Topic Categorization based on User behaviour in Random Social Networks Using Firefly Algorithm

S. Jayapratha and Dr.P. Pabitha

Abstract--- *During an intercommunication period in social network participants either upgrade the interest in a topic through their positive inputs, catchy comments, tag lines or they simply do not show any interest. The activity of the participants motive for refinement in the topic state after a certain period of time. Moreover, the scenario of interaction could be leverage as an imprecise and non-crisp scheme. So the process is all about to know the transition state of a topic over a certain period of time. There are huge amount of data available in the random social networks such as facebook, twitter etc., In these data from the conversation block is used to place the topic in concern category. These topics categorization is fully based upon FA(Firefly Algorithm) using Matlab. The firefly Algorithm gathers the data on social network and based on the information it groups the data those are all having the similar attributes.*

Keywords--- *Machine Learning, Firefly Algorithm, Clustering, Random Social Networks.*

I. INTRODUCTION

WITH rise technological improvements, it has become more than that and easier to halt connected to people care about, even if they are on the other side of the world, and the entire world as such. In [9] Social media is get instant on anything and everything, making it an integral part of our lives. In social networks, flow of discussion on particular topic, that is specified using the likes and dislikes [11], [14]. There are various social networks platform available in the online such as Bebo, Classmates, Facebook, Google+, Instagram, LinkedIn, Myspace, Path, Pinterest, Whatsapp, Reddit, Twitter and Stumbleupon. Topics are classified by the interesting topic and not interesting topic based on the user behavior [5]. Fig.1 shows the social networks architecture. There are six users connected together that the form is called network. It each user does not connected in the entire user. So, it is also known as random social networks [25]. In [23] Uncertain data prediction are to find the which topic is user interested topic and the which topic is user not interested topic using the firefly algorithm. There are many machine learning algorithms using the existing research works. In [1], [2] Classify the topic related to user behavior using firefly algorithm. In this algorithm, calculate the fitness value and then finally produce the optimal result.

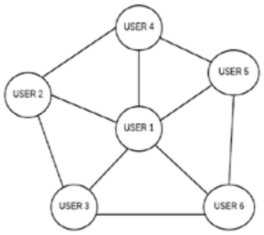

Fig. 1: Social Networks

In [7] State transition conversation over social networks fully based upon the events that currently happened or it may be solely focused upon trending social issues or emergency needs. Relevant data can be extracted from the huge amount of data involved during the discussion session. It eliminates the redundancy of data and it simplifies the sample data for training. Sample the data by comparing with the given labels then it can be able to classified as user behavior.

Select the topic which has the likes, share for input data and analyze the topic thoroughly with large amount of sample data about a topic. It will make the prediction about a topic as more accurate. During an interaction session in social network like facebook, twitter etc. The users either boosts the interest in a topic through their positive inputs catchy comments, wall posts, blogs, tag lines or simply do not show any interest [14]. Based on the scenario, the topic under discussion is broadly classified into two classes interesting topic and not interesting topic, each topic given a topic rank lies between 0-1. Based on the active participation of the participants new topic rank is given.

Classification based on the flow of conversation among different participants on a concept at a specific period that can be classified into different classes using density based clustering technique [21]. This technique is used to form a collection of different pattern from the record set with the participants current state assigning values for their activities about the particular concept. Also it is possible to calculate the contribution of the participants towards a topic.

In [16] social networks it could be randomly collect the huge data from the participants involvement like blog entry, comment and wall post. To organize the interesting and not interesting topic grouped by using density based clustering algorithm and to identify the initial state of a topic before attending the discussion session. These types of data are classified into the topics, that data does not produce the accurate results. Based on the collected data, it is hard to arrive at any decision.

S. Jayapratha, PG Student, Dept. of Computer Technology, Anna University, MIT Campus, Chennai, India

Dr.P. Pabitha, Assistant Professor, Dept. of Computer Technology, Anna University, MIT Campus, Chennai, India.

DOI:10.9756/BIJSESC.8389

The remainder of this paper is organized as follows. Section II discusses the briefly some related work applying firefly algorithm in similar setting. Our approach is detailed in section III. Experiments and results are provided in section IV. Section V is dedicated to a discussion of conclusion and future work.

II. RELATED WORK

Kraetzl et al. [24] describes the arbitrary degree distribution is used to random graph model. In [25] Random graph model give the simple unipartite networks is acquaintance networks and bipartite networks is affiliation networks. Netwoks examples are friendship, families and business relationships. In particular networks consult the properties of clustering, diameter and degree distribution with respect to these models. Charu C. Aggarwal [23] discusses the uncertain data management is a traditional database management. In this traditional database management contains the join processing, query processing, selectivity estimation, OLAP queries and indexing. Mining problems are frequent pattern mining, outlier detection, classification and clustering. The data may contain errors or may be partially complete the data.

Mitchell [6] describes the quick development is certainly occurs using the Machine learning and data mining concepts. Machine learning concepts are capable of extracting valuable knowledge from large data stores. It is discovery of models, patterns and other regularities in data. Machine learning approaches categorized into two different approaches are the statistical or pattern-recognition methods including k-nearest neighbor or instance-based learning, Bayesian classifiers, neural network learning and support vector machines. Seigo Baba et al. [11],[14] discusses the social media can classified the retweets without text data, these method used demerits are the large amount of calculations. Users who retweet the same tweet are interested in the same topic, it can classify tweets similar interests based on retweets. It related tweets based on retweets to make a retweet network that is connects similar tweets and extracted clusters that contain similar tweets from the constructed network by our classification method.

Florian Michahelles et al. [9] gives the additional marketing channel that contains the professional marketing and traditional marketing . Demerit is small dataset extracted from only one facebook page. First analyze the user discussion with topics, intentions for participation and emotions shared by the users on a facebook page. Second, analyze the user activities and interactions in terms of their evolution over time and dependency based on the community size. Further, the user classification is based on the interaction patterns and then analyzed the interactions in relation to the community.

Iztok Fister et al. [2], [10], [13] tells about the Firefly Algorithm (FA) is a one of the optimization algorithm which is based on the social flashing behavior of fireflies. There are various optimization algorithm available such as Artificial Bee Colony, Cuckoo Search Optimization algorithm, Ant colony Optimization and Particle Swarm Optimization. In this algorithm, the flashing light helps fireflies for finding mates and attracting their potential prey and protecting themselves from their predators. The swarm of fireflies will move to brighter locations and more attractive locations by the flashing light intensity that associated with the objective function of problem considered in order to obtain efficient optimal solutions. Athraa Jasim Mohammed et al. [24] tells about the content of online social netwoks communicated through text, blogs, chats, news dynamically updated. Relevant information are grouped by one clusters based on the similar attributes. A text clustering that utilizes firefly algorithm is introduced. The proposed, aFA_{merge}, clustering algorithm automatically groups text documents into the appropriate number of clusters based on the behavior of firefly population and cluster combining process [4]. Dataset are the twenty newsgroups and Reuters news collection.

Thomas Meller et al. [1] proposed to two nearest neighbor classification algorithms like naïve bayes algorithm and J48 algorithm for making recommendations to students based on their academic history. Goldina Ghosh et al. [22] discuss about the density based clustering technique is able to classified the four different topics. These classes are Interesting Motivating, Not Interesting Motivating, Interesting Deviating and Not Interesting Deviating. In the ranges between the 0-1, then Interesting Motivating range is 1-0.8, Not Interesting Motivating range is 0.8-0.6, Interesting Deviating range is 0.5 and finally Not Interesting Deviating range is 0.4-0.2, below 0.2 is noise.

Bo Wu et al. [16] propose the analyzing users behavior social roles to specify the collective decision-making process. These model propose the three layers are relation layer, individual profile layer, content layer. In this layers related to on real world-based roles and cyber world-based roles. The relation layer are the negotiation and voting. Individual profile layer are the negotiation, voting and opinion collecting. Content layer are the voting and opinion collecting. These are all the decision-making process [15]. S. A. M. Felicita tells about the achieved optimal resource utilization, maximize throughput, minimize response time. Mosab Faqeeh et al. [18], [21] proposed to document classification, there are comparing the classification algorithms such as Naïve Bayes, Support Vector Machines, Decision Trees. These are classified into the two topics through the facebook comments.

Athraa Jasim Mahammed et al. [13], tells about the drawback of swarm-based algorithms such as particle swarm optimization and ant colony optimization. Proposed another swarm algorithm is firefly algorithm in text clustering. In this algorithm discuss for two different ways, that namely weight-based firefly algorithm (WFA) and weight-based firefly algorithm. weight-based firefly algorithm is a more restricted condition finding user of a cluster compared to WFA. R. Senthamil Selvi, M. L. Valarmathi [20] discusses the efficient feature selection in the improved firefly heuristics. There are following domains using the big data health care machine, bank transaction, social media. Problem of big data is NP-hard and accordingly search intractable. There are huge amount of twitter dataset available in the online.

III. System Design

Proposed architecture is given Fig. 2 shows process are the topic categorization based on user interest using the firefly algorithm. Fig. 2 tells about the various types of topics under discussion may be interesting or not interesting. The participants discussing about the topic can make an interesting topic more interesting, that is, motivating others to talk in support of the topic. In other case, a not interesting topic can also be made interesting by the participants intervention or it may deviate away.

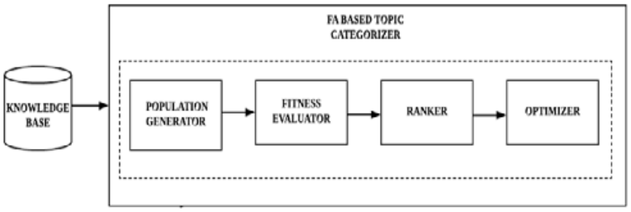

Fig. 2: Firefly Algorithm based Topic Categorizer

Topic categorization using firefly algorithm related components are the population generator, fitness evaluator, ranker and optimizer.

Population Generator

First analyze, the firefly population generation. That is specified in the input dataset. Inputs are facebook user discussed topics through comment and posthour. It is used by the interesting or not interesting that particular topic. Also, specified the maximum generation of fireflies and input range. This is known as generation of fireflies.

Fitness Evaluator

Fitness evaluator are evaluates by using the formula shown in below. That is calculated light intensity value. Input comment is greater value than posthour. so, calculated fitness value or else input posthour is greater value than comment, that is calculated movement of fireflies. Light intensity shown in eq.(1),

$$I=I_0e^{-\gamma r2} \quad \text{....(1)}$$

Then, find the movement of firefly. It is less brighter location move to the higher brighter location in eq.(2),

$$x^{comment} = x^{comment} + I * (x^{posthour} - x^{comment}) + \alpha\varepsilon_{comment} \quad \text{....(2)}$$

Ranker

In this step, rank the fireflies according to their fitness value. Attractiveness is calculated by using the euclidean distance in eq.(3), Distance (r) represents,

$$(x_{comment}, \qquad x_{posthour}) = \sqrt{}(x_{comment} - x_{posthour})^2 \quad \text{....(3)}$$

That is calculate the distance (r), and update the light intensity value. Compare the light intensity of firefly and moving of fireflies, in these two values ranking the current best of value.

Optimizer

Topic categorization end process is an optimization of fireflies. There are many values iterated in ranking the firefly. Finally, higher rank of firefly that the optimal result will be produced.

Firefly Algorithm for Topic Categorization

Firefly algorithm is invented by Xin-She Yang and get radiate behavior of fireflies. In this algorithm mainly focused on feature selection and clustering. Specifically, this research clustering the topics using firefly algorithm through user comment and posthour.

```
Algorithm: Firefly function fireflies(comment, posthour);
Input: Two inputs are comment and posthour, where (posthour>comment)
Output: clusters using comment and posthour.
Step1: Objective function f(x), x = ($x_1$, ..., $x_d$)
Step2: Generate initial population of fireflies $x_{comment}$=(comment = 1, 2, ..., n) and $x_{posthour}$ = (posthour = 1, 2,…., n)
Step3: Light intensity $I_{comment}$ at $x_{comment}$ is determined by f($x_{comment}$)
          $I = I_0e^{-\gamma r2}$
Step4: Define light absorption coefficient, initial $\gamma$=1
Step5: Define the randomization parameter $\alpha$=0.2
Step6: Define initial attractiveness $I_0$=1.0
Step7: while (t < MaxGeneration)
Step8: for comment = 1 : n all n fireflies
Step9: for posthour = 1 : comment all n fireflies
Step10: if ($I_{posthour} > I_{comment}$), Move firefly comment towards posthour in d-dimension;
          $x^{comment} = x^{comment} + I * (x^{posthour} - x^{comment}) + \alpha\varepsilon_{comment}$
Step11: end if
Step12: Attractiveness varies with distance r via exp[−r] Distance (r)
        ($x_{comment}$, $x_{posthour}$) = $\sqrt{}(x_{comment} - x_{posthour})^2$
Step13: Evaluate new solutions and update light intensity
Step14: end for posthour
Step15: end for comment
Step16: Rank the fireflies and find the current best
Step17: end while
Step18: return Postprocess results and visualization.
Step19: end
```

Fig.3: Pseudo Code of Firefly Algorithm for Topic Categorizer

Firefly algorithm process are put the input data, population generator, fitness evaluator, ranker and optimizer. Initially define the values are light absorption coefficient, randomization parameter and attractiveness. MaxGeneration means maximum value of input dataset to be generated. Comment range between the 1 to n and posthour range between the 1 to comment both for all n fireflies. Check $I_{posthour} > I_{comment}$ satisfy this constraints their automatically produce result, not satisfy this constraints move from less brighter location to higher brighter location using eq.(2). In Fig.3 shows the firefly algorithm, inputs are facebook conversation through comments and output are the topic clustering. Initial input must be 1 to n rank the firefly then finally produce optimal result.

IV. Experiments and Results

In this section discuss about the three dimensional view of surface, moving of fireflies and topic categorization. First, input data clustered by four groups and then fireflies are moving into the center point that means path tracing. Finally, similar topics are grouped itself.

Surface in 3-D View

In this graph four topics in the groups are classified in the surface. Fig.4 shows the 3-D view of the graph, x-axis show the input of comment, y-axis show the input of posthour and z-axis show the surface. That the range between the -5 to 5. In each and every point clustered in using the grid surface of the function. There are 100 grid lines specified in this graph.

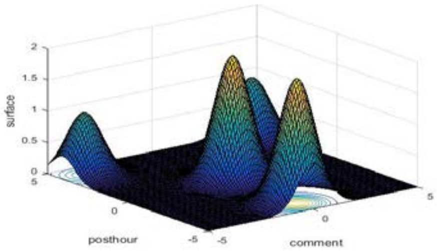

Fig. 4: 3-D View Surface

Grid value is 100, that use in the value plot of input produce the accurate results. The center of the surface value is 0 and x-min, y-min value is -5 and x-max, y-max value is 5. These four groups based on the distance between radius values. Each peak value represents the outside of surface with maximum value. There are four values 0.835, 1.3, 1.9 and 1.07.

Moving of Fireflies

In the fireflies are step by step moving the data. It group the same topics are one clusters. In graph x-axis are the comment and y-axis are the post hour, that specifies the number of users reply and post in hours.

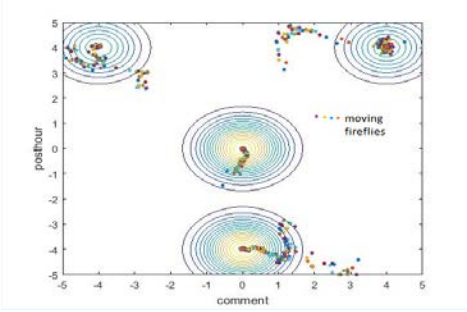

Fig. 5: Fireflies Moving in the Center

Input to give the number of users participants in the discussion session that based on the users comment and share of the communication. In Fig.5 shows tracing the moving of fireflies. Two clusters are between the same range and x-axis center of the cluster value is zero then same as y-axis center of the cluster value is zero. In other two cluster are top of the graph, x-axis and y-axis center of the cluster value is 4. All fireflies are stimulating to the center of the cluster value 0 and 4.

Grouping of Topics

The grouping of topics in x-axis are the comment and y-axis are the posthour, same topics are grouped by the one type of cluster. Fig.6 shows the topics categorization. The fireflies are moving the particular range, that arranging the fireflies moving to center of the clusters. In circle specify the clusters, particles are the fireflies, that particles closed into the contour function.

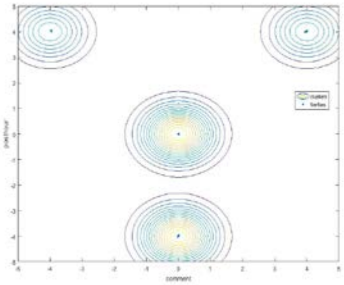

Fig. 6: Topic Categorization

Similarly in the each clusters moving to the same attributes of fireflies. It is using the contour function, that contour function specifies the number of rounds in clusters.

Performance Parameter

The inputs and outputs are specified the multiple users, that users comments and posthour with the number of inputs and outputs. Twelve inputs and maximum of fireflies is 58. In output values are produce the positive and negative values with the multiple users. Accuracy measured calculated by using the sensitivity (True Positive Rate), specificity (True Negative Rate) and F-score (Accuracy) with three dimensional view of the surface. Sensitivity and Specificity ranges are 0 to 1 and F-score range is the 0 to 4 in Fig.7. Sensitivity value is 1, Specificity value is 0.1 and F-score value is 3.2.

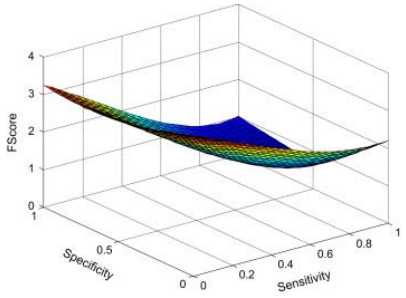

Fig.7: Comparison of Sensitivity, Specificity and F-Score

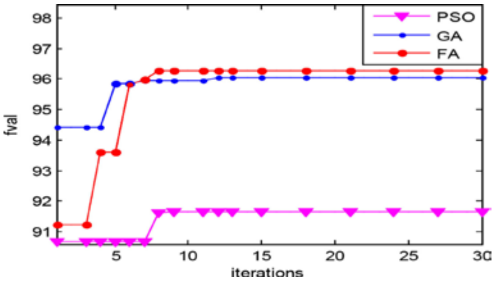

Fig.8: Comparison of Fitness Value with PSO, GA and FA

In Firefly Algorithm compare with the algorithm of Particle Swarm Optimization(PSO) and Genetic Algorithm (GA). Firefly algorithm accuracy value is high examine to

PSO and GA. In Fig.8 shows the test result analysis. In x-axis define the thirty iteration iterated from the dataset and y-axis define the fitness value of three algorithms in comparison graph.

Table 1: Result Analysis for Topic Categorization

INPUT (Multiple Users)		**OUTPUT (Multiple Users)**		
Number of comments	Number of posthour	Number of comments (x-axis)	Number of posthour (y-axis)	Surface (z-axis)
3	4	3.94	3.98	0.99
10	5	-3.97	4.05	0.99
3	0	4.00	4.03	0.99
10	50	4.00	4.02	0.99
58	3	3.99	4.00	0.99
19	9	0.01	-3.95	1.99
1	3	-0.02	-4.01	1.99
3	9	0.01	-0.00	1.99
0	3	0.00	-3.98	1.99
0	10	-0.00	-0.00	1.99
3	6	0.00	-0.00	2.00
4	5	-0.00	-3.99	1.99

Topic categorization inputs are the comment and posthour for the multiple users and similarly on outputs then additionally, surface of the clusters in Table.1. These values are plot the cluster of the topic. Output values range between the 0 to 1 based on light intensity and attractiveness. If the value is zero firefly grouped by center of the cluster.

V. Conclusion and Future Work

Firefly algorithm for classification was able to classify the topics according to the participants activity used for the blogs, comments, photos. It was able to classify the topics based number of users that communicate in the particular topics. The numeric values of share and comments are produced for obtaining the results. The topic similarity is measured within the range of the comment and posthour, the range of value is calculated using the firefly algorithm and ranking is done based on value for fireflies. Propose to select the features from the classification data which is necessary to identify whether the topic either interesting or not interesting. Predict the best possible decision based on the conversation of different participants in random social network using ranking method, it will take a long period time. So, it will enhance the limited time period.

References

[1] A. Lamba and D. Kumar, "Optimization of KNN with Firefly Algorithm", BIJIT-BVICAM's International Journal of Information Technology, Vol. 8 No. 2, 2016.

[2] A.J. Mohammed, Y. Yusof and H. Husni, "Determining Number of Clusters Using Firefly Algorithm with Cluster Merging for Text Clustering", International Visual Informatics Conference, Pp. 14–24, 2015.

[3] C.C. Aggarwal, "Network analysis in the big data age: Mining graphs and social streams", IBM T J Watson Research Center, ECML/PKDD, 2014.

[4] M.D. Choudhury, A. Monroy-Hernandez and G. Mark, "Narco Emotions: Affect and Desensitization in Social Media during the Mexican Drug War", CHI, 2014.

[5] M. De Choudhury, W.A. Mason, J.M. Hofman and D.J. Watts, "Inferring relevant social networks from interpersonal communication", Proceedings of the 19th International Conference on World Wide Web, Pp. 301–310, 2010.

[6] E. Elmurngi and A. Gherbi, "An Empirical Study on Detecting Fake Reviews Using Machine Learning Techniques", International Conference on Innovative Computing Technology, 2017.

[7] Gómez, A. Kaltenbrunner and V. Lopez "Statistical analysis of the social network and discussion threads in Slashdot", Proceeding Proceedings of the 17th International Conference on World Wide Web, Pp. 645–654,2008.

[8] http://en.wikipedia.org/wiki/Cluster_analysis,wikipedia.

[9] I.P. Cvijikj and F. Michahelles, "Understanding the user generated content and interactions on a Facebook brand page", Int. J. Soc.Humanist. Comput., Vol.2, No.1–2, Pp. 118–140, 2013.

[10] I. Fister, I. Fister Jr, X.S. Yang and J. Brest, "A comprehensive review of firefly algorithms", Swarm and Evolutionary Computation, Vol. 13, Pp. 34–46, 2013.

[11] J.W. Treem and P.M. Leonardi, "Social media use in organizations exploring the affordances of visibility, editability, persistence, and association", Commun. Yearb., Vol.36, Pp. 143–189, 2012.

[12] A. Kaltenbrunner, V. Gomez and V. Lopez, "Description and prediction of Slashdot activity", Proceedings of the Latin American Web Conference, IEEE Computer Society, Pp. 57–66, 2008.

[13] K. Bhatt, A. Singh and D. Singh, "An Improved Optimized Web page Classification using Firefly Algorithm with NB Classifier", International Journal of Computer Applications, Vol. 146, No.4, 2016.

[14] E. Kiciman, M. De Choudhury and B. Thiesson, "Analyzing social media relationships in context with discussion graphs", Eleventh Workshop on Mining and Learning with Graphs, ACM, 2013.

[15] I. King, M.R. Lyu and H. Yang,, "Online Learning for Big Data Analytics", Tutorial presentation at IEEE Big Data Santa Clara, CA, 2013.

[16] M. Lieberman, "Visualizing big data: Social network analysis", Data, Proceeding of the CASRO Digital Research Conference, on line+ Mobile+Big San Antonio, Texas, 2014.

[17] X. Ling, Q. Mei, C.X. Zhai and B Schatz, "Mining multi-faceted overviews of arbitrary topics in a text collection", Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pp. 497–505, 2008.

[18] R. Duriqi, V. Raca and B. Cico, "Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA", Mediterranean Conference on Embedded Computing, 2016.

[19] S. Arora and S. Singh, "The Firefly Optimization Algorithm: Convergence Analysis and Parameter Selection", Int. J. of Com. Appli, Vol. 69, No. 3, 2013.

[20] V. Subha and D. Murugan, "Opposition-Based Firefly Algorithm Optimized Feature Subset Selection Approach for Fetal Risk Anticipation, Machine Learning and Applications", An International Journal, Vol. 3, No. 2, 2016.

[21] T. Meller, E. Wang and F. Lin, "New Classification Algorithms for Developing Online Program Recommendation Systems", Int. C. on Mobile, 2009.

[22] V. Virgilio, "Exploring Big Data in Social Networks", Key Note address INWEB-National Science and Technology Institute for Web Federal University of Minas Gerais-UFMG, 2013.

[23] Y. Cheng, R. Chi and S. Zhu, "An Uncertain Data Model Construction Method Based on Non-parametric Estimation", IEEE International Conference on Electronic Information and Communication Technology, 2016.

[24] Y. Zhou, X. Guan, Z. Zhang and B. Zhang, "Predicting the tendency of topic discussion on the online social networks using a dynamic Probability model, webscience", Proceedings of the Hypertext Workshop on Collaboration and Collective Intelligence, 2008.

[25] M.E.J. Newman, D.J. Watts and S.H. Strogatz, "Random graph models of social networks", Proceedings of the National Academy of Sciences of the United State of America, Vol. 99, 2002.