Enhancement of Existing Community Detection Method using Similarity Based Link Prediction

Pallavi Gupta

Computer Science & Engineering, M.I.T.S. Gwalior, MP, India
Department of CSE & IT M.I.T.S. Gwalior, MP, India
gptpallavi@gmail.com

Sanjiv Sharma

Computer Science & Engineering, M.I.T.S. Gwalior, MP, India
Department of CSE & IT M.I.T.S. Gwalior, MP, India
dr.s.sanjiv@gmail.com

Abstract- Link prediction and community detection are two major area of link mining which separately studies over decades. Link prediction is used to predict the possible future links and hidden links in complex network while community detection deals with topological structure of the network to find new communities and change in the network structure over time. Community detection refers to the problem of finding such groups in real world social network. Link prediction and community detection both are based on relations between vertices in the network. Related work shows that the use of link prediction is beneficial for detecting new communities in complex network that has given very less attention. This paper is based on a different approach to community detection through link prediction that uses the Girvan Newman method.

Keywords: Social Network, Link Mining, Link Prediction, Community Detection.

1. Introduction

Data mining has established itself as the demanding field for studying social networks by applying one of its feasible techniques named Social Network Analysis. Social Network Analysis usually analyse the various social networks present on web that allows us to understand the structure and behaviour of these networks. In real life, individuals are not independent. They are mutually connected and affected by each other. Social network analysis focuses on individuals and ignoring their relationships with other individuals affect the accuracy and comprehensiveness of analysis. Thus social network corresponds to link between social entities such as person or social group. From a data mining perspective, it is also called link analysis or link mining that focuses on exploring and explaining the hidden patterns in the network and effect of these relationships that is based on this assumption that social entities are interdependent. A social network contains set of objects and relationships that can represented by graph $G_{V,E}$ in which V is set of nodes that represent entities and E is edge set that represent relationship between these entities. Social network describes interaction or collaboration between entities that are highly dynamic for developing relationships among individuals, they grow and change quickly overtime through addition of new link. Link prediction is a very important problem that is an aspect of social network analysis. Two comprehensive area on which social network mainly focuses are link prediction between entities and community detection in the network.

Link prediction problem is a very significant task of social network analysis. It is based on observed existing link information and attributes information and it predicts the correlation or links between two nodes. Link prediction is not limited to social network it has

DOI- 10.18486/ijcsnt.2015.4.3.02
ISSN-2053-6283

wide range of scenario. As in electronic commerce, it can be used to create recommendation system [1]; in the field of bioinformatics, it can be used to predict interaction between proteins [2] and it can be also used to find hidden terrorist network and criminals gangs [3] in the field of security. Link prediction is based on similarity between two entities that used correlation algorithms to solve the problems. For this purpose a similarity function is defined as $Similarity\,(x, y)$ between the nodes x and y. Greater the value of this similarity function is greater the possibility of link between nodes.

The community structure captures the tendency of nodes in the network to group together with other similar nodes into communities. This property has been observed in many real-world networks. Despite excessive studies of the community structure of networks, there is no consensus on a single quantitative definition for the concept of community and different studies have used different definitions. A community, also known as a cluster, is usually thought of as a group of nodes that have many connections to each other and few connections to the rest of the network. Identifying communities in a network can provide valuable information about the structural properties of the network, the interactions among nodes in the communities, and the role of the nodes in each community.

II. Background and Related Work

Social network [4] consists of a group of people and connections between them and this social link makes a relationship between two people. It is popular way to model the interactions among the people in group or community. Social networks are highly dynamic in nature they grow and change with time and can visualize as graph.

M. Girvan and M.E.J. Newman [5] proposed the property of community structure, in which networks nodes are joined together in tightly knit groups, between which there are only looser connections. These authors propose a method for detecting such communities, built around the idea of using centrality indices to find community boundaries. Further they extend their work and divisive an algorithm which uses between ness score to find the inter community edges [6].There algorithms share two ultimate features: first, they involve iterative removal of edges from the network to split it into communities, the edges removed being identified using one of a number of possible "betweenness" measures, and second, these measures are, crucially, recalculated after each removal.

Bowen Yan and Steve Gregory [7] proposed a novel method for improving existing community detection algorithms by using a simple vertex similarity measure (common neighbours) to calculate the scores for the existing edges in a network. Fenhua Li et al. [8] proposed a method for improving existing link prediction algorithm by using clustering information. Clustering information is used to improve the accuracy of link prediction.

The link prediction problem was first proposed by Liben-Nowell and Kleinberg [9] who used a model for link prediction based on node similarity. There are several categories of node similarity. First one is the neighbourhood based similarity like common neighbours of two nodes and the other one similarity based on path which tries to determine shortest path distance between two nodes. Tylenda et al. [10] proposed methods to incorporate temporal information available on evolving social networks for link prediction. In this paper, results unequivocally show that timestamps of past interactions significantly improve the prediction accuracy of new and recurrent links over rather sophisticated methods proposed recently. Liyan Dong et al. [11] proposed a new algorithm for link prediction on social network in which they improved the algorithm of common neighbour based on local similarity and Katz algorithm based on global similarity and provides novel approach based on nodes multiple attributes information under some guidance force. Linyuan Lu and Tao Zhou [12] investigate the behaviour of local similarities index in case of weighted and unweighted scenarios. There experimental studies show that weak ties play significant role in link prediction problem they remarkably enhance the predicting accuracy.

Karate club is constructed by Wayne Zachary [13]. It is a network of friendships between members of the club using a variety of measures to estimate the strength of ties between members.

DOI- 10.18486/ijcsnt.2015.4.3.02
ISSN-2053-6283

III. Basic Preliminaries

There are two basic purpose of social network analysis i.e. link prediction and community detection both uses some basic indices for the evaluation and analysis of complex networks. Link prediction uses indices based on the local and global similarities of nodes to predict the new links between them and finding known links that are already present in the network. Community detection is used for finding groups in the network with similar behaviours. This section introduces some indices that are used for experiment and analysis purpose.

3.1 Local Similarity Indices

Link prediction uses the topological information for the measure of similarity based on local information between the nodes. Here this paper introduces some local similarity indices for link prediction.

3.1.1 Common Neighbours

Common neighbour is a node neighbourhood based technique. The size of common neighbourhood of two nodes x and y can be defined as

$$S_{xy}^{CN} = |\Gamma x \cap \Gamma y| \dots \dots \dots (1)$$

Equation (1) represent the number of neighbours that x and y have in common. This technique is based on the intuition that if there is a node that is connected to x as well as y, then there is high probability that vertex x be connected to vertex y. Thus, as the number of common neighbours grow higher, the probability that x and y have link between them increases. In other words two nodes x and yare more likely to have a link if they have many common neighbours. Kossinets and Watts [14] works to analyse a large-scale social network like Facebook. In their work, they found that two students having many mutual friends are very probable to be friend in the future.

3.1.2 Jaccard's Coefficient

Paul Jaccard introduces Jaccard coefficient [15] which determines the association between two words. Jaccard index is a name often used for comparing distance, similarity and dissimilarity of the data set. It presents a normalized form of common neighbour. To measure the Jaccard similarity coefficient between two data sets is

$$S_{xy}^{JC} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \dots \dots \dots (2)$$

It defines the probability that a common neighbour of a pair of vertices would be selected if the selection is made randomly form the union of the neighbour sets.

3.1.3 Adamic Adar

This technique [16] was firstly proposed for the metric of similarity between two web pages. It calculates the probability when two personal homepages are strongly related. It computes features that are shared among nodes and then defines the similarity between them. In this similarity index, small degree node contains a higher similarity function value. It is formulated as

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \dots \dots \dots (3)$$

Finally this equation shows that the common neighbours of smaller degree more heavily.

3.1.4 Resource Allocation

In the pair of node x and y that have no direct link the resource allocation from x to y is done through their common neighbour. This similarity metric is influenced by the idea that complex network resources are dynamically allocated. The common neighbours plays role of parsers that have unit of resources in simplest case thus the similarity of node x and node y can distinct as the number of resources that node x receives from node y. This similarity measure is formulated as

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{d(z)} \dots \dots \dots \dots (4)$$

Intuitively this is similar to the common neighbour but node with higher degree are not much important than the node with low degree because it assume that a high degree node connected to node x and node y are less meaningful than low degree vertices.

3.1.5 Preferential Attachment

In generation of scale-free network evolution model preferential attachment [17] mechanism is used. It describes that probability of new link of node x is directly proportional to the degree of node. Therefore the probability of the link between node x and node y is directly proportional to multiple of their node degree.

This similarity metric is not required neighbours node information hence having low computational complexity. It is defined as

$$S_{xy}^{PA} = d_x \times d_y \dots\dots\dots (5)$$

3.2 Global Similarity Indices for Link Prediction

Compared to local similarity indices, global similarity indices requires more network topological information hence calculation is very time consuming and when network is large the calculation program of global similarity does not work.

3.2.1 Katz Metric

Katz [18] describes the similarity using global path. Katz metric is based on the ensemble of all paths which directly sums over the collection of paths and is exponentially damped by length to give the shorter paths more weights. The mathematical expression is defined as

$$S_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l . |paths_{xy}^{<l>}| \dots\dots\dots (6)$$

Where $paths_{xy}^{<l>}$ is the set of all paths with length l connecting x and y and β is a free parameter that has value between 0 and 1 controlling the path weights. A very small β yields a measurement close to common neighbour, because the long paths contribute very little.

3.3 Community Detection Methods

This section introduces a community detection method that is useful for identifying communities in Zachary's karate club dataset. This method is mostly based on graph clustering and partitioning, hence many community detection methods partition the network into disjoint communities. In recent development, researchers have begun to develop community detection methods that identify overlapping communities.

3.3.1 Girvan Newman Method

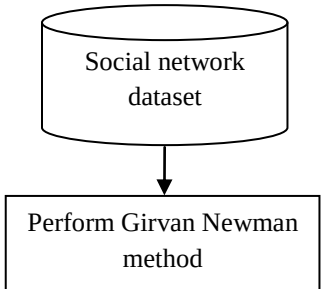
Girvan and Newman [2] proposed a divisive algorithm which uses betweenness score to find the inter community edges. Betweenness score for an edge is calculated based on the fraction of the shortest path between all pairs of nodes in the network that contains this edge. Inter community edges will have high betweenness score as it will be the part of many shortest paths in the network. This method suffers from high computational cost.

3.4 Role of Link Prediction in Community Detection

Link mining has two forms i.e. community detection and link prediction that are concerned in discovering the relations between vertices in the network. Several node similarity indices used in link prediction methods are closely related to the perception of community structures and hence provide important contribution into community detection methods. Communities are group of nodes in the network such that link between nodes are denser in same community and sparser in different communities. Varieties of algorithms are invented over years to find community structure in the network. For example Girvan Newman algorithm used betweenness to select link that should be removed between the communities, modularity optimization algorithm uses a function that measures the quality of an entire partition and tries to optimize it and so on. Link prediction used techniques of node similarity to predict the existing or future links that focus on the common features that share between pair of nodes. It is efficient and effective to predict missing links by considering the concept of community structure but the hypothesis also explored about enhancing community structures of a network with some extra information like node similarity even if the structure of the network is known.

IV. Proposed Methodology

This section proposes a novel approach for detecting communities with the support of link prediction that uses node similarity techniques based on local structure and path based techniques based on global structure of the network. This section analysed that link prediction plays vital role for finding communities within the network.



DOI- 10.18486/ijcsnt.2015.4.3.02
ISSN-2053-6283

Figure 1 Proposed framework of community detection using link prediction

Here T_p is prediction time. This illustration clearly provides a new idea of finding communities at time T to T_p by using link prediction methods. The algorithm takes input of network dataset i.e. karate club to perform Girvan Newman community detection algorithm. It is an attempt to identify the factions involved in the split of club and then apply various link prediction algorithm based on local and global structure of the network to compute the similarity measures afterword based on new linking probability it discover best communities within the network using link prediction methods at time T_p.

Algorithm: Community detection based on link prediction

Procedure:
Input : Karate club dataset
Output : new communities within network

1. Start
2. Select the dataset.
3. Perform Girvan-Newman algorithm to detect the communities.
    - Calculate the betweenness for all edges in the network.
    - Remove the edge with the highest betweenness.
    - Recalculate betweenness for all edges affected by the removal.
    - Repeat from point 2 until no edges remains.
4. Perform similarity based link prediction methods.
5. Resulting community shows all nodes are available at time T_p .
6. End

a. Dataset

In order to be able to apply all selected methods and taking into account the types of datasets available, the network is represented as a binary unweighted network. This enables us to reach a consistent and comprehensive review of the existing methods for detecting communities. Zachary's Karate Club [13] network is a well-known real world dataset for which the community structure is already known from other sources. This is small real world network, in which data is collected from member of university karate club over the period of two years by Zachary. In this, each node of the network represents the member of club and each edge represents the link between these members. This network contains only 34 nodes and 78 undirected edges. During the period of two years, Zachary observed that a disagreement developed between the administrator of the club and the club's instructor, which ultimately resulted in the instructor's leaving and starting a new club, taking about a half of the original club's members with him. Zachary constructed a network of friendships between members of the club, using a variety of measures to estimate the strength of ties between individuals. The network contains the description as follows

Parameters for Analysing Social Network

1. If a vertex v_i has k_i neighbours, $k_i(k_i - 1)/2$ edges could exist among the vertices within the neighbourhood.
2. Avg. degree is defined as

$$< k >= \frac{1}{N}\sum_{i=1}^{N} k_i \ldots\ldots\ldots (7)$$

3. The density [19] of a binary network is the total number of links divided by the total number of possible links. For a valued network, it is the total of all values divided by the number of possible links. In this case the density gives the average value. It is defined as

$$D = \frac{\text{total no. of links}}{\text{total no. of possible links}} = \frac{2|E|}{|V|(|V|-1)} \ldots\ldots\ldots (8)$$

where V is the number of nodes in the network.

b. Performance of Link Prediction Methods

The main goal of this research is to explore the correlations between the accuracy of different prediction methods and network indices to detect the communities. This section compares the performance of these simple prediction methods and evaluate the similarity score S_{xy} for Jaccard Coefficient (JC), Common Neighbour (CN), Adamic Adar (AA), Resource Allocation (RA), Preferential Attachment (PA) based on local structure and Katz based on global structure in which $\beta = 0.0005$ that controls the contribution of path to the similarity when applied to real world social network.

Table 1: The basic topological characteristics of karate club network

Link Prediction Methods	AUC
JC	0.5659
CN	0.6068
AA	0.6238
RA	0.6256
PA	0.6289
Katz	0.6196

Table 2 The results from karate club network

Vertices |V|	34
Edges |E|	78
Avg. Path Length	2.408
Max. Degree	17
Avg. Degree<k>	4.588
Density D	0.139

Table 2 shows that the performance of a given prediction method highly depends on the type of network being analysed.

V. Analysis

Figure 2 shows a Girvan Newman method for detect the community that divide the network into two communities. The colors correspond to the best partition found by optimizing the modularity of Girvan Newman. Girvan and Newman focused on the concept of betweenness, which is a variable expressing the frequency of the participation of edges to a process. Edge betweenness is the number of shortest paths between all vertex pairs that run along the edge. It is an extension to edges of the popular concept of site betweenness, introduced by Freeman and expresses the importance of edges in processes like information spreading, where information usually flows through shortest paths.

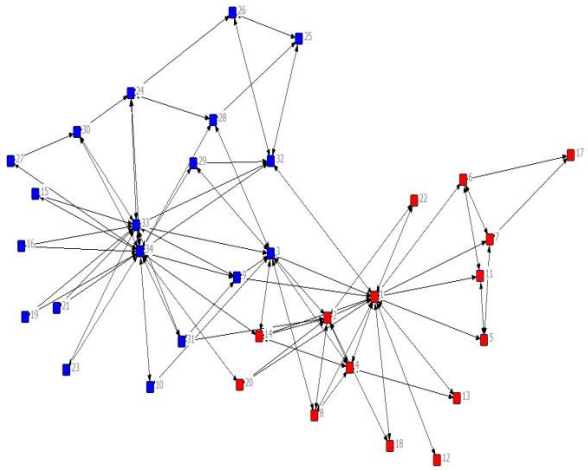


Figure 2 Communities in karate club after applying Girvan-Newman

a. Comparison among Various Link Prediction Algorithms

The area under the receiver operating characteristic curve is standard metric for measuring link prediction accuracy. AUC statistic is used to quantify the accuracy of prediction algorithms and test how much better they are than pure chance, similarly to the experimental protocol. Provided the score for all non-observed links, the AUC value can be interpreted as the probability that a randomly chosen missing link is given a higher similarity value than a randomly chosen non-existent link. To calculate AUC randomly select a missing link and a non-existent link to compare their scores at each time. If among n independent comparisons, there are n′ times the missing link having a higher score and n″ times they have the same score [20], the AUC value is

$$\text{AUC} = \frac{n' + 0.5n''}{n} \dots\dots\dots (9)$$

If all the scores are generated from an independent and identical distribution, the AUC value should be about 0.5. Therefore, the degree to which the value exceeds 0.5 indicates how better the algorithm performs than pure chance. The AUC values give the comparison of various link prediction methods perform as follows

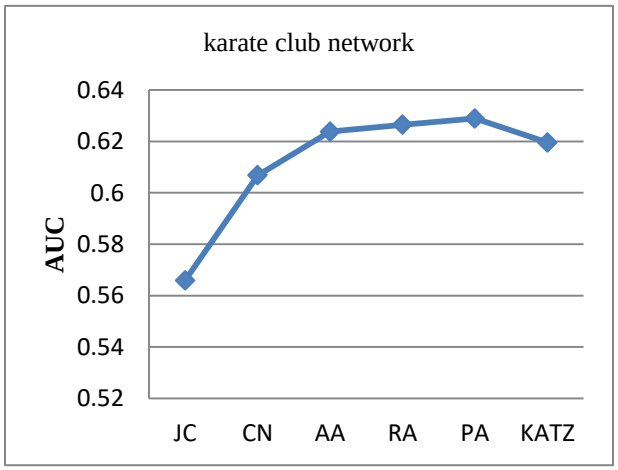


Figure 3 Prediction accuracy of karate club dataset

By comparing the variance of each method PA also provides the most stable prediction and for karate club network some of the prediction methods could provide a good prediction result. This type of network is called the "prediction friendly" network.

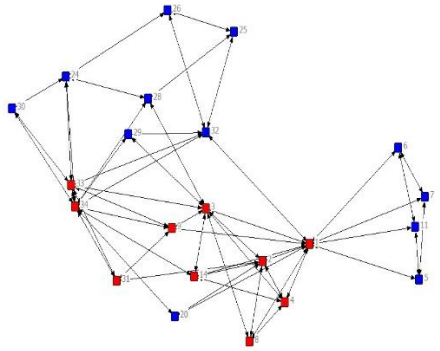


Figure 4 Communities in karate club at time T_p

This section tries to find out why the link prediction based community detection approach can give very good accuracy at future time T_p on the karate club dataset. Using the similarity based Girvan Newman community detection approach, it partition the network into communities and note that some node pairs with the community have a relatively highly connection likelihood. This implies that these node pairs are possible to form links within the communities and nodes with lower interactions are removed at time T_p. In similarity based link prediction algorithms, Preferential Attachment provides good prediction accuracy together with a relative immovability. The performance measure of Preferential Attachment is best on the karate club network which nearly same as Resource Allocation and Adamic-Adar. The performance of Jaccard Coefficient method is not suited in this network. Since the network is small the Katz method is perform better than the Jaccard Coefficient and Common Neighbours method but this is not always true for all type of networks while the size increases the performance of Katz method is decrease rapidly. Based on the best suited method new possible links were finds and then applying Girvan-Newman method to detect the communities that divide the network into best communities.

VI. Conclusion and Future Work

This paper shows an experimental analysis of karate club data set and provides an enhancement to the past Girvan Newman algorithm for finding communities. One of the most suitable and efficient mechanism available for detecting communities is similarity based community detection method using link prediction. In this paper, six link prediction algorithms are compared on the basis of similarity measures. It can be observed

that among the six algorithms, Preferential Attachment (PA) is the only algorithm that produces the better prediction accuracy. This article is based on the theory for network partitioning. It observes that the algorithm of link prediction based community detection can partition a karate club network into communities efficiently. This is because some measures of vertex similarity are nicely relevant to the definition of community detection. By applying it, to predict communities, the similarity based community detection algorithm exhibits slightly better accuracy of prediction and overwhelming better computational efficiency that the Girvan Newman approach. This research work focuses only on link prediction based community detection approach. Furthermore, for improvement of community detection algorithm to achieve higher accuracy more data set can further investigated from other domain for future perspective where some networks do not satisfy the power law distribution.

References

[1] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," in Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, NK, pp. 141–142, June 2005.

[2] E. M. Airoldi, "Mixed membership block models for relational data with application to protein-protein interactions," in Proceedings of International Biometric Society-ENAR Annual Meetings, 2006.

[3] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in SDM Workshop of Link Analysis, San Francisco, Calif, USA, 2006.

[4] Pallavi Gupta, Sanjiv Sharma, "A Survey on Link Prediction Problem in Social Network", International Journal for Science and Advance Research in Technology (IJSART) vol. 1 Issue 1, 2015.

[5] M. Girvan and M.E.J. Newman, "Community structures in social and biological networks" vol. 99, no. 12, pp. 7821-7826, 2002.

[6] M.E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks", Physical Review E, 69, page 026113, 2004.

[7] Bowen Yan and Steve Gregory, "Detecting community structure in networks using edge prediction methods", CoRR, abs/1201.3466, 2012.

[8] Fenhua Li et al. "A Clustering-based Link Prediction Method in Social Networks" vol. 29, pp. 432–442 2014.

[9] Liben-Nowell, D., Kleinberg, J., "The link prediction problem for social networks" In Proceedings 12th International Conference on Information and Knowledge Management (CIKM'2003).

[10] Tylenda, T., Angelova, R., Bedathur, S., "Towards time-aware link prediction in evolving social networks" In Proceedings 3rd Workshop on Social Network Mining and Analysis (SNA-KDD'2009), pp. 1–10.

[11] Liyan Dong et al. "The Algorithm of Link Prediction on Social Network", Hindawi Publishing Corporation, Mathematical Problems in Engineering, 2013, Article ID 125123, 7 pages.

[12] Linyuan Lu and Tao Zhou, "Role of weak ties in link prediction of complex networks", arXiv:0907.1728v2 [cs.IR] Aug 2009.

[13]Zachary W. "An information flow model for conflict and fission in small groups" Journal of Anthropological Research, vol. 33, pp. 452-473, 1977.

[14] G. Kossinets, Effects of missing data in social networks, Social Networks, vol. 28, no. 3, pp. 247-268, 2006.

[15] P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et des Jura, Bulletin de la Societe Vaudoise des Science Naturelles vol.37, no. 547, 1901.

[16] Lada A Adamic and Eytan Adar "Friends and neighbours on the web" Social networks, vol. 25, no. 3, pp. 211-230, 2003.

[17] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek "Evolution of the social network of scientific collaboration" Physica A, vol. 311(3-4), pp. 590-614, 2002.

[18] L. Katz, "A new status index derived from sociometric analysis", Psychmetrika, vol. 18, no. 1, pp. 39 -43, 1953.

[19] Anderson, B.S., Butts, C., and Carley, K. "The interaction of size and density with graph-level indices" Social Networks, 21, 239-267, 1999

[20] Liu Z.,Zhang Q.-M.,Lü L.,Zhou T. "Link prediction in complex networks: a local naive bayes model" EPL (Europhysics Letters), 96, 2011.