Graphical Approach for Social Network Mining

Priyanka Asrani
BE Computer
Department of Computer Engineering
VESIT, Mumbai

Shradha Randad
B.E Computer
Department of Computer Engineering
VESIT, Mumbai

Sneha Wadhwa
B.E Computer
Department of Computer Engineering
VESIT, Mumbai

Manisha Gahirwal
Professor at
Department of Computer Engineering
VESIT, Mumbai

ABSTRACT
Social Networking sites are a rage in today's world. Its importance has transcended the sole purpose of keeping in touch. In the quest of outweighing their contemporaries, social networking websites in today's world needs to incorporate new and interesting features to have an edge. Data Mining plays an important role in providing many such features. The social network structure consisting of numerous users can be best implemented using a graph data structure. It takes into consideration the data regarding various forms of interaction between two users to compute their association. The basic features such as friend suggestions, community suggestions, etc. have been incorporated. Also, commercial features such as targeted advertisement have been included.

General Terms
Graph based data mining.

Keywords
Interaction index, Interaction ratio, Association Edge Weight, Community point, Graph based data mining.

1. INTRODUCTION
When social networks were newly introduced their main purpose was just to be in touch with people you know. Over the years, social networking sites have successfully adapted to the changing trends in the society .It has evolved from being a basic means of socializing to a full-fledged system incorporating the commercialization aspect .The first social networking website 'The WELL'(Whole Earth 'Lectronic Link),launched in 1985 included basic features such as Internet forums, email services, etc. The most recent social networking websites, Facebook and Twitter provide various functionalities such as posts/tweets, messages, communities, groups, etc. Data mining is used to further enhance this experience in various ways tailored to individuals' interests. For example, in the absence of data mining, if a user wishes to join a community, then there exists an overhead of searching the community of his/her interest. With data mining, this additional task is avoided by providing him/her with suggestions to join communities which he/she will be most likely to join. Also, advertisements by communities are sent to the user by email. These features are incorporated based on the degree of association with other entities in the network. E-commerce websites can tie up with various commercial brands and use our model to obtain data and optimize their collective sales. They can do so by providing offers in accordance with the popularity of those brands, thus increasing their profit margins.

2. DESIGN OF PREDICTION MODULE

2.1 Objective of the Model:
User experience can be enhanced by encouraging social networking. Personalization can be used to achieve this.

2.2 System Module Diagram
The system module diagram shown below takes user interest, user groups and activity pages as inputs. These are fed to the prediction module. Likes, posts and comments on a user's profile are also used as inputs. These inputs are derived from the database. All these inputs are processed inside the prediction block. The processed output is used to provide friend and community suggestions. Further, this data can be processed to provide targeted advertisements and marketing.

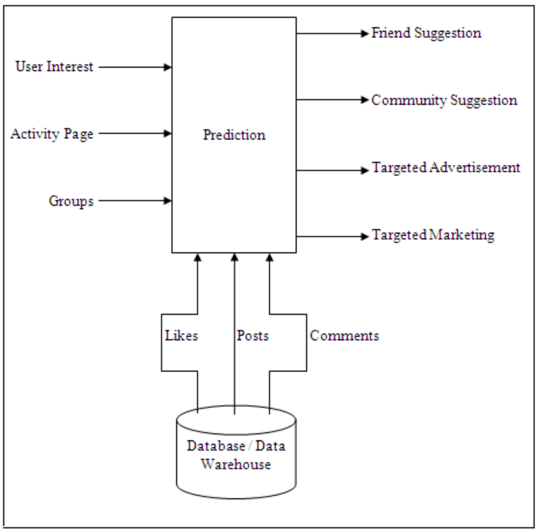

Figure 1: System Module Diagram

3. STRUCTURE OF PREDICTION MODEL

3.1 Selection of Prediction Model
The proposed system uses weighted bidirectional graph model for representing the data pertaining to user associations. The graph consists of a set of nodes representing system entities such as people and communities. The edges between the nodes represent the association between the participating entities. Further, the weights on the edges represent the degree of association. The graph is automatically altered depending on the changing participations among the users of the website. The Graph model uses the following parameters:

- Interaction Index
- Scaling Factor
- Degree of Association

3.2 Algorithm

Consider a social network comprising three users: X,Y and Z.

3.2.1 Interaction Index

The interaction index from user X to user Y, I(X Y)is defined as the sum total of posts, likes and comments made by user X on user Y's profile.

I(X,Y)=(∑Post + ∑Likes + ∑Comments)

Table 1. Table of Interaction Index

Users	Interaction Index		
	X	Y	Z
X	-	10	60
Y	40	-	20
Z	60	80	-

3.2.2 Interaction Ratio

The interaction ratio from user X to user Y, IR(X,Y) is defined as the ratio of Interaction index from user X to user Y to summation of interaction indices from user X to every other user. Interaction Ratio is required to determine the degree of connection from one person to other relative to his connection with other users.

IR(X,Y) = I(X,Y) / (I(X,Y) + I(X,Z))

Table 2. Table of Interaction Ration

Users	Interaction Ratio		
	X	Y	Z
X	-	0.143	0.857
Y	0.667	-	0.333
Z	0.429	0.571	-

3.2.3 Scaling Factor

Scaling factor (SF) is defined as the ratio of base index to the highest interaction ratio. This factor when multiplied with Interaction Ratio translates the ratio in range [0, Base Index].

SF= Base Index / (IR_{max})

Where IR_{max}= Highest Interaction Ratio

(In proposed model, Base Index is taken to be 10)

For the above example, IR_{max}= 0.857

Scaling Factor turns out to be 11.66

3.2.4 Association Edge Weight

Association Edge Weight (AEW) from user X to user Y is defined as the product of Interaction Ratio from user X to user Y and the Scaling Factor.

AEW(X,Y)= IR(X,Y) * SF

Table 3. Table of Association Edge Weight

Users	Association Edge Weight		
	X	Y	Z
X	-	1.67	10
Y	7.78	-	3.88
Z	5.0	6.66	-

Association Edge Weight is further rounded off to integral values to get Approximate Association Edge Weight(AAEW). This step is incorporated so as to account for cases where the difference in Association Edge Weight is less. So the above table gets converted to following table:

Table 4. Table of Approximate Association Edge Weight

Users	Approximate Association Edge Weight		
	X	Y	Z
X	-	2	10
Y	8	-	4
Z	5	7	-

The data in the above table is converted and represented using a graph.

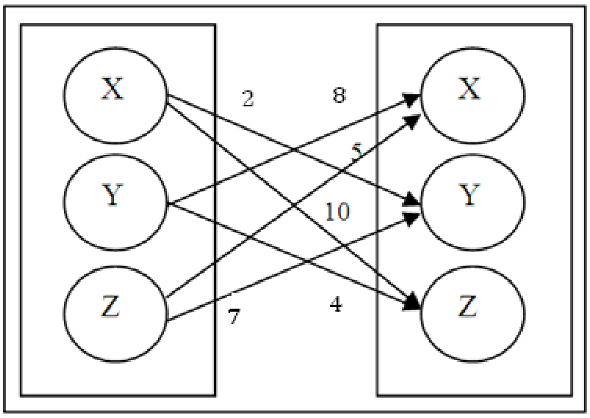

Figure 2: Association between Users

In the above graph, each node represents a user and an edge between two users denotes the degree of association between them.

4. GENERAL GRAPH STRUCTURE

In general a social network consists of users, communities, applications, etc. Various friends, community suggestions are based on activeness of user with his friends and also within communities. Consider the following graph, in which there are two communities C1 and C2. User Y and Z are part of community C1and C2 respectively. The individual association between various users is shown along the edges connecting the users. Also the participation level or the association level of a user in a community is shown along the edge between user and a community.

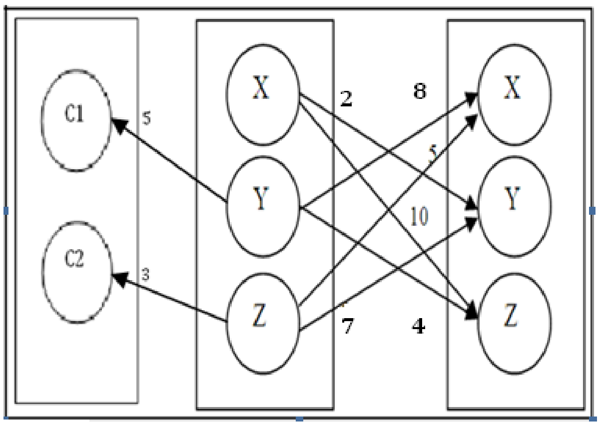

Figure 3: Association between users and communities

Also from the Figure 3,

Table 4. Table of Association Edge Weight

Users	Association Edge Weight	
	C1	C2
Y	5	-
Z	-	3

The logic behind community suggestion takes the associated edge weight into consideration between two users. In figure 3, user X has 2 friends Y and Z who are part of communities C1 and C2. To find which community will be suggested to user X, community point is calculated.

4.1 Community Point

Community Point (CP) is the product of associated edge weight (AEW) from User X and his friend and the assigned point for that associated edge weight. It is calculated for all friends of user X which are part of any community. The maximum value of all community points is selected and community associated to that is suggested to user.

CP(C1)= AEW(X, Y)* AAEW(X, Y)=1.67*2= 3.34

CP(C2)= AEW(X, Z)*AAEW(X, Z)=10*10= 100

CP(C2) is maximum. So C2 is the community which will be suggested to user X.

In case of a tie, the participation of user's friends in their communities is taken into consideration. In that case, the formula changes to

CP(C1)=AEW(Y, C1)* AAEW(X,Y)

CP(C2)=AEW(Z, C2)* AAEW(X,Z)

The maximum of community is point is the community that is suggested to user X. In this way, the community which a user is most likely going to join is calculated and is the one that is suggested to the user.

5. RESULTS

Most prevalent social networking sites today use a dilute version of the concept by taking into consideration only the fact of being associated, whereas some others consider the interests of an individual alone as the deciding factor. The proposed system on the other hand, considers not only the association between two individuals but also the degree of association between them. This largely improves the accuracy of the prediction as this metric is significantly important in social networks. Also the accuracy is bettered by further taking into account the common interests, activities, groups, etc. In order to test the performance of the proposed model, a comparison was made between the results of the proposed model and the results of the traditional model.The comparison was made based on the Accuracy and the Coverage resulted from each model.

- Accuracy is defined as the ratio of total number of successful predictions made by the system to the total number of predictions made.
- Coverage is defined as the ratio of count of successful predictions to the total advances actually made.

As an example, accuracy of providing friend suggestions is defined as the ratio of the count of accepted suggestions to the total number of suggestions made. The coverage of providing friend suggestions is defined as the ratio of the count accepted suggestions to the total number of friend requests sent.

5.1 Statistics

The proposed system was tested on our own department students comprising of 210 students. They formed a social network of their own via our website. It was found that out of a total number of 550 friends suggested to these 210 students, 498 were accepted and used for sending a friend request. Thus the accuracy of our system turned out to be:-

Accuracy=498/550=0.9054

Also, the total number of friend requests these 210 users sent over the week were 589. Out of these 589, 498 were those that were suggested by our system to these users. Hence coverage of our system turned out to be:

Coverage=498/589=0.8455

This data collected spanned an entire week. The accuracy and coverage of the proposed system were satisfactory. The accuracy of 0.9054 implies that 90.45% of the friend suggestions made by our system are successful. This accuracy ensures that a person never gets vague and undesirable friend suggestions that we often find on the social networking websites these days. The coverage of 0.8455 implies that 84.55% of the total friend suggestions made by the user are constituted by the suggestions made by our system. This coverage minimizes manual searching for friends on the website by a member. In simple words, a high coverage indicates that the suggestions have been successful in covering a larger part of a user's social network.

6. LIMITATIONS

- A prime limitation of this system as a whole is the randomness that comes with social networks. Social networking sites are used by the people to connect with their friends. No two people are same, so one single algorithm cannot cater to the entire crowd.
- The actions performed by a person may not always reflect his actual choices. A person may like a friend's post on the website but may not actually like it so much. This leads to inaccuracy. This inaccuracy increases with increase in the size of the database. Thus it depends on the social network of a person.
- A person may just remain inactive for a while, but may still be as good friends with an active person as before. The relative method of finding association makes it difficult to retain the same association level as before in such cases.

7. CONCLUSION AND FUTURE SCOPE

In the proposed model, the technique for providing friend and community suggestion is provided. The basis used to achieve this is the association level between two users. Commercialization and personalization can also be achieved through community advertisements that are provided to users having highest degree of association with a user having highest participation in a community through email. Association level concept can be further extended in future to increase the sales of commercial brands in tie up with various E-commerce websites. This can be achieved through our technique by putting various offers and discounts for the brands that have gained highest popularity as compared to others. Thus, this will help in boosting their sales and also to increase their brand value in the market.

8. REFERENCES

[1] Marjaneh Safaei, Merve Sahan, Mustafa Ilkan, "Social Graph Generation & Forecasting using Social Network Mining", in 2009 Annual IEEE International Computer Software and Applications

[2] Alvaro Ortigosa, José Ignacio Quiroga, Rosa M. Carro, "Inferring User Personality in Social Networks: A Case Study in Facebook", in 2011 11th International Conference on Intelligent Systems Design and Applications

[3] Juan J. Cameron, Carson Kai-Sang Leung, Syed K. Tanbeer , "Finding Strong Groups of Friends among Friends in Social Networks", in 2011 Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing

[4] Jerzy Surma, Anna Furmanek, "Data mining in on-line social network for marketing response analysis", in 2011 IEEE International Conference on Privacy, Security, Risk , and Trust, and IEEE International Conference on Social Computing

[5] Aleksandra Doniec, Albert Hupa, Radoslaw Nielek, "Web of Friends – Discovering a Social Network By Mining Data from Instant Messengers" , in 2009 International Workshop on Social Informatics

[6] Ruhaizan Ismail, Zalinda Othman, Azuraliza Abu Bakar, "Associative Prediction Model and Clustering for Product Forecast Data", in 2010 10th International Conference on Intelligent Systems Design and Applications

[7] Kyung Soo Cho, Jae Yoel Yoon, Iee Joon Kim, Ji Yeon Lim, Seung Kwan Kim, Ung-Mo Kim, "Mining Information of Anonymous User on a Social Network Service", in 2011 International Conference on Advances in Social Networks Analysis and Mining

[8] Slah Alsaleh, Richi Nayak, Yue Xu, "Finding and Matching Communities in Social Networks Using Data Mining", in 2011 International Conference on Advances in Social Networks Analysis and Mining

[9] P.M. Zadeh, M.S. Moshkenani, "Mining Social Network for Semantic Advertisement", Third 2008 International Conference on Convergence and Hybrid Information Technology, IEEE Press, Nov. 2008, Volume 1, pp. 11-13.

[10] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proc. ACM SIGMOD 2000, pp. 1–12.

[11] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," in Proc. ACM SIGMOD 1993, pp. 207–216.

[12] S. Wasserman and K. Faust, "Social Network Analysis: Methods and Applications", Cambridge University Press, 1994.

[13] S. Mitra, A. Bagchi, A.K.Bandyopadhyay, "Complex Queries on Web Graph representing a Social Network", 1st International Conference on Digital Information Management, IEEE Press, Dec. 2006

[14] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, "Mining Newsgroups Using Networks Arising from Social Behavior", Proc 12th Int'l Conf. on World Wide Web, 2003.