Study of Data Mining Algorithm in Social Network Analysis

Chang Zhang[1,a], Yanfeng Jin[1,b], Wei Jin[1,c], Yu Liu[1,d]

[1]Shi Jiazhuang Post & Telecommunication Technical College，Shijiazhuang，050021, China

[a]airmagicz@163.com, [b]yanfengjin@163.com

Keywords: Data Mining, Social Network, Logic Programming, Social Structure

Abstract. Some mining methods and algorithms will be introduced to describe the social network. Method based on similarity measurement and inductive logic programming are useful here to analyze social networks, moreover, some specific datasets are used to analyze the characters of social networks by Ucinet which is very powerful data mining software. The Zachary club social network's property can be found from analysis.

Introduction

Traditional data mining uses a "property-value" table to represent data. Data is shown as vector; each dimension of vector corresponds to a value of conditional property. Social network data is a structured relational data, except the properties of each node, the more important thing is the link between nodes. The links contain a lot of information. However, the vector form shows the independence of nodes, ignore the link must not be good to knowledge discovery. Therefore, to analyze the social network data, the relational data mining methods should be used.

Social network analyze is a major application of relational data, the development of relational data mining provides a more effective tool for social network analysis, which promotes the development of social network analysis. The society network analysis pays attention to the link, which is a very important character. From the aspect of data mining, the society network analysis is also called link mining [2]. Through the mining of links, richer and more accurate information of instances can be got. At the same time, the link itself is often of concern to us as information. For example, in certain circumstances, not all links are observed, thus we possibly interested in predicting the existence of link between instances. When a social network model is established, some very useful information can be analyzed from it.

Method Based on Inductive Logic Programming

Many algorithms of relational data mining are all come from the ILP [3] (inductive logic programming). In order to study useful models from relational data, ILP method primary uses logic programming language, ILP is a crossover field of machine learning and logic programming, which mainly concern how to find new knowledge from data. Social network analyze is a major application of relational data, the development of relational data mining provides a more effective tool for social network analysis, which promotes the development of social network analysis.

ILP is an important method in learning relationship, which construct first order logic statements inductively from samples and background knowledge. Because it adopts logic as a representation, ILP overcomes two difficulties in traditional machine learning: the limited expression of propositional logic; background knowledge can not be added in the learning process. Furthermore, the result of ILP learning is easy to understand. ILP research focuses on the induction of

relationship rule before. In recent years, ILP research has been expanded to cover almost all of the learning tasks, such as classification, regression, clustering and correlation analysis.

Kings [4] uses the ILP method to classify the graph. First, convert graph representation to relation representation, use predicates below to describe graph:

$$vertex(graphID, VertexID, VertexLabel, VertexAttributes) \quad (1)$$

$$edge(graphID, VertexID1, VertexID2, BondLabel) \quad (2)$$

Then, ILP system can be used to find a good assumption in this space. ILP has been used in mining subgroup and graph classification.

Method Based On Similarity Measurement

Many data mining methods based on similarity measure. The definition of similarity associated with tasks, the best definition of similarity is likely to be different when the same data set under different tasks. Sometimes it is difficult to choose an appropriate similarity measurement, especially when there are lot of attributes whose relationships are not clear with aim and task. However, if given the appropriate similarity measure, such algorithm has a good intuitive explanation.

Similarity measure [5] is very useful in the link prediction. Link prediction is to determine whether there is a link between two actors. In the social network G, the similarity measure function for each pair of nodes: <x, y>, is given a possibility of a link:score(x, y) In some applications, the function can be seen as the topology structure of network G, for each node x and y calculated the degree of similarity between them. However, in some social network analysis tasks, the weight is not calculated the degree of similarity between nodes, but in order to do appropriate changes for specific target. Some of these weights are based on node neighborhood of node; others are based on the ensemble of all paths.

Now moving on the weight based on the neighborhood, if two authors' colleagues have a large intersection, the possibility of their future cooperation will be greater than the same two who do not have same colleagues. Two people who have overlapping social circles have more probability to become friends. Starting from this intuitive observation, the probability that node x and node y contact each other in the future is related to their neighborhood nodes. Γ(x) can be used to denote neighborhood of x in graph G. Some methods measure the intersection level of Γ(x) and Γ(y) as the probability of two node intersection.

Common neighbours [6] use the number of neighbour's intersection as a measurement of intersection degree, which is a very straight idea. It defines:

$$score(x, y) := |\Gamma(x) \bigcap \Gamma(y)| \quad (3)$$

Jaccard coefficient refers a similarity measure which is often used in information retrieval:

$$score := \left| \frac{\Gamma(x) \bigcap \Gamma(y)}{\Gamma(x) \bigcup \Gamma(y)} \right| \quad (4)$$

These two methods are simple counts, which treats all the neighbours equally, but Adamic/Adar method takes neighbours property into consideration, which weighted the neighbours:

$$score(x, y) := \sum_{z \in \Gamma(x) \bigcap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \quad (5)$$

In the sort of object, according to the node characters itself, as well as local or global structure of

the networks, the importance of nodes can be judged. For example, node degree can simply be used as the importance standards of local standards. Overall standards can use approach of eigenvectors to describe node importance which is related to the important nodes they linked.

Specific Interpretation of Social Network Analysis

Social networks analysis is a set of norms and methods to analyse social network structure and its properties. It is also called as structural analysis. Because it mainly analyse structures and attributes of social relation which is constituted by different social unit, such as individuals, groups, organizations and so on. Hence, social network is not only a set of technologies to analyse the structure and relation, but also a theoretical approach, which is called as structural analysis idea. Network analysis is to explore the deep structure which is a certain network mode hidden under the complex social system surface.

Basic Principles of Social Network Analysis: As a basic approach to research social structure, there are several principles of social network analysis below.

First, relationship ties are often interactive and asymmetric; it is different in content and intensity. Second, relations link directly or indirectly connected to the network members; so we must analyse it under a larger the network structure context. Third, social ties' structure produces non-random networks, which generates network clusters, network boundaries and cross-correlation. Forth, cross-correlation links network groups and individuals together. Fifth, asymmetrical ties and complex networks make the distribution of scarce resources unequal. Sixth, network produces the action of cooperation and competition for the purpose of scarce resources.

This structural methodology [7] means: Social science research should be targeting the social structure, rather than the individual. Through research networks relationships, we can associate relationship between the individual, "micro" social networks and large-scale social systems of "macro" structure together. Therefore, the British scholar J.Scott said: "Social network analysis has laid the foundation of new theories emergence of social structure."

While the researchers emphasis on those methodology of social structure, but they use the concept of structure is also much different. In fact, in sociology, social structure is used in different levels. It can be used not only to illustrate the micro social interaction relation model; but also explain the macro-social relations model. In other words, from the social role to the all society, there are structural relationships.

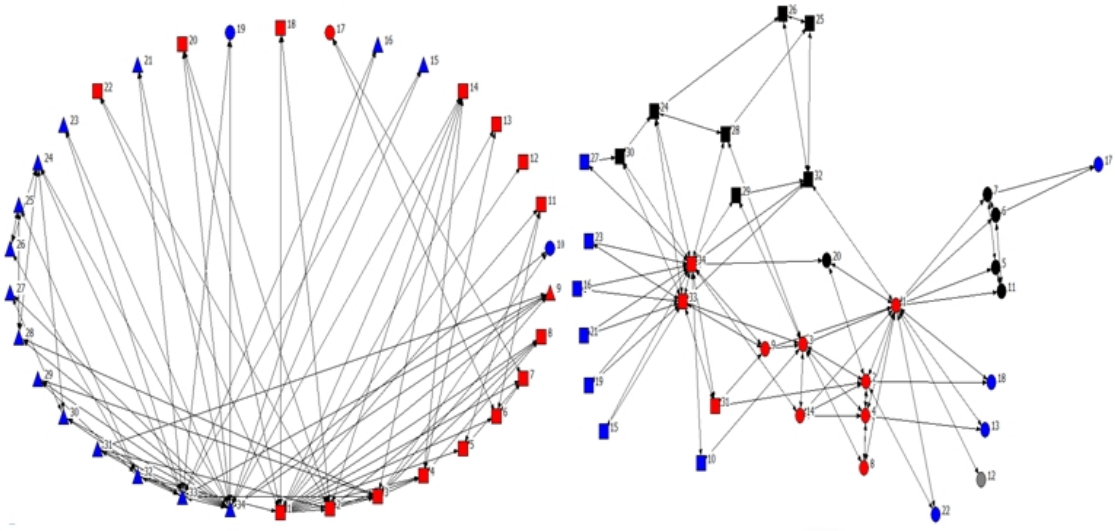
The Levels of Social Structure: In general, sociologists use the concept of social structure in the following levels: First, the level structure of the social role (micro structure): the most basic social relation is role relation. Role is often not single, isolated, but the form of role bundle. It reflects people's social position or status relationships, such as teacher - student. Second, the level structure of organization or group (middle structure): refers to the relationship between the constituent elements of society, this structure is not reflected in the individual relationship between activities. Such as occupational structure, it reflects the social professional status, the relationship with the resources and so on. Third, the level structure of the social system (macro structure): refers to the social macro structure as a whole. For example, the class structure of society, it embodies the relationship between the major interest groups, or the characters of social system.

Therefore, the social structure has multiple meanings. However, from the new concept of structure, the social structure is the general form of social existence, rather than specific content. Therefore, many sociologists have advocated the study object of sociology should be the social relations, rather than specific social individual. Because as individuals people are different and changing, but only the relationship is relatively stable.

Data Analysis

Here software NetDraw [8] can be used to draw graphs, which is very useful to render a directed graph of the social network data. The NetDraw supports a VNA data format which is just ordinary text file format. They have three start sections which are node data, node properties, and tie data. The three sections consist of the structure of VNA file. The first section "Node data" contains variables that describe the actors in a network, which means it contains some actors' features which can be used to divide actors into different subgroups. The second section "Node properties" contains some of the actors' attributes which can be shown on the graph. The third section "Tie data" contains the relations between actors. It may include the link relation and strength relation.

"Zachary" data file [9] can be used in data analysis, the data shows relationships among people in Zachary club. Different data sampling are extracted in this example. The ucinet can show a circle graph by using samplings. The nodes here are located at equal distances around a circle, and nodes that are highly connected are very easy to quickly locate because of the density of lines. If transforming from attribute "ID", the result is shown as left part of graph 1. K-cores which is a definition of group or subgroup to show the graph's subgroup can be used too, nodes in same group have same color. A K-core is a maximal group of actors, all of whom are connected to some number (k) of other members of the group. In this example there are four different subgroups as right part of graph 1.



Graph 1: Circle and K-cores

Using in-degree statistics analyse ZACHE data by ucinet, shown as table 1, an actor's in-degree is the sum of connection from other actors to this actor, which shows how many actors send information to the specific one. The in-degree is very meaningful, because actors that receive much information from sources seem to be powerful and prestigious. But sometimes it maybe means information overload or noise interference.

Descriptive Statistics

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Mean	0.485	0.273	0.303	0.182	0.091	0.121	0.121	0.121	0.152	0.061	0.091	0.030	0.061	0.152	0.061	0.061
2 Std Dev	0.500	0.445	0.460	0.386	0.287	0.326	0.326	0.326	0.359	0.239	0.287	0.171	0.239	0.359	0.239	0.239
3 Sum	16.000	9.000	10.000	6.000	3.000	4.000	4.000	4.000	5.000	2.000	3.000	1.000	2.000	5.000	2.000	2.000
4 Variance	0.250	0.198	0.211	0.149	0.083	0.107	0.107	0.107	0.129	0.057	0.083	0.029	0.057	0.129	0.057	0.057
5 SSQ	16.000	9.000	10.000	6.000	3.000	4.000	4.000	4.000	5.000	2.000	3.000	1.000	2.000	5.000	2.000	2.000
6 MCSSQ	8.242	6.545	6.970	4.909	2.727	3.515	3.515	3.515	4.242	1.879	2.727	0.970	1.879	4.242	1.879	1.879
7 Euc Norm	4.000	3.000	3.162	2.449	1.732	2.000	2.000	2.000	2.236	1.414	1.732	1.000	1.414	2.236	1.414	1.414
8 Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9 Maximum	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
10 N of Obs	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000	33.000

Table 1: In-degree

Use the similar way, there are many tables and graphs can be made by using ucinet analyzing data. Many quotas which help people to analyze social network can be found too. Link proportion shows proportion of each actor link to another, adjacency shows which two actors have one actor to share, out-degree shows actor sends information to others, reachability shows there are connections where can be traced from the source to the specific actor, the distance is important for us to understand actors' different in the constraint and the probability that they have their position, maximum flow shows the number of paths between two actors, group shows how the network as a whole is likely to behave, tree diagram shows how many actors are overlapped in different cliques.

Conclusions

Application of social network analysis is a very strong sociological research approach; it can be used to describe the nature of the community, and can solve many practical problems. Social network data's collection, both to the field with an unprecedented opportunity for data analysis also presented enormous challenges. Data mining as a tool help people to discover useful knowledge from a lot of data. Through constant development, it has been able to deal with the structure of social networks as network data. Different from the traditional tasks of data mining, which assume the instances are independent; instances in social network are dependent. Such dependence can be described as links. Mining from links can provide us more accurate and richer information about the social network.

Social network adopts graph as a representation, the data have structural characters, it also emphasize the links between the actors are more valuable than the actors themselves. Therefore, we must use relational data mining methods, give full consideration to the important role of contact, and then we can complete the analysis tasks well.

Social graph's representation and some methods of mining the social network data are introduced here. Some algorithms of network analysis are given to enhance understanding for the data mining of social networks. If analyzing a specific dataset of social network, the software ucinet could be a very useful tool, which can be used to mining the social structures and find some useful information.

Acknowledgment

This work is supported by the research of science and technology research project of education department of Hebei Province, the contents and conclusions are based on deep study of sub-topic of the research project. The project No. is Z2014167.

References

[1] P. Domingos, P. Mika and J. Golbeck, Social networks applied, IEEE Intelligent Systems, 20 (1), 80-93, 2005

[2] Ulrike Gretzel. Social network analysis: Introduction and resource [EB/OL]. http://lrs.ed.uiuc.edu/tse-portal/analysis/socialnetwork-analysis/,2001.

[3] Freeman L C. The development of social network analysis[M].Vancouver: Empirical Press, 2004.

[4] Whittaker S, Jones Q, Terveen L. Contact management: Identifying contacts to support long term communication[C]. Proceedingsof Conference on Computer Supported CooperativeWork.New York: ACM Press, 2002:216-225.

[5] L. Freeman, Centrality in social networks: Conceptual clarifications, Social Networks, 10 (1), 215-239, 1979.
[6] Carley K.“Dynamic Network Analysis”in the summary of the NRC workshop on social network modeling and analysis [C]. Ron Breiger, Kathleen M Carley. National Research Council, 2003.
[7] Goh K I, Oh E, Kahng B, et al. Betweenness centrality correlation in social networks [J]. Physical Review E, 2003,67 (1-2): 017101-1-017101-4.
[8] Girvan M, Newman M. Community structure in social and biological networks [C]. USA: Proc Natl Acad Sci, 2002:8271-8276.
[9] Brandes U. A faster algorithm for betweenness centrality [J]. Journal of Mathematical Sociology, 2001,25(2):163-177.