Mining in Social Network for Friend Suggestion List Providing: SNSs

Chit Hsu Wai, Khin Mar Myo
University of Computer Studies, Mawlamyine
chithsulay@gmail.com

Abstract

Nowadays Social Network analyses are receiving a lot of attention for researchers. By knowing the knowledge of social network, it can be applied in many application areas such as co-author finding, research group finding, friendship finding, terrorism finding, etc. All these tasks are based on frequent sub-graph patterns, graph classification, clustering and other techniques. This system intends to provide social network services(SNS) for online community social site. Firstly, framework for SNS is described. Then a prototype for social network is constructed to evaluate the efficiency of our framework. Secondly, methods and techniques used in framework are detail explained. Social network members with similar features are mined using ***k-means clustering algorithm*** *to provide find friend facility. Finally, suggestion friends list are provided by* ***mining frequent itemset (person) using vertical data format approach****.*

1. Introduction

Social networks are receiving a lot of attention on the webs because it allows users to post their personal information directly into online networked information space. The users of such websites form virtual or online community which becomes part of the modern society in many contexts such as social, educational, political, medical and business. Interest continue to grow in social networking services (SNS), notable due to the popularity of community web sites where user can develop their network by adding known individual to it. Perhaps because of this, research emphasis has focus on the development of communities online, where data mining can be done to analysis the network of memberships and relationships [1,4].

As a consequence, social network analysis (SNA) becomes an interested research areas because it can be applied in community mining, criminal group detecting, research and co-author group detection, etc. For business case, by detecting the customer and consumer groups from existing large network, it can be applied in viral marketing.

Generally the analysis of social networks involves graphs and complex networks modeling to study many complexes human and natural phenomena [2].

For *node boundary specification*, defining which nodes to include in a network poses an early challenge. [10] identified three approaches to addressing this boundary specification problem: *A position based approach, event based approach and relation based approach.*

For *link analysis*, relations between nodes must be identified. These could include collaborations, friendships, trade ties, web links, citations, resource flows, information flows, exchange of social support, or any other possible connection between nodes. [2] identified four broad categories of relations: *similarities, social relations, interactions, and flows.*

In this paper, a small prototype for social network is firstly constructed. Then dataset getting from social network concerning with members (nodes) are analyzed based on their personal information using partitioning clustering approach. These clustering results can be considered as groups with similar characteristics and events. Then individual member networks (subgraphs) are mined from large network using frequent itemset mining. These all results are used to provide as friend suggestion lists to the users.

The remainder of the paper is organized as follows. Some related works are discussed in the next Section 2. Section 3 describes the proposed system of friendship group finding system. In Section 4 System Implementation and performance evaluation is presented. Finally, the paper ends with the Conclusion in Section 5.

2. Related Works

Some of the SNA tasks are concerned with community mining. Most of these tasks are based on data mining techniques. The author [6], has proposed a method to find the research groups by research area and co-author from bibliography network using keyword clustering and author clustering approach. Another author [9], detected the community (sub-graph) from large graph using SOM clustering, k-means and efficient spectral algorithm). The author [10] has proposed a framework for automated network data analysis and deduction from multiple social networks by converting to transaction dataset, applying association mining, and statistical methods.

The author [7] tried to build a model of prediction for viral marketing. The main idea is that massive quantities of data on very large social networks are now available from blogs, newsgroups, chat rooms, knowledge-sharing sites, collaborative filtering systems, etc.

Another SNA tasks are link mining and link prediction. Mostly link mining and link analysis researches are based on mathematical model such as graph and set theories [4].

Other SNA tasks are based on semantic. The paper[8] tried to mine "social semantic association" between two networks by mining two sites and extant semantic with ontology approach. [5] explored the user profile elements as signals for making community. They used the user profile(datasets) getting from Facebook social network for analysis.

According to their research, network members are more prefer to connect with another member who have similar profile, for example similar location or similar University, etc. According to paper[2], a new network user is willing to make friends by viewing other persons' profile. If the two persons on the net have same University or same location or same interest and hobby, they are probable to become friends on the net. Therefore profile data such as address, Education status, etc are used as signals for user grouping in our system.

3. Concepts for Proposed System

This section describes some concepts for our system. These concepts are used in our system design.

3.1. Vertex and Edge in Social Network

In social network, vertexes (nodes) are the social network members and edges (links) are the relation between members as shown in Figure.

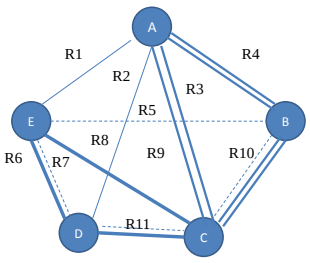


Figure 1. Members and Relations in Social Network

As shown in Figure 1, a social network can be considered as a collection of sub-networks. Here S is a large social network and consists of event-based sub-networks. By mathematical form,

S= {E1, E2, E3,……, En}

Events may be occurring when mouse click, message exchange, note left, gifts offer, photo share, etc. Each event may have set of member and relation.

E1={M1,R1}, M1={E,D,C}, R1={r6,r7,r11}

3.2. Mutual Friend in Social Network

Usually online community site provide social network services (SNS) such as friend suggestion lists, mutual friend lists providing. Mutual friend can be considered with the following scenario.

Let Kyaw, Maw, Zaw are the friend of Chit Su and Kyaw, Maw, Thaw are the friend of Swe Swe. Then Kyaw and Maw are the mutual friend of Chit Su and Swe Swe.

Chit Su ={Kyaw, Maw, Zaw}

Swe Swe ={Kyaw, Maw, Thaw}

The mutual friends of Chit Su and Swe Swe can be found by intersection of these two sets as follow.

Chit Su∩Swe Swe={Kyaw, Maw}

4. System Framework for SNS

The main objective of this system is to provide some of SNSs, the friend suggestion lists, to the user by taking the social network analysis using data mining techniques. There are two main components in system design as shown in Figure 2.

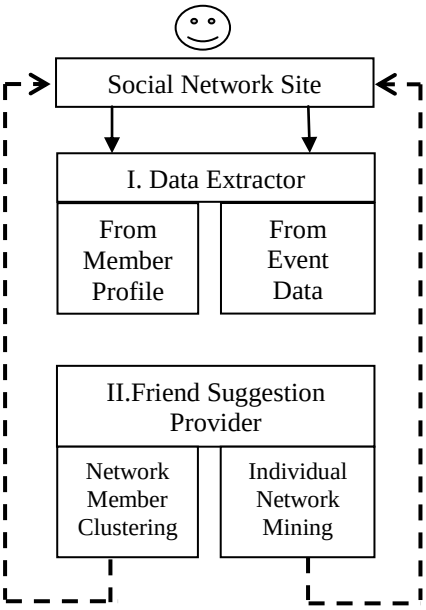


Figure 2. Framework for Social Network Service (SNS)

4.1. Data Extractor

This component responsible to extract and collect the required data from social network site. The required data may be profile of site members, event log data or user transaction data according to the types of social network analysis. The source of the data may have different forms of structure such that structure, semi-structure or unstructured database. These data must be transformed into relational data format for analysis.

4.2. Friend Suggestion Provider

A green user of social site may want to know who are the existing users and which persons have similar interest and similar characteristics with her. This component responsible to provide some of Social Network Services(SNSs) concerning with suggestions for new or green users of social sites.

To provide the above facility, this component must use one or more of the data mining and other techniques.

5. Implementation and Results

A prototype system for social network site is implemented using PHP and MySQL to evaluate the efficiency of proposed framework. Firstly social network design is explained. Then how to implement the framework and what are the results is described.

5.1. Social Network Site Design

The design of our social network site facilitate some of SNSs: (i) wall publish, (ii) photo uploading and (iii) user discussion. If a person wants to use social network site, he/she must firstly register to the site. After that he/she can make invitation to existing site members to become friends. By this way he/she can get networked friends.

Registration members can participate and access only in public events while networked members have participation rights to both public and private events. The following table shows the access right to the events types for site member.

Table 1. Member Types and Event Types

Types of Events	Types of Members
Public Wall	All registered members
Public Photo Upload	
Private Wall	Restrict to only networked members.
Private Photo Upload	

Members' transaction data for correspondent event are collected for analysis purpose.

5.2. Data Extracting

This component extract two forms of data from social site as shown in Table 2: Profile dataset and Table 3: Event transaction dataset to be analysis.

Table 2 . Profile Dataset

UID	Name	Age	Degree	City	Religious
1	Hein	18	Physics(2nd)	Bago	Buddhisms
2	Chit Oo	25	M.C.Sc	Lashio	Buddhisms
3	Aung Aung	29	B.Sc(Chem)	Dagon	Islam
4	Thar Thar	20	AGTI	Meiktila	Christianity
5	Komin	18	B.C.Sc	Myeik	Buddhisms
6	Mandalar	28	B.Com	Mandalay	Buddhisms
7	Sai Zeyar	24	B.Sc(Phy)	San Chaung	Christianity
8	Swan Htet	24	ME	Meiktila	Buddhisms
9	Win Zan Lin	23	B.PA	Botataung	Buddhisms
10	Kyaw Soe	29	Myanmar(second)	Dagon	Islam

Some attributes such as sex, interest, hobby, degrees obtained, degree getting date, occupation, etc are omitted to be described in Table-2 for saving space.

Table 3. Event Transaction Dataset

TID	UID(from)	UID(To)
001	1	**3**,5,
002	2	**1**,
003	5	**2,3,4**
004	6	**1**,
005	8	**2** , **1**, 5,
006	9	**1**,**6**,8
007	10	**1**,**2**,**7**,**3**,5,**6**,11,
008	12	**1**,**2**,10

5.3. Friend Suggestion Providing

The extracted data getting from component one is used by these second components. The main objective of this component is to provide one of

SNSs, friend suggestion list providing, to users. *K-means clustering* method is used to analysis the user profile(Table 1) and *frequent item set mining(vertical approach)* is used to analysis the event transaction data(Table 2).

5.3.1. Members Clustering

In this step, the well known k-means algorithm is used to cluster the site members based on their characteristics such as age, degree, city, etc.

The *k*-means algorithm starts with *k* random mean vectors and then, in turns, assigns each instance to the cluster with nearest mean vector and re-calculates the means by averaging over the assigned instances as long as there is a change in the cluster assignments. [The steps of k-means algorithm is describes in Appendix 1]

Euclidean distances measured are used to determine the similarity of the user characters.

Euclidean distance $\|\boldsymbol{x}-\boldsymbol{y}\|$ where x= $(a_1, a_2, \dots, a_n)$ and y= $(b_1, b_2, \dots, b_n)$ is given by the equation:

Euclidean distance=$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_n-a_n)^2}$

The cluster output may be differing based on number or k value or number of attributes contained in clustering.

If the k value is 2, the table 2 becomes two clusters as shown in Appendix 2. The cluster output groups are used to provide as friend suggestion list to user as shown in Appendix 3.

5.3.2. Mutual Friend finding

The event transaction dataset are used to find the mutual friend by using frequent item set mining algorithm (vertical approach). Firstly event dataset (Table3) is converted into vertical format as in Table 4.

Table 4. Frequent Itemset

UID	Transaction
1	2,6,8,9,10,12,
2	5,8,10,12,
3	1,5,10,
4	5,
6	9,10,
7	10,

After finding 2 frequent item set, Table 4 become Table 5.

Table 5.Two Frequent Itemset

UID	Transaction
1,2	8,10,12,
1,3	10,
1,4	-
1,5	-
1,6	9,10,
1,7	10,
2,3	5,10
2,4	5,
2,5	-

According to Table 4, UID-8,10 and 12 are mutual friend of member UID-1 and 2. Beside these, we can say that Table 4 represent the individual sub friend network in large network. UID-2,6,8,9,10 and 12 are the network of UID-1. UID-5,8,10 and 12 are the network of UID-2. The interface for individual network from admin mode is described for Appendix 4.

The results getting from this second phase is used to facilitate the users as SNSs.

5.4. Performance Measure

The quality of the clustering can be evaluated by means of internal cluster and external cluster quality measuring techniques.
For this system, F-measure is used to evaluate the external cluster quality.

Recall (i, j) $=n_{ij}/n_i$

Precision (i, j) $=n_{ij}/n_j$

Where n_{ij} is the number of members of class i in cluster j, n_j is the number of cluster j and n_i is the number of members of class i.

F measure

F(i, j)$=\frac{(2*Recall\ (i,j)*Precision(i,j))}{((Precision\ (i,j)+Recall(i,j)))}$

F(i, j)$=\sum_i \frac{n_i}{n} max\{F(i,j)\}$

Where the max is taken over all clusters at all levels, and n is the number of members.
The accuracy results for different k-values are illustrated in figure 3.

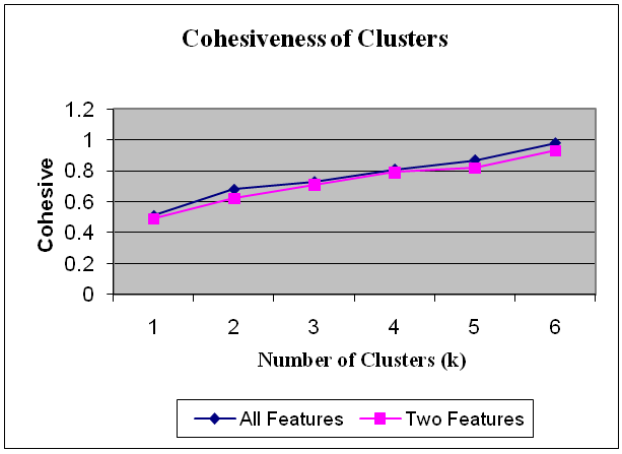


Figure -3. F-measure for different k-values

According to the above diagram, the accuracy is increased when the k-value is increased.

6. Conclusion

According to our experiment, it is obvious that the framework worked well for SNSs. Although there may have many services for social network (SNSs), only suggestion friend lists are provided by our system. K-means clustering algorithm is useful for existing social network site members clustering and grouping. The clustering results are used for suggestion to similar user. Our experiment of mutual friend finding using frequent-item set mining contains many limitations. Only small amount of transaction data set is applied for this. It will be difficult to mine on large numbers of network members to find mutual friend. Others alternative approaches or models are required.

REFERENCE

[1]. L.A Adamic, O. Buyukkokten, and E. Adar.: A Social Network Caught on the Web. First Monday.

[2]. Admin Milani Fard, Martin Ester "Collaborative Mining Multiple Social Networks Data for Criminal Group Discovery" school of Computer science Simon Fraser University of BC, Canada V5A 1S5 {milanifard ester}@cs.sfu.ca.

[3]. Anil K. Jain, Department of Computer Science & Engineering," Data Clustering**:** 50 Years Beyond K-Means", Michigan State University, East Lansing, Michigan 48824 USA.

[4]. Sören Preibusch-Alastair R.Beresford Privacy-Preserving Friendship Relations for Mobile Social Networking W3C Workshop on the Future of Social Networking-Position Paper University of Cambridge. Computer Laboratory William Gates Building, 15 JJ Thomson Avenue Cambridge CB3 0FD, United Kingdom. Firstname.Lastname@cl.cam.ac.uk

[5]. Cliff Lampe, Nicole Ellison, and Charles Steinfield, "Familiar Face (book): Profile Elements as Signals in an Online Social Network" Dept. of Telecommunication, Information Studies, and Media Michigan State University, East Lansing, MI

[6]. B. Wellman, J. Boase and W. Chen.: The Network Nature of Community Online and Office. IT and Society 1(1), Summer.pp.151-165.2002.

[7]. Pedro Domingos "Mining Social Networks for Viral Marketing" Department of Computer Science and Engineering, University of Washington.

[8]. Ivan Cantador and Pablo Castells, Multilayered Semantic Social Network Modeling by Ontology-Based User Profiles Clustering: Application to Collaborative Filtering Escuela Politecnica Superior, Universidad Autonoma de Madrid Campus de Cantoblanco, 28049 Madrid, Spain.

[9]. M. E. J. Newman, Detecting Community structure in networks , Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109-1120

[10]. Jianhua Ruan and Weixion Zhang, An Efficient Spectral Algorithm for Network Community Discovery and Its Applications to Biological and Social Networks, Department of Computer Science and Engineering Washington University in St Loius One Brookings Dr, St Louis, MO 63130.

Appendix 1. Steps of k-means algorithm

Input: The number of clusters k and a database containing n objects.
Output: A set of k clusters that minimizes the square-error criterion;
Method:

1. Arbitrary choose k objects as the initial clusters;
2. Repeat
3. (Re) assign each objects to the cluster to object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, i.e., calculates the mean value of the objects of the each cluster;
5. until no change;

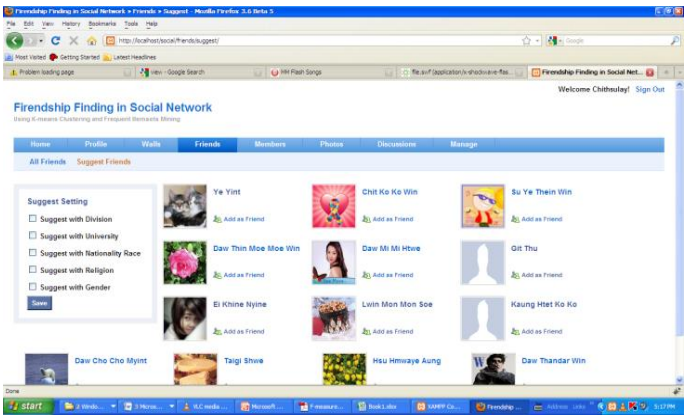
Appendix 2(a). Example table for cluster 1

UID	Name	Age	Degree	City	Religious
1	Hein	18	Physics(2nd)	Bago	Buddhisms
2	Chit Oo	25	M.C.Sc	Lashio	Buddhisms
3	Aung Aung	29	B.Sc(Chem)	Dagon	Islam
4	Thar Thar	20	AGTI	Meiktila	Christianity
7	Sai Zeyar	24	B.Sc(Phy)	San Chaung	Christianity
9	Win Zan Lin	23	B.PA	Botataung	Buddhisms

Appendix 2(b). Example table for cluster 2

UID	Name	Age	Degree	City	Religious
5	Komin	18	B.C.Sc	Myeik	Buddhisms
6	Mandalar	28	B.Com	Mandalay	Buddhisms
8	Swan Htet	24	ME	Meiktila	Buddhisms
10	Kyaw Soe	29	Myanmar(second)	Dagon	Islam

Appendix 3. Friend suggestion for new user



Appendix 4. Individual network from admin mode