proposed recommendation system is evaluated by conducting various experiments using the Face book dataset (Latest) which is prepared on our own and also the Weblog dataset (Timeworn). The conducted experiments confirmed the efficacy of the proposed Recommender System.

Key words: Criminal behavior, negative selection, recommendation system, weighted K-means clustering.

1. Introduction

Our society is undergoing rapid renovation in almost all aspects due to the innovation of computers and computer networks. We are buying online, gather information by search engines and live a significant part of our social life over the Internet. Nowadays, the global network provides a huge amount of diverse data useful for social network analysis due to the fact that many of our actions and interactions are stored electronically. Internet-based social networks can be either directly maintained by dedicated web systems like Face book, Friendster, MySpace, and LinkedIn or extracted from data about user activities in the communication networks like e-mails, chats, blogs, homepages connected by hyperlinks, etc. The role of recommender systems is to crack data on users and their preferences into predictions of user's behaviour based on their likes and interests.

Clustering is the process of grouping a set of objects which are "similar" with one another or "dissimilar" from the objects of other clusters. The distance measurement is a major task between objects in clustering. Euclidean distance may be ambiguous in certain instances when the components of the data instance vectors are present in the same group of a cluster. Therefore, different distance measures can be used to form clusters. Clustering algorithms are categorized into four such as exclusive clustering, overlapping clustering, hierarchical clustering, and probabilistic clustering. In exclusive clustering grouped in an exclusive way and hence if certain data item belong to a particular cluster then it cannot be included in any other cluster. Overlapping clustering uses uncertainty to be applied to the cluster data so that, each point

can belongs to more than one cluster with different levels of membership. Next, the hierarchical clustering algorithm on the union between the two nearest clusters is considered to form a cluster. Finally, the probabilistic clustering approach uses the probability value to form cluster.

The criminal behavioural perspective promises to improve the understanding of the complexities of criminal activity and enhance intervention effectiveness. However, as we attempt to monitor criminal behaviour, understanding the criminal behaviour from the amount of available data which becomes less manageable for the human analyst, it is possibly send-off a knowledge gap that hinders effective decision-making. To improve a decision support in recommender system we are combining the nine factor analysis method, clustering technique called Genetic Weighted K-Means clustering (GWKMC) and the existing classification algorithm namely Negative Selection Algorithm (NSA). In this paper, we describe the proposed approach and the dataset is taken from recent activities of 1000 users for analysing the criminal behaviour of the users over social networks and, for this purpose, we present a comprehensive computational framework for criminal behavioural analysis defined in terms of a process that combines data mining and machine learning approaches.

In this paper, we propose a new recommendation system for predicting and recommending the criminal behavioral users on social network. Rest of this paper is organized as follows: Section 2 discusses about various past works done in this direction. Section 3 explains the overall system architecture. Section 4 described the proposed method. Section 5 contains the results and discussion. Finally, Section 6 gives the conclusion and future works.

2. Related Works

A number of extensive studies have been made over the past few years in behaviour analysis on Social Networks such as, Rafa Drezewski [1] presented a social network analysis component for detecting money laundering that makes use of data from bank statements and the National Court Register and construct and analyse social networks during an investigation into money laundering cases. The paper presented by Mamoun Alazab [2] examines the evolution of malware including the nature of its activity and variants, and the implication of this for computer security industry practices. Proposed a framework to extract features statically and dynamically from malware those reflect the behavior of its code such as the Windows Application Programming Interface (API) calls. Similarity based mining and a machine learning method has been employed to profile and classify malware behaviours.

Robert C. McMahon [3] hypothesized that cluster subgroups with more extensive criminal conduct would reveal more troubled social histories, less favourable out-of-home placement experiences, more mental health problems, and fewer social bonds and current support structures than those with less criminal conduct. José I. Castillo Manzano [4] examined the records of the 28 current member states of the European Union over the period from 1999 to 2010 to test the hypothesis that crime rates that can be considered as predictors of fatal road traffic accidents. The effect of the severity of the legal system applied to traffic offenses is analysed.

The study proposed by Kwang-Ho Lee [5] is a comprehensive hybrid model of the use of online travel communities for social andemotional loneliness (OTS-SEL), identification with the peer group (IPG), peer communication (PCO), user satisfaction (USAT), and behavioural intentions to follow travel advice (INFTA). Determines whether OTS-SEL is composed of three sub-dimensions of social loneliness, friend loneliness, and romantic loneliness based on a second-order stricture; tests a structural equation model to examine the relationships between OTS-SEL, IPG, PCO, USAT, and INFTA; and provides a multi-group analysis to investigate the moderating effect of emotional expressivity (EME) on the relationship between USAT and INFTA. Richard K. MouleJr [6] examines the patterns of Internet use among a sample of 585 individuals at-risk for and involved in street crime and compares on the general population, similar predictors and

lower rates of Internet participation and observed the usage, and suggests participation in criminal lifestyles contributes to digital inequality.

Patrick Lussier [7] proposed a concept of achievement in sexual offending defined as the ability to maximize the payoffs of a crime opportunity while minimizing the costs and showed a wide variation in criminal achievement, a variation that is not correlated with the severity of sentences meted out or the actuarial risk scores obtained by those offenders. The offenders who specialize in sex crimes were shown to be the most productive and least detected offenders. Criminal thinking styles were examined by Lorraine E. Cuadra [8] as mediational links between different forms of child maltreatment and adult criminal behaviours in 338 recently adjudicated men. Analyses revealed positive associations between child sexual abuse and sexual offenses as an adult, and between child physical abuse/neglect and endorsing proactive and reactive criminal thinking styles. Analysis showed associations between overall maltreatment history and adult criminal behaviours. Ram Dantu [9] proposed a methodology for vulnerability analysis of a network based on attacker behaviour, based on the sequence of actions carried out by the attackers and their social attributes, described a five-step model of vulnerable device detection and risk estimation of a network using attack graphs and attack behaviour. Here, we introduced an optimization technique of the network by patching the identified vulnerable devices or reconfiguration of network components for guaranteed security.

The role of civil unrest on social network were examined by Elhadj Benkhelifa [10] due to the similarity in monitoring online social media and digital cloud forensics, a framework was developed which spawned as a combination of these. This framework is applied in order to analyse such networks and produce datasets in order to potentially predict new incidents of civil unrest. To validate this framework a proof-of-concept implementation was given, which monitored twitter for signs of civil unrest in order to determine potential locations and dates. Watson [11] reports profiling information for speeding offenders and is part of a larger project that assessed the deterrent effects of increased speeding penalties in Queensland, Australia, using a total of 84,456 speeding offences. The speeding offenders were classified into three groups based on the extent and severity of an index offence such as once-only low-rang offenders, repeat high-range offenders and other offenders.

3. System Architecture

The overall architecture of the proposed system is shown in Fig. 1. The proposed system architecture is consists of seven major components such as Weblog dataset, User Interface Module, Feature selection module, Recommendation system framework and result. The recommendation system framework consists of two modules namely clustering and negative selection.

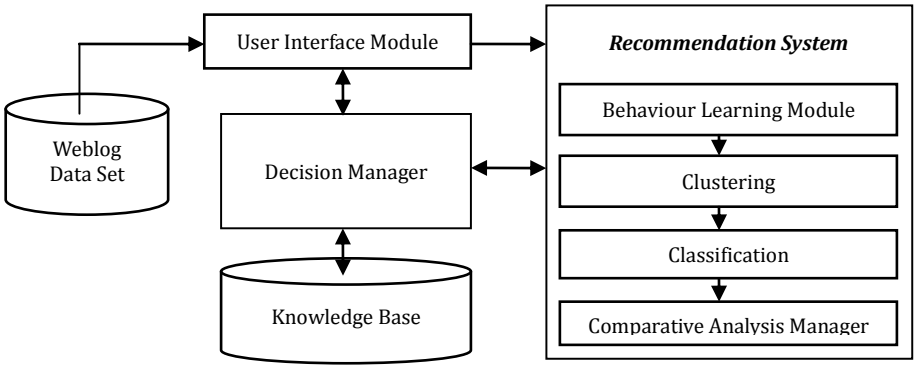

Fig. 1. System architecture.

Weblog Dataset contains the standard benchmark data set of social network user's which is collected

from all over the world. The user interface module collects the necessary data from the standard bench mark dataset and forward it to the recommendation system for further processing. The recommendation system contains four sub modules such as behaviour analysis module, Clustering module, Classification module and comparative analysis manager. Among them, the behaviour analysis module selects the necessary attributes from given dataset based on the factors determined and send these records to the classification module for further process. The classification module classifies the data by the help of decision manager and rule base.

Finally, the comparative analysis manager takes final decision whether the product is suitable for the society now or not by the help of decision manager and rule base. The decision manager takes decision over the social network user is having criminal behavioral or not using rules and the classification result. The knowledge contains the past experience of social network users and present youngster interest over social network which are helpful for taking decision over the users by the decision manager.

4. Proposed Work

In this paper, we propose a new recommendation system for recommending the criminal behavioural Facebook user's over social network. This recommendation system is the combination of the proposed behaviour analysis procedure which is functioning with nine factors, the proposed Genetic Weighted K-Means Clustering algorithm (GWKMC) and the existing Negative Selection Algorithm (NSA) [12]. We introduced two new factors for identifying the criminal behaviour over the social network and also proposed a new clustering algorithm according to [13], [14].

4.1. Nine Factor Analysis

In the past, social networks were formed using a seven factor analysis [14]. In that model, seven factor parameters namely frequency, duration, friends, gender, qualification, age and area were considered. Indira Priya [14] used their own ranges for all seven factors for clustering the data.

Table 1. List of Factors with Description

Parameters	Description	Range of values
Frequency	Daily session	1 = one, 2 =two, three, 3=4-6, 4 = more
Duration	Typical length of a session	1 = few minutes, 2 = up to 1 hr, 3 = 1-3, 4=>3, 5=always online
Friends	Number of friends	1=<10, 2=10-20, 3=20-30, 4=30-50, 5=50-80, 6=80-100, 7=100-200, 8=200-400, 9=400
Gender	Male or Female	1=M, 2=F
Qualification	Arts or Engineering	1= Arts, 2=Engineering
Age	Age group	18-35 = Young, >35 = Senior
Area	Continent	1=Asia, 2=Europe, 3=Africa, 4=North America, 5=South America, 6=Australia
Post status	Number of (blocked/banned) messages post in the past 1 month	1 = <25%, 2 = 25% - 50%, 3 = 50% - 75%, 4 = 75% - 100%
Share status	Number of (blocked/banned) messages shared in the past 1 month	1 = <25%, 2 = 25% - 50%, 3 = 50% - 75%, 4 = 75% - 100%

We have monitored the users' activities such as Likes, Shares and Posts. This monitoring report is also considered for taking final decision over the Social Network. They have used the qualification, age and area

factors for easy identification of the members in the given dataset. In addition to that, two more new factors were introduced namely number of posts and shares. Table 1 lists the nine important factors for identifying the criminal behaviour.

4.2. Genetic Weighted K-Means Clustering

A Genetic based Weighted K-Means Clustering Algorithm (GWKMC) is proposed in this paper for solving high dimensional multiclass problems according to [15], [16]. In the existing New Weighted Fuzzy C-Means Clustering Algorithm (NWFCMA) [13], weighted means are calculated based on all the sample points whereas in the proposed WKMC weighted mean is calculated which is based on cluster centres and the rest of sample points. As the weighted mean is calculated based on the cluster centres, this proposed algorithm is less computationally exhaustive than the existing FWCM.

4.2.1. Fitness function

Each member of the cluster population represents a competing user's feature subset that must be evaluated to provide fitness feedback to the Weighted K-Means according to [16]. This is achieved by invoking cluster on a set of training data with the particular users feature of Facebook/Weblog dataset. We aim to enhance the clustering accuracy of the recommender system which is indirectly achieved by maximizing the sensitivity and specificity of the classifier. Hence, this knowledge is incorporated into the recommender system through the fitness function of the clustering module. The fitness function is formulated as follows:

$$\text{Fitness} = a * \left(\frac{1}{\text{count of ones}}\right) + \beta * \text{Sensitivity} + r * \text{Specificity} \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

Fitness of a chromosome (Feature of Facebook/Weblog data) is evaluated based upon the sensitivity and specificity from the validation dataset and number of features present in a chromosome. Here, TP and TN are the number of records correctly classified in normal and abnormal classes respectively. Similarly FP and FN are the number of records incorrectly classified in normal and abnormal classes respectively.

4.2.2. Genetic weighted K-Means clustering algorithm

The proposed Genetic Weighted K-Means algorithm has been proposed according to the existing Weighted K-Means clustering [16] and Genetic Algorithm [15]. Our contribution is involved in genetic process so as to introduce a new fitness function for effective record selection over the given datasets. And also combined a genetic algorithm and clustering.

Step 1: Initialize the values for the parameters such as population size, the maximum number of iteration and the number of clusters, etc.

Step 2: Generates m number of chromosomes randomly.

Step 3: Each and every chromosome represents a set of initial cluster centres to form the initial population.

Step 4: A Procedure showed by every chromosome and compute weights according to the initial cluster centres perform the weighted k-means result using the fitness function (Eqn. (1)).

Step 5: Carry out the selection, crossover and mutation operator to produce a new generation of the group for each group.

Step 6: Determines whether the conditions meet the genetic termination conditions or not.

Step 7: If meet the genetic termination condition then withdraw genetic operation by agent and proceed to step 6, else go to step 5.

Step 8: Calculate the fitness of the new generation of group.

Step 9: Compare the fitness of the best individual in current group with the best individual fitness so far to find the individual with the highest fitness.

Step 10: Carry out the Weighted K-Means Clustering according to the initial cluster centre represented by the chromosome with the highest fitness and then output clustering result.

4.3. Negative Selection Algorithm

We have used Negative selection algorithm (NSA) for classifying/identifying the criminal behavioural users on Social network. NSA is a most successful method for many serious applications in the construction of the artificial immune system [12]. Initially, the standard NSA was proposed by Forrest [17] for analyze the samples. It consists of three different phases namely the data representation phase, the training phase and the testing phase. The data representation phase is responsible for represent the data in a binary or in a real valued representation. The training phase or the detector generation phase of the algorithm is randomly generate detector with binary or real valued data. In addition, they are subsequently used to train the algorithm [18], while the testing phase evaluates the trained algorithm. The random generation of detectors by a negative selection algorithm makes it impossible to analyse the type of data needed for the training algorithm. Finally, affinity matching is performed for identify the attacks. Artificial Immune System (AIS) researchers have shown that the importance and the role of affinity matching distance on NSA performance [12].

5. Experimental Results and Discussion

This section discusses about the dataset used in this work, experimental scenario and also about obtained result and discussion of the proposed system and reason for achievements.

5.1. Data Set

We have used two categories of datasets namely Facebook data that is collected manually and the weblog datasets which is available for research over the Social Network.

5.1.1. Weblog dataset

The weblog dataset is a bench mark dataset which is released for research purpose and it contains offline dataset of the particular duration of the past online database like face book, twitter, etc. This dataset has been collected from internet using the standard program which is released for retrieve the data.

5.1.2. Facebook dataset

We have prepared our own dataset from 1000 Facebook users with our own questionnaires. This dataset contains the detailed information about their Likes, Comments, Posts and shares of every individual Facebook users. We have asked many questions regarding their frequent activities such as Like, Comment, Share and Post. The questionnaires include what related message/post you give like frequently? How many like you given out of total received message/post?, How many messages you posted / shared and what type of information/post those?, etc.

5.2. Experimental Setup

We have used the Pentium IV personal computer with Intel Core i3 Processor 2.20 GHz for evaluating the proposed system. We have used two kinds of datasets for evaluating the proposed behavioural analysis model which is used by the proposed recommendation system.

5.3. Results and Discussion

The various experiments have been conducted for evaluating the proposed recommendation system. This section discusses the various experimental results obtained by the proposed recommendation system and other methods. Precision and Recall values are calculated by using the following formula.

$$Precision = \left(\frac{TP}{TP + FN}\right) * 100$$

$$Recall = \left(\frac{TP}{TP + FP}\right) * 100$$

Table 2 shows the performance evaluation of the proposed clustering algorithm and the existing clustering methods. The various experiments have been conducted by using different datasets and the precision and recall values are calculated.

Table 2. Performance Evaluation of the Clustering Algorithms

Datasets	IGA-NWFCM		GNWFCMCA		GWKMCA	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Weblog	97.23	97.23	97.92	97.34	98.32	98.25
Facebook (Present)	97.56	97.76	98.45	98.31	99.12	99.43

From Table 2, it can be observed that the proposed clustering algorithm provides better performance than the existing clustering algorithms such as IGA-NWFCM [13] and GNWFCMCA [17].

Fig. 2 shows the performance of the proposed recommendation system for identifying the criminal behavioural social network users. Five experiments have been conducted for evaluating the model which is the combination of nine factor based behavior learning, the proposed GKMCA and the existing NSA [12].

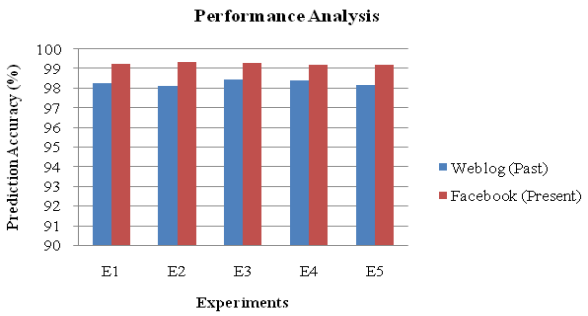

Fig. 2. Performance of GWKMCA+NSA.

Table 3. Comparative Analysis

Data Set	GWKMCA		GWKMCA+NSA	
	Precision	**Recall**	**Precision**	**Recall**
Weblog (Past)	98.32	98.25	98.32	98.25
Facebook (Present)	99.12	99.43	99.12	99.43

From Fig. 1, it can be observed that the proposed classifier better performance on Facebook dataset which is collected by own when it is compared with other dataset. The reasons for the changes of various performances over every experiment are based on the consideration of dataset for the particular experiment.

From Table 3, it can be observed that the performance in the form of calculating precision and recall for GWKMCA and GWKMCA with NSA. We have conducted five experiments for evaluating these two categories of methods. Finally, calculate the average value of precision and recall separately based on the given results

using the past and present datasets.

The reason for this performance difference is the introduction of two factors for behavioural learning, the uses of new fitness function and performed genetic operation during clustering process and the uses of Negative selection algorithm. The decision making agent also contributed reasonably for improving the performance over the dataset with the help of knowledgebase. The recommendation system consists of all the above said methods and it identified / predicted the criminal behavioural users correctly over social network. Here the task of recommender system is to forecast and detect the criminal behavioural users.

6. Conclusion and Future Enhancements

Recommendation system has been proposed and implemented in this paper for recommending the criminal behavioural users on social network. The proposed recommender system uses the proposed nine factor analysis method, clustering technique called Genetic Weighted K-Means clustering (GWKMC) and the existing classification algorithm namely Negative Selection Algorithm (NSA). The various experiments were conducted on our system using the Facebook dataset which is collected on our own along with the weblog datasets for the evaluation. Future works in this direction could be the introduction of temporal fuzzy rules for effective classification and identification of the criminal behavioural users over Social Networks.

References

[1] Drezewski, R., Sepielak, J., & Filipkowski, W. (2015). The application of social network analysis algorithms in a system supporting money laundering detection. *Information Sciences, 295*, 18–32.

[2] Alazab, M. (2015). Profiling and classifying the behavior of malicious codes. *The Journal of Systems and Software, 100*, 91–102.

[3] Robert, C., *et al*. (2015). Criminal conduct subgroups of "aging out" foster youth, *Children and Youth Services Review, 48*, 14–19.

[4] José, I., Manzano, C., Castro-Nuño, M., & Fageda, X. (2015). Are traffic violators criminals? Searching for answers in the experiences of European countries. *Transport Policy 38*, 86–94.

[5] Lee, K., & Sunghyup, H. S. (2015). A model of behavioral intentions to follow online travel advice based on social and emotional loneliness scales in the context of online travel communities: The moderating role of emotional expressivity. *Tourism Management, 48*, 426-438.

[6] Richard, K., *et al.* (2013). From 'What the F#@% is a Facebook?' to 'Who doesn't use Facebook?': The role of criminal lifestyles in the adoption and use of the Internet. *Social Science Research,* 42, 1411–1421.

[7] Lussier, P., Bouchard, M., & Beauregard, E. (2011). Patterns of criminal achievement in sexual offending: Unravelling the "successful" sex offender. *Journal of Criminal Justice, 39*, 433–444.

[8] Lorraine, C. E., Anna, J. E., Thomas, R., & DiLillo, D. (2014). Child maltreatment and adult criminal behavior: Does criminal thinking explain the association? *Child Abuse & Neglect, 38*, 1399–1408.

[9] Dantu, R., Loper, K., & Kolan, P. (2004). Risk management using behavior based attack graphs. *Proceedings of International Conference on Information Technology*: *Coding and Computing: Vol. 1* (pp. 445-449).

[10] Benkhelifa, E., Rowe, E., Kinmond, R., Oluwasegun, A. A., & Welsh, T. (2014). Exploiting social networks for the prediction of social and civil unrest: A cloud based framework. *Proceedings of International Conference on Future Internet of Things and Cloud* (pp. 565- 572).

[11] Watson, B., Watson, A., Siskind, V., Fleiter, J., & Soole, D. (2015). Profiling high-range speeding offenders: Investigating criminal history, personal characteristics, traffic offences, and crash history. *Accident Analysis and Prevention, 74*, 87–96.

[12] Balthrop, J., Forrest, S., & Glickman, M. R. (2002). Revisiting LISYS: Parameters and normal behavior. *Proceedings of the 2002 Congress on Evolutionary Computing*.

[13] Ganapathy, S., Kulothungan, K., Yogesh, P., & Kannan, A. (2012). A novel weighted fuzzy C–Means clustering based on immune genetic algorithm for intrusion detection. *Procedia Engineering Journal, 38,* 1750-1757.

[14] Indira, P. P., Ghosh, D. K., Kannan, A., & Ganapathy, S. (2014). Behaviour analysis model for social networks using genetic weighted fuzzy c-means clustering and neuro-fuzzy classifier. *International Journal of Soft Computing*, *9(3)*, 138-142.

[15] Siva, S. S., Geetha, S., & Kannan, A. (2012). Decision tree based light weight intrusion detection using a wrapper approach. *Expert Systems with Applications, 39,* 129-141.

[16] Reda, E. M., Elsayed, S. A., *et al*. (2013). A hybrid network intrusion detection framework based on random forests and weighted k-means. *Ain Shams Engineering Journal, 4,* 753–762.

[17] Forrest, S., Perelson, A. S., & Allen, L. (1994). Self-nonself Discrimination in a Computer. *Proceedings of the IEEE Symposium on Research in Security and Privacy* (pp. 202–212).

[18] Wang, C., & Zhao, Y. (2008). A new fault detection method based on artificial immune systems, *Asia-Pac. J. Chem. Eng.*, *3(6)*, 706-711.

V. Soundarya is working as an associate professor in the Department of Computer Science and Engineering at Dhanalakshmi Srinivasan College of Engineering and Technology, Chennai. She is pursuing her Ph.D in the Faculty of Information and Communication Engineering, Anna University, Chennai, India. She received her M.E. degree in computer science and engineering from Anna University, Chennai in 2011 and B.E. in computer science and engineering from Anna University of Technology, Trichy in 2009. Her fields of interests are social network analysis, information retrieval, opinion mining, and sentiment analysis. She has published 3 papers in national/international conferences and journals.

U. Kanimozhi is currently working as a teaching fellow in the Department of Computer Science and Engineering in Anna University, Chennai, India. She is pursuing her Ph.D. in the Faculty of Information and Communication Engineering, Anna University, Chennai, India. She received her M.E. degree in computer science and engineering from Anna University, Chennai, India in 2012 and B. Tech. degree in information technology from Anna University, Chennai, India in 2010. Her fields of interests are behavioral analysis in social networks, social network analysis, text data mining, machine learning and big data analytics. She has published 3 papers in national/International conferences and journals.

D. Manjula is currently working as a professor in the Department of Computer Science and Engineering in Anna University, Chennai, India. She received her Ph.D. degree in the Faculty of Information and Communication Engineering from Anna University, Chennai, India in 2004, M.E. degree in computer science and engineering from Anna University, Chennai, India in 1987 and B.E. degree in electronics and communication engineering from Thiagarajar College of Engineering, Madurai, India in 1983. She has published three books. Her present research interests include social network analysis, machine learning, big data analytics, cloud computing, virtualization techniques, information retrieval, NLP, text data mining, parallel processing, grid computing and databases. She has published more than 200 papers in national/international conferences and journal.