SOCIAL NETWORKS USER MODELING

Wasan Abdallah* **Shahenda Salah El-Din Sarhan**** **Samir EldesokyElmougy***

* Department of Environment - Faculty of Science, Al-qadisiyah University, Al-qadisiyah, Iraq
** Department of Computer Science- Faculty of Computers and Information, Mansoura University, Mansoura, Egypt

Shahenda_sarhan@yahoo.com samirelmougy@yahoo.com wasan.alawsi@qu.edu.iq

Abstract:

Social Networks (SN) is a source of great interest to researchers in the fields of scientific research, because of spread a very large number of SN and the attention of quite a lot of peoples, as well as to the content of SNS of a large quantity and variety of transmitted data between one site and various sites (such as photos, messages, and personal information, news, websites, scientific research, and other information). Within this study we study the attributes and activities of users of social networking sites (Facebook) and identify the most important and effective elements in activating the site, where accumulate information on social sites users generally speaking and information on facebook in particular, the methodology incorporates a two-stage cross mechanism. neural network for clustering data and rough sets theory (Johnson Reducer and naïve Bayes Classifier) for classification and analysis. The simulation model which we carried out, demonstrated really good brings about the diagnosing process that contacted (0. 94)% of the accuracy diagnosis..

Keywords: *user modeling, social network modeling Artificial Neural Networks, Rough set, Reducts.*

1. Introduction:

user models, the suitability of the data, the noise within that data, and the necessity of taking the imprecise nature of human behavior.

Data mining and machine learning techniques have the ability to handle large volumes of data and to process improbability. These characteristics make these techniques suitable for automatic generation of user models that simulate human verdict making [2]. In the Internet stage of development, we have witnessed the rapid growth of social network sites (SNS) such as Facebook, MySpace, LinkedIn, and Orkut, in recent years [3].

Users across the world have signed up for accounts on SNWs in order to discover other people with similar interests or experience, to share

A SN service is an online service platform that helps in building and maintaining social relations among people with similar interests and/or activities. What makes SNS (SNS) unique is not only that they allow individuals to meet strangers, but rather that they enable users to articulate and make visible their SNalso. The first social networking website Friendster came up in 2002 and became a big success.

The SNS has about 1.5 billion parts of information present on Facebook, more than 140 million of tweets present on Twitter, over 2 million of videos are present on YouTube and nearly 5 million of photos in Flickr [1].Some of the difficulties that user modeling faces are the amount of data available to create

the proposed model, results, and discussion are introduced next in Section 4 and the last section present our conclusions and related works.

1.1 Self-Organization Map:

The Self-Organizing Map (SOM) was blueprint as an unsupervised competitive learning algorithm of the artificial neural networks is additionally named a Kohonen map add up to the explorer'sthe Finnish Professor (Teuvo Kohonen, 1980).

SOM algorithm is a standout amongst the most capable calculations in information perception and concentrate on and is broadly utilized in clustering analysis in data mining. Its ability to guide high-dimensional info vectors onto a two-dimensional framework of model vectors and requests them topologically significantly encourages the elucidation by bare human eyes [5].

1.2 Rough Set Theory:

RST was makeknown to by Polish logician, Professor Z.Pawlak in mid 1980s. It is an expansion of the set hypothesis for the investigation of clever framework described by mistaken, indeterminate or ambiguous data and can help as another numerical device to softcomputing [6].

The RST has been down to earth in more than a couple fields image processing, data mining, expert systems, pattern recognition, knowledge discovery and medical informatics. In the cutting edge writing, a few exploration works have been joined the RST with other manmade brainpower strategies for example neural network[7].

Interestingly, if the information is sure, rough sets can figure out if there are any developments in the information and locate the base required for characterization of personal information with both friends and strangers, every marketer knows that the most modern barometers of popular culture are SNS like Facebook. Along with other forms of computer mediated communication, they have transformed consumers from mute, sequestered and invisible individuals, into a noisy, public, and even more impossible than usual, collective. At the same time, grappling with social media strategies has been difficult for many concerns [4].

The explanation of the importance of this study at following points:

1. Find and discover the side of the concerns of Facebook users.
2. Identification of the link between the personal aspects demographic profile of users and potential roles that appear in their behavior online.
3. Determine the possibility that SN is a substitute for social relations.
4. Knowledge discovery from databases and convert raw data into useful and the use of modern and sophisticated algorithms to build the model users to SN by using a new community sample.

Obtaining the user's concerns within the daily activity on the internet is a difficult task because of the complexity of the data, the fact that the user's activity may vary from one to another location, and transmitted large amount of data in the social networks, which adds considerable complexity in the process of building a model for users of SNS. These difficulties encouraged us in this work to use NNs with RST to get the lowest possible number of attribute, which keeps the dependency of the subject.

This work is structured as follows. Section 2 explores the background and related work while Section 3 provides a brief proposed methodology (data collection, preprocessing, clustering, reduction, and classification). Then,

recurrence numbers from an "expert" decision table.

The NB classifier frequently works extremely well in preparing, and excellent order results might be obtained notwithstanding when the likelihood gauges contain extensive errors [10].

Most widely utilized classifier is the NB. This classifier constructed the possibility of probabilistic Classification where the likelihood is a figure for every archive. It demonstrates the having a place with the classifications indicated [11][12][13]. Numerous techniques utilizing NB classifier, multinomial NB is utilized where the likelihood P(Cj|di) of a record has a place in the classification.Cj is figured through the accompanying equation:

$$p(c_j \mid d_i) = \frac{p(c_j)\Pi_{k=1}^{|d|} \mathrm{P}(w_{di,k} \mid C_{j)}}{\sum_{r=1}^{c} p(c_j)\Pi_{k=1}^{|dj|} \mathrm{P}(w_{di,k} \mid C_{r)}} \ldots\ldots\ldots\ldots\ldots (2)$$

category is calculated according to the equation.

where $|d_i|$ be the length of the document .|C| is the number of categories. P(Cj) is the probability of

$$p(c_j) = \frac{1 + \sum_{-1}^{|D|} p\left(c_{j \mid d}\right)}{|C| + |D|} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The probability of attribute gives that the category occurred $P(w_i|C_{)})$is calculated by the equation.

$$p(Wi \mid \mathrm{Cj}) = \frac{1 + \sum_{i=1}^{|D|} N(Wi, di) P\left(yi = c_{j \mid di}\right)}{|V| + \sum_{i=1}^{|D|} \sum_{j=1}^{|D|} N(Wi, di) P\left(yi = c_{j \mid di}\right)} \ldots\ldots\ldots\ldots (4)$$

Given the enormous expansion in social sites giving them a variety of different functions, which helped researchers to delve into the different types of modeling.

In [15] Empirical study of Facebook users (n = 182) revealed that Intention to use online SN is strongly determined by social presence. Among

where |D| is the number of objectin the training set, |v| is the number of the attribute in the training set [14].

2. Related Works:

The growth and popularity of online SN have created a new world of collaboration and communication,

information. This property is essential for applications where learning is exceptionally restricted or immoderate information accumulation/challenging in light of the fact that it makes of the information gathered is sufficient just to fabricate a model decent evaluating without giving up exactness or squandering time and push to assemble extra data about the items [8].

In RST, can be lessened the measure of the dataset either by showing an entire class given by a confusion connection (by a method for a solitary component of this class) or by erasing characteristics that don't participate to the classification [9].

1.3 Naïve Bayes (NB) Classifier:

The probabilities convoluted in creating the last gauge are figured as

and dynamical processes taking place on the networks and reports improvements on the structure and function of complex networks. Mathematical models based on epidemiological procedures have influenced the research on information diffusion.

In [21][22] In user classification, Individual Parameter Estimates fluctuate less over time than they do across individuals.

In [3] social networks, studies show that Multi-User Queries result in multiple operations and are expensive in terms of performance.

3. The Proposed Methodology:

The observation and analysis of user behavior on the web is usually a preliminary stage to infer information about user interests and preferences. Adaptation and personalization is an important part of the modeling to learn about user behavior. But the complex part is the user behavior that varies depending on the type of network that can be cared the same users on Twitter to specific event, in YouTube to music or action film, and in Wikipedia for literature subject. In addition, user's interests are different within a certain period of time that may be radically from another time period. Therefore, in this thesis, a common general model SN user is introduced and implemented.

Figure (1) illustrates the general workflow of users modeling that deals with the user in the web taking into account the user's activity on other sites.

The main work stage of the proposed modeling system is shown in Figure (2) in which it is composed of six stages (data collection, data preprocessing, clustering using SOM, reduction, simulation model, and the five values, social related factors had the most significant impact on the intention to use. Implications for research and practice are discussed.

In [6] work on variation in the spread of content has been carried with a focus on categories of twitter hashtags (similar to keywords). This work is aligned with ours in its attention to importance of content in variations among popularity, however they consider categories only, with news being one of the hashtag categories.

In [16] analyzed SNS data using K-mean classification algorithm. The experimental results of text document classification on SNS dataset shows that 46% user preferred Facebook, 15% user preferred Orkut, 11% user preferred Google+, 10% user preferred Twitter, 9% user preferred Skype, 8% user preferred Hi5.

In [17] a unique extension of prior work on clustering SN users that attempts to correlate the latent roles associated with these clusters with demographic and psychological profiles. We also listed properties common to the most popular broadcasts.

In [18] the problem of joint modeling of users' queries and clicks in the search log data, and proposed a generative model.

In [19] important research on social media, has fixated on the measurement and analysis of network structures, user interactions, and traffic characteristics of social media with experimental approaches which use data mining and statistical modeling schemes. There is a significant exertion to use mathematical models to understand and predict information flow over a period in online social networks.

In [20] discussed dynamical processes on complex networks, dynamical models of network progress

uncorrected forms, translating character data to digital value, storing in excel file, and completion (in Rosetta) as shown in Figure (3).

To predict the user behavior from dataset, we had selected (16) attributes (age, gender, education, post average, average Facebook time, average internet, time, friends number, friend like sharing, find new friends, interest play games in Facebook, proposal page, interaction, research, like, preference, post).

classification) which are discussed below.

3.1. Data Collection:

The used dataset was collected from 680 Facebook users through making information form (family, friends circle, and college students from the third and fourth years at the Faculty of Computers and Information, Mansoura University, Egypt.

3.2. Data Preprocessing:

The preprocessing stage steps are divided into removing blank and

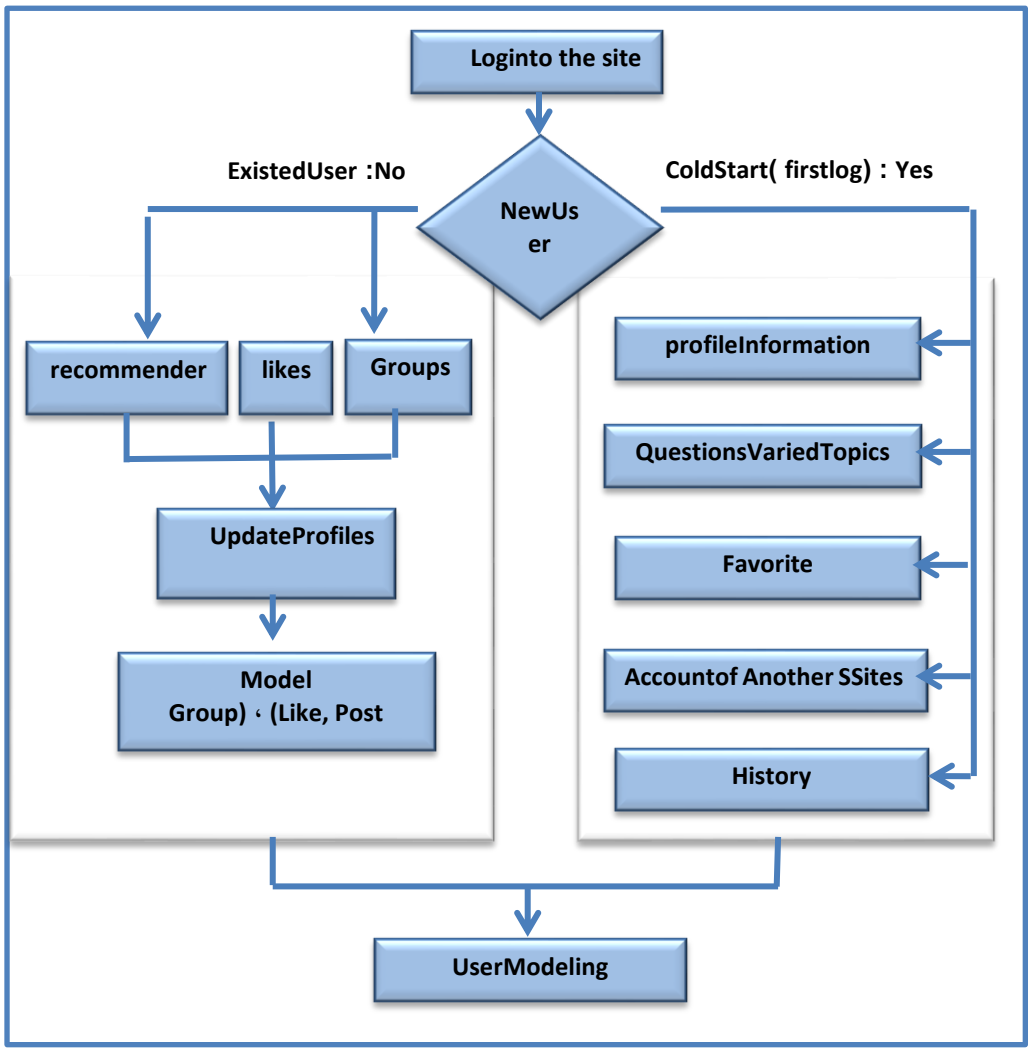


Figure 1: The general workflow of users modeling in social networks

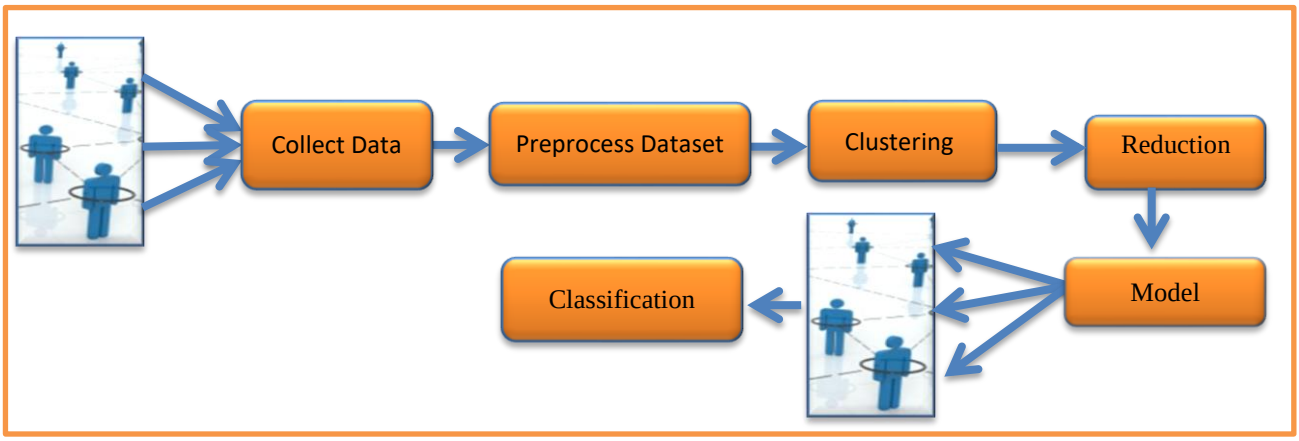


Figure 2: Work stage of apply the proposal model

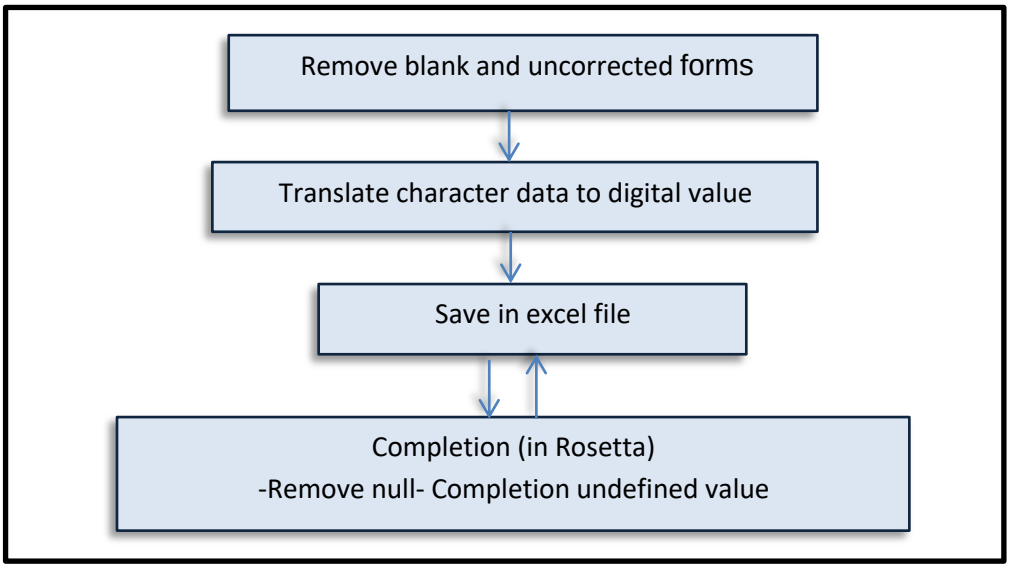
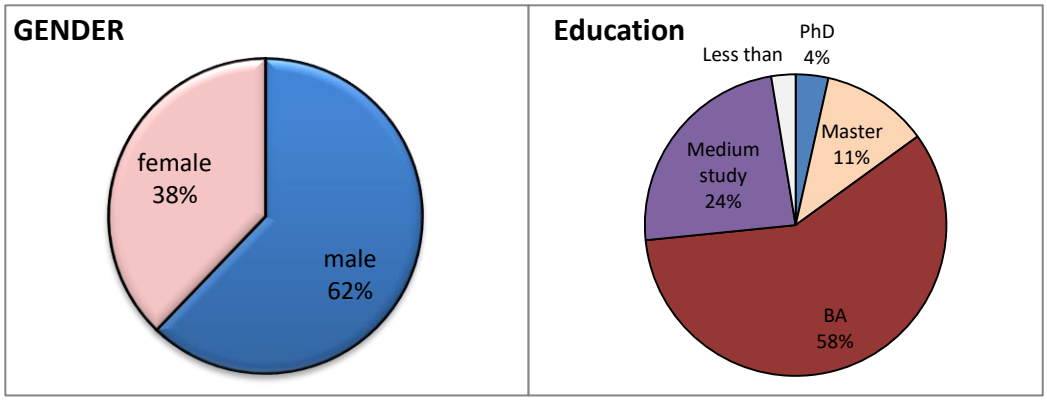


Figure 3: The preprocessing data stage



A B

Figure 4-A: The gender distribution of users in the dataset
B: The educational distribution of users in the dataset

9. **Proposal page:** How much proposal page is helpful to the user?range (0,5).
10. **Facebook games:** It measures user interests of the communion game in Facebook site, range (0,5).
11. **Search for new friends:** The user attention for searching new friends in the verity site, range (0,5).
12. **Share like friend:** It measures user attention to new or unknown publication, range (0,5).
13. **Topics preference**: Some of the main topics that users used in social sites (blog, world news, local news, sciences, technology, business, sport, industry, lifestyle, universities, jobs, music, celebrities community, art, health, fine stuff, game, video, event, entertainment, travel, shopping, religion, programming, commentators, literature, history, geography, design, family and child and general culture), range (1,32).

As a result of data collected using information form and after preprocessing, we got a set of characteristics that had been specified as follows: -

1. **Age** (13-52).
2. **Gender** (male, female), range (0,1).
3. **Education**: Including PhD, Master, BA, Medium study, Less than, range (1,5).
4. **The average daily use of the network**: It represents the daily average time spent by the user to browse the various sites, except Facebook, range (1,5).
5. **The average daily use of the Facebook**: Represents the daily time spent browsing Facebook user rate, range (1,5).
6. **The average posts daily in Facebook**: It represents the average number of daily publications for the user within the Facebook site, range (1,5).
7. **Average number of friends**: It represents the average number of friends to the user within the Facebook site, range (1,5).
8. **Interaction:** (active, enactive), range (0,1).

16. **Search**: Same topics preference subject, range (1,32).

14. **Likes topics**: Same topics preference subject, range (1,32).
15. **Posting topics**: Same topics preference subject, range (1,32).

Table 1: Demographic information profiles

Attribute	Percentage	Value
Age		13 – 52 years, Mean 39 years
Gender	62.2 % 37.8 %	Male Female
Marital status	64.8 % 35.2 %	Single Married
Nationality	66.5 % 32.7 % 0.8 %	Iraq Egypt Other
Education	3.5 % 11.5 % 58.4 % 24 % 2.6 %	PhD Master BA Medium study Less than

select the best reduction based on reduction, rule, cardinality numbers, and support.

Each one of these algorithms has two options: full and object related. Full results are a set with less number of properties that can determine functional dependencies. Object related results are a set of decision rules or general patterns through minimal attribute subsets that discern on a per object basis.

3.5. Simulation Model:

This stage used the attributes of the resulting from reduction algorithms, usage in building a simulation model to Facebook site. By using the programming language C# and SQL, where used (32) image of different topics. The simulation model includes two page, the first is register page of profile user in the simulation

3.2. Clustering:

The following stage is to use SOM to cluster the selected attributes. It uses the 16 attributes as one input vector for each user and the maps input matrix of all input vectors to tow dimension to cluster them. The first dimension is the input vector itself and the other is the cluster type.

The cluster type is number from 1 to 6 that represents the classified (scientific, policy, religion, sporty, general culture, education). That represents anew attribute for each user based on this new matrix, rough set algorithms are applied.

3.3. Reduction:

In this stage, attributes in clustering data are reduced through applying different rough sets (GA, JA, Holte1R, Manual, SVA Genetic, Johnson Reducer, Holte1R Reducer, and Manual Reducer) algorithms to

This matrix contains information about actual and predicted classifications done by a classification algorithm. The accuracy is the proportion of the total number of predictions that were correct where TP represents True Positive, FP is the false positive, TN is the True Negative, and FN is the False Negative.

model and the second get proposal image in the simulation model.

3.6. Classification:

In this stage, NB classifier was applied to the reduced data of the previous step. The classification accuracy measure used in this experiment was computed using the confusion matrix shown in Figure (5).

		Predicted class	
		Positive	**Negative**
Actual class	**Positive**	**TP**	**FN**
	Negative	**FP**	**TN**

Figure 5: General confusion matrix [69]

daily activities) in Facebook site, for getting the lowest number of attributes which are more effective and can give us a clear idea of the interest, and recommendations of the SN users.

4.1. SOM Algorithm:

The main benefit of using SOM is to add a new feature to data property clustering. It combines similar properties within a single type.

An example of using SOM algorithm is shows in Figure (6) where the training input is16 attribute and the output is 6 clusters. Figure (7) shows the results of executing SOM algorithm where the number of objects belongs in cluster one are (142), cluster two are (139), cluster three are(83), cluster four are (111), cluster five are (117), and cluster six are (89).

4. Experimental Results:

An adaptive system capability for the creation of environments is mainly determined by the collected and correctness of the stored data in every user model. Among the problems that confront user modeling is the collecting data of more than one site for the creation of user models, the noise inside that data, and the inevitability to capture the vague human conduct nature. The methods of completion data and machine learning are capable of handling large data and processing for uncertainty.

Such features enable these methods to be fit for user models' automatic generation which mimics the decision making of a human. The data collected, it enables us to study the user attributes (favorites, profile and

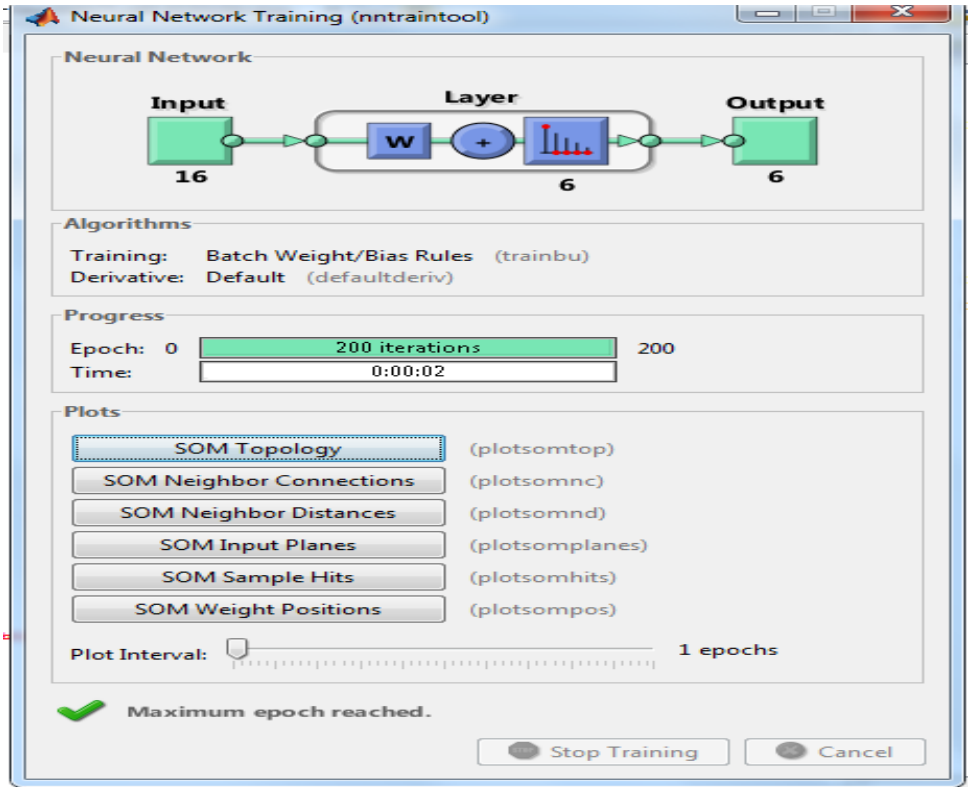


Figure 6: Example of using SOM algorithm

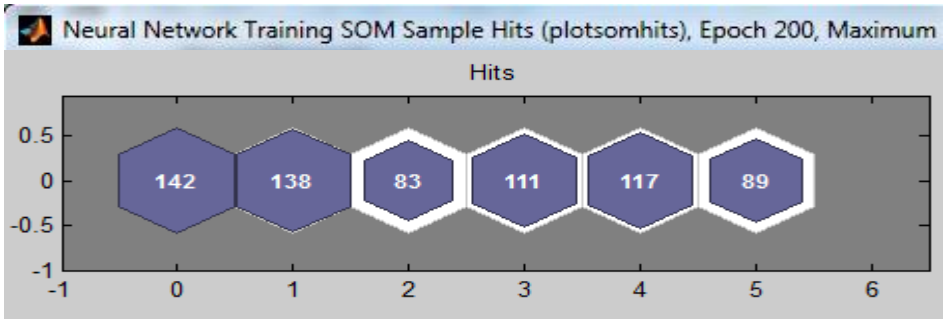


Figure 7: Results of executing SOM

output a set contains less number of properties can determine functional dependencies) is chosen. Table (2) presents the results of the implementation of these algorithms to get attribute data reduce. From this table, JA gives a better result in which it gives less number of reducts, less number of rules, less number of cardinality, and high number of support.

4.2. Rough Set:

The second stage is to apply RS with reduction algorithm (Genetic algorithm, Johnson Algorithm (JA), Holte1R, Manual, SINE-VNTR-*Alus* (SVA) Genetic, Johnson Reducer, Holte1R Reducer, and Manual Reducer), each at a time.

In this stage, reduction algorithm is applied, and the full option (intended

Table 2: Evaluation measurements of reduction and rules produced by different algorithms

Reduction Algorithm	No. reduct	No. rule	Cardinality	Support
Genetic algorithm	28	19040	7,8,9,10	100
Johnson algorithm	1	680	7	100
Holte1R	16	942	1	1
Manual	1	678	15	0
SVA Genetic	34	23120	8,9,10,11	100
JohnsonReducer	1	559	3	100
HolteRReducer	16	942	1	1
ManualReducer	1	678	15	0

reduct with three attributes. This means that the percentage of reduction is 84%.

Figure (8) shows the percentage of reduction resulting from the implementation of JA on the dataset features in which it that gives one

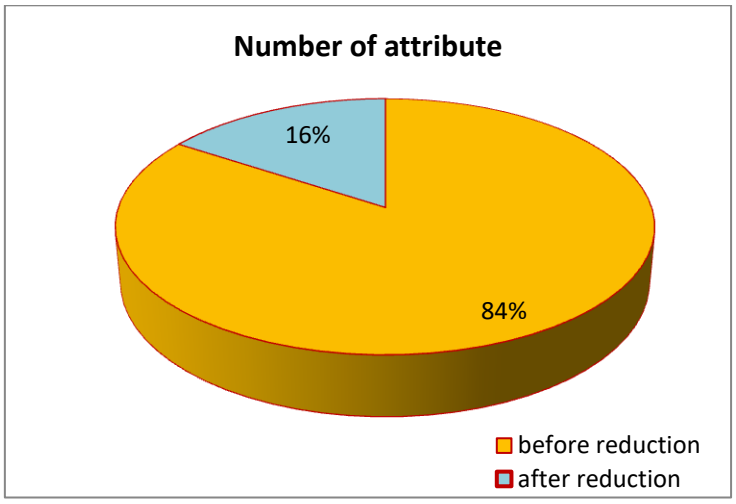


Figure 8: The percentage of the attributes before and after reduction

topics and the user selects whether the picture is (like, dislike). Result execute getting opinions to (70) users.

4.4. Discussion:

To test the proposed model, 70 participants are executing the system. They distributed: Age from 13 to 68,

4.3.Model Results:

Features that resulted from the reduction process had been used to build a simulation model. This model consists of two page. First, the user is recorder profile information (name, E-mail address, and age. Second page, a picture showing a group of different

interests represented (sport, commentators, games, hobbies, travel),

4. (18%) of the general culture and interests of the participants was represented by (blog, a variety of information, religion, literature, history, general culture),
5. (14%) of the participants with an interest in the policy field (local news, world news, politics, events),
6. (11%) trade and business, industries, jobs, shopping.

0.14% participants outside Facebook site, and (0.86%) from Facebook site.

Figure (9) shows participants descriptions in Facebook site where:

1. (21%) of the participants with an interest in the art field such as (drawing, video, celebrity news, design, life style , and music)
2. (20%) of the participants had scientific interests represented (scientific news, education, technology, software, geology, health, family and child),
3. (16%) of the participants were concerns hobbies,

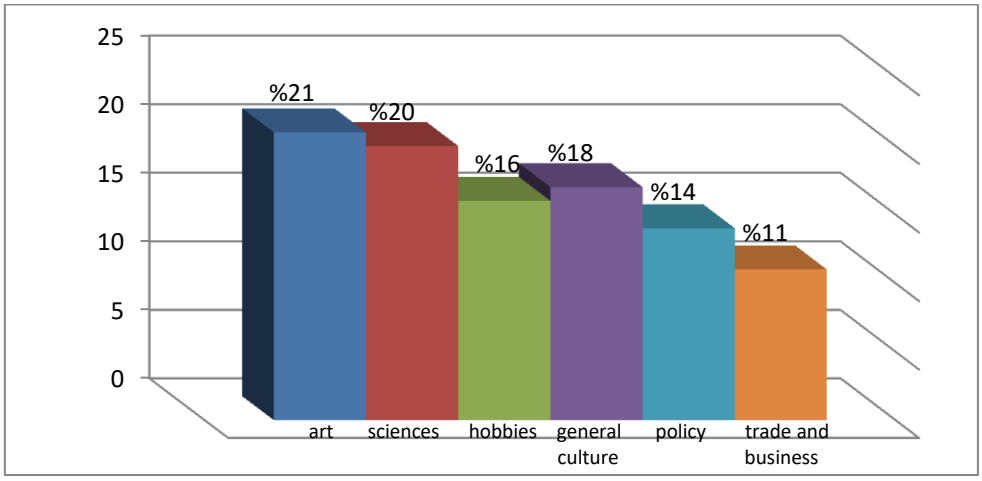


Figure 9: Description of the particular distribution in testing model

4. (11%) from participants interested with general culture.
5. (12%) of the participants their policy interests.
6. (14%) of the participants concerns of trade and business

Similarly, Figure (10) describes the participants outside Facebook site in which:

1. (26%) of the participants interest in art field.
2. (21%) participants interest concerns of scientific interests.
3. (16%) of hobbies interest.

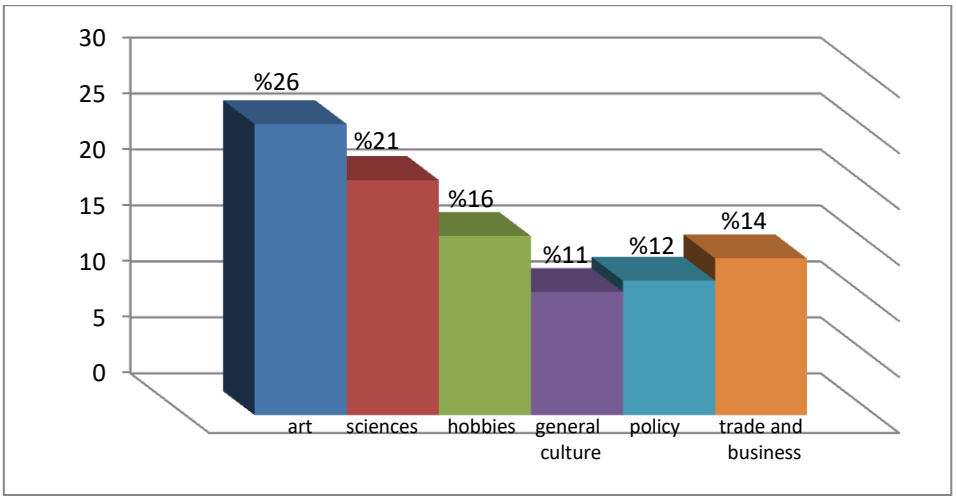


Figure 10: Description of the particular distribution in testing model (users outside Facebook site)

Finally Table (3) shows the confusion matrix in which the overall accuracy of model is (0.94%).

Table 3: The output confusion matrix

No name

		Predicted		
		0	1	
Actual	0	32	1	0.969697
	1	3	34	0.918919
		0.914286	0.971429	0.942857

Properties that have been identified of the study (age, like, friends number) candescribe a significant impact on the prediction of user's interests in different locations.The result of classification based on the used attributes in a user simulation Facebook model is 94.28%.

In future works, we plan to evaluate our work within the framework of the proposed model applied on ready dataset.

For the purpose of obtaining more data comprehensiveness of users of communication networks, social

5. Conclusion and Future works:

User modeling material in SN is not simple for many reasons, including: -

1.User information does not have to be real (the lack of restrictions on the user's personal registration information).

2.Users may use different names and information from one site to another.

3.User concerns may change from time to time (depending on the need, the psychological, the user events surrounding the private and public).

topics: idioms, political hashtags, and complex contagion on twitter", in: proceedings of the 20th international conference on world wide web. acm, (2011), PP. 695-704.

[7] Thabet Slimani, "Application of rough set theory in data mining", arXiv preprint arXiv:1311.4121, (2013).

[8]Enas Elharir,Shahenda Sarhan, Magdi Zakaria ,"A Hybrid Rough-Neuro model For Diagnosing Erythemato-Squamous Diseases",IJCSI International Journal of Computer Science Issues, 11(1), (January 2014), PP.143-147.

[9] Pessoa, Alex Sandro Aguiar, and Stephan Stephany, "An Innovative Approach for Attribute Reduction in Rough Set Theory ", Intelligent Information Management 2014, (2014).

[10] Aleksander Øhrn, "Rosetta technical reference manual", Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway (2000), PP.1-66.

[11] Bhatt, Rajen B., and M. Gopal, "On the compact computational domain of fuzzy-rough sets", Pattern Recognition Letters, 26(11), (2005), PP.1632-1640.

[12] Pawlak, Zdzisław, "Rough sets: Theoretical aspects of reasoning about data", 9, Springer Science & Business Media, (2012).

[13] Hassan Abu-Donia, and Amgad Salama, "Approximation operators by using finite family of reflexive relations", International Journal of

should be taken the attributes from the proposed model as essential attributes to building future systems.

References:

[1] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, Virgílio Almeida,"Characterizing user behavior in online social networks." Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, (2009), PP.49-62.

[2] Martinez Enrique Frias, Sherry Y. Chen, and Xiaohui Liu, "Survey of data mining approaches to user modeling for adaptive hypermedia." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, 36(6), (2006),PP.734-749.

[3] Ata Turk, R. Oguz Selvitopi, Hakan Ferhatosmanoglu, and Cevdet Aykanat, "Temporal workload-aware replicated partitioning for social networks", Knowledge and Data Engineering, IEEE Transactions on, 26(11), (2014), PP.2832-2845.

[4] Tiberiu Chis, Peter G. Harrison, "Modeling Multi-user Behaviour in Social Networks," Modelling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2014 IEEE 22nd International Symposium on , PP.168-173.

[5] Peter Leow, 2014 , "Self-organizing Map Demystified". http://www.peterleowblog.com/self-organizing-map-demystified/ at[15/8/2015]

[6] Daniel M. Romero; Brendan Meeder, Jon Kleinberg, "differences in the mechanics of information diffusion across

Applied Mathematical Research,4(2), (2015), PP.376-392.

[14] Leena H. Patil, and Atique Mohammad, "A multistage feature selection model for document classification using information gain and rough set", International Journal of Advanced Research in Artificial Intelligence (IJARAI), 3(11), (2014).

[15] YU, Bei; CHEN, Miao; KWOK, Linchi, "Toward predicting popularity of social marketing messages", In: Social Computing, Behavioral-Cultural Modeling and Prediction, Springer Berlin Heidelberg, (2011), PP. 317-324.

[16] G. S. Thakur & Neeraj Sahu \ D. S. Rajput, R. S. Thakur, "Analysis of Social Networking Sites Using K- Mean Clustering Algorithm", International Journal of Computer & Communication Technology (IJCCT) ISSN (ONLINE), 3(3), (2012), P. 90.

[17] Francis T. O'Donovan , Connie Fournelle, Steve GafFigurean, Oliver Brdiczka, Jianqiang Shen, Juan Liu, and Kendra E. Moore,"Characterizing user behavior and information propagation on a social multimedia network", In: Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on. IEEE, (2013), PP.1-6.

[18] Hongning Wang1 , ChengXiang Zhai1, Anlei Dong, Yi Chang ,"User modeling in search logs via a nonparametric bayesian approach", In: Proceedings of the 7th ACM international conference on Web search and data mining, ACM, (2014), PP. 203-212.

[19] Haiyan Wang,Feng Wang, and Kuai Xu. "Modeling information diffusion in online social networks with partial differential equations." arXiv preprint arXiv:1310.0505 (2013).

[20] Aaron Clauset, M. E. J. Newman, and Cristopher Moore, "Finding community structure in very large networks", Physical review E ,70(6), (2004), PP. 1-6.

[21] Chis T.; Harrison P. G., "Modeling Multi-user Behaviour in Social Networks, Modelling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS)", 2014 IEEE 22nd International Symposium on , l , (9-11,Sept., 2014), PP.168-173.

[22] Mihail Cocosila And y Igonor, "How important is the "social" in social networking? A perceived value empirical investigation", Information Technology & People, 28(2), (2015), PP. 366 – 382.