Link Prediction of Social Networks Based on Weighted Proximity Measures

Tsuyoshi Murata and Sakiko Moriyasu
Department of Computer Science, Graduate School of Information Science and Engineering
Tokyo Institute of Technology, Japan
murata@cs.titech.ac.jp

Abstract

Question-Answering Bulletin Boards (QABB), such as Yahoo! Answers and Windows Live QnA, are gaining popularity recently. Communications on QABB connect users, and the overall connections can be regarded as a social network. If the evolution of social networks can be predicted, it is quite useful for encouraging communications among users. This paper describes an improved method for predicting links based on weighted proximity measures of social networks. The method is based on an assumption that proximities between nodes can be estimated better by using both graph proximity measures and the weights of existing links in a social network. In order to show the effectiveness of our method, the data of Yahoo! Chiebukuro (Japanese Yahoo! Answers) are used for our experiments. The results show that our method outperforms previous approaches, especially when target social networks are sufficiently dense.

1. Introduction

Question-Answering Bulletin Boards (QABB), such as Yahoo! Answers and Windows Live QnA, are gaining popularity recently. Questions are submitted on QABB and let somebody in the internet answer them. As Kautz pointed out, the search for information often must come down to the search for person who holds the information privately [6]. Communications on QABB connect users, and the connections of users as a whole can be regarded as a social network. If the evolution of social networks can be predicted, it is quite useful for encouraging communications among users. For example, suitable questions can be recommended to potential answerers based on the structures of previous communications. Another example is to predict future "hot" questions that will attract many users.

Link prediction is one of the challenging research topics of link mining [3]. There are two main data sources for predicting links between nodes: 1) attributes of nodes, and 2) structural properties of networks that connect nodes. In the case of online social networks, nodes represent users and their attributes (personal information) are not always available. The latter data source (structural properties) is preferable for the purpose of predicting links of online social networks.

Although the links of practical social networks are not always uniform, previous approaches based on structural properties, such as Newman's common neighbors [8] and Adamic/Adar [1], do not take weights of links into consideration. In general, weights of links between users correspond to the number of times they meet or communicate.

This paper proposes new graph proximity measures, which are called weighted graph proximity measures, for improving the performance of link prediction for social networks. The measure is based on an assumption that new links can be predicted better by using both graph proximity measures and the weights of existing links in a social network. The weight of a link between two users in a social network is defined as the number of encounters of the users on QABB. The data of Yahoo! Chiebukuro (Japanese Yahoo! Answers) are used for our experiments. The results show that our method outperforms previous approaches, especially when target social networks are sufficiently dense.

2. Link Prediction for QABB

Data sources for link prediction can be broadly divided into the followings: a) link prediction based on attributes of nodes, and b) link prediction based on structural properties of graphs. The former approach is taken by and Popescul [10] and Hasan [4]. The latter approach is taken by Liben-Nowell [7] and Huang [5]. An approach using both data sources is attempted by O'Madadhain [9]. In this paper, we take approach b) for the link prediction of evolving online social networks. As mentioned above, this is because node attributes correspond to personal information and available personal information is limited.

Liben-Nowell presents a survey of predictors based on several graph proximity measures and compares their performance using academic co-authorship networks of physics [7]. In general, online communities of question-answering bulletin boards are

0-7695-3026-5/07 $25.00 © 2007 IEEE
DOI 10.1109/WI.2007.52

more "open" than academic co-authorship networks, and they are therefore more dynamic. In this paper, we would like to investigate 1) whether the predictors based on graph proximity measures are appropriate for predicting links of open and dynamic online social networks and 2) whether the predictors can be improved by taking weights into consideration.

3. Weighted Graph Proximities

As described above, link prediction based on graph proximity measure relies solely on structural properties of given network. The basic approach for predicting links is to rank all node pairs based on proximities in their graph. A connection weight score(x,y) is assigned to each pair of nodes x and y, and then produce a ranked list in decreasing order of score(x,y). For a node x, let $\Gamma(x)$ and w(x,y) denote the set of neighbors of x in a social network, and the weight of link between x and y respectively.

Several definitions of score(x,y) are proposed. Common neighbors [8] define score(x, y) as the number of neighbors that x and y have in common:

$$score(x,y) = \left|\Gamma(x) \cap \Gamma(y)\right|$$

This is based on an assumption that the more neighbors are in common, the more likely that nodes x and y will be connected. Adamic and Adar [1] refine the common neighbors by taking rarer neighbors more heavily. In other words, common neighbors of low degrees are taken more seriously in the following Adamic/Adar score:

$$score(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log\left|\Gamma(z)\right|}$$

Preferential attachment is based on an assumption that the probability that a new link involves node x is proportional to the number of its neighbors. The idea is famous as the growth model of the Web network [2].

$$score(x,y) = \left|\Gamma(x)\right| \times \left|\Gamma(y)\right|$$

In this paper, we propose new scores that take weights of links into account. Figure 1 shows an example of weighted common neighbors. Definition of the score of weighted common neighbor is given as follows:

$$score(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{2}$$

In Figure 1, each node represents a user, and a link between two nodes represents encounter(s) on QABB. Each number indicates the weight of nearby link, and a thick link represents more than one encounters on QABB. According to the definition of original common neighbors, score(x,y) is 2 (the number of intermediate nodes between x and y). For the calculation of weighted common neighbors, the upper intermediate node is weighted rather than the lower one because of the weight of the link between x and upper intermediate node. The score of weighted common neighbor is 2.5.

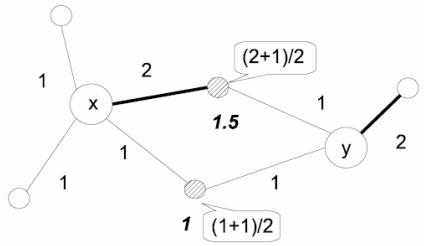

Figure. 1 Weighted common neighbors

Weighted Adamic/Adar and weighted preferential attachment are introduced in the same manner. Their definitions are given as follows:

$$score(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{2} \times \frac{1}{\log(\sum_{z' \in \Gamma(z)} w(z',z))}$$

$$score(x,y) = \sum_{x' \in \Gamma(x)} w(x',x) \times \sum_{y' \in \Gamma(y)} w(y',y)$$

4. QABB Data

The service of Yahoo! Chiebukuro (Japanese Yahoo! Answers, http://chiebukuro.yahoo.co.jp/) started on April 2004, and it is one of the most popular question and answering sites in Japan. A bulletin board is generated for each submitted question, and answers to the questions follow on the board.

The data we used for our experiments were recorded from September 1, 2005 to September 30, 2005. The data are divided into two groups, and the former (September 1 - 15) is used for training and the latter (September 16 - 30) is for testing. The total number of questions or answers is 1,081,104, and the number of users during the period is 58,755. The data is composed of encrypted user ID, message ID, categories, contents of the questions or answers, date, time, and so on. We have used encrypted user ID, categories, date and time in our experiments. A social network is generated by putting links to all the pairs of the answerers in each question. Contents of questions or answers are not used in our experiments.

Links between users who already exist in training period are the target for link prediction, which is the same as Liben-Nowell's experiments. For link prediction, proximities between all the pairs of users have to be calculated. We divide the whole QABB data into categories, and generate a social network for each category. This is because the whole social network is too big to analyze, and because more than 1/3 of users submit questions or answers to only one category.

5. Experiments

Based on the above graph proximity measures, experiments of link prediction for QABB social networks are performed.

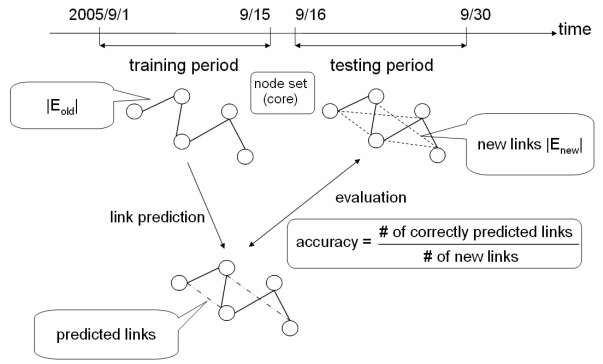

Figure. 2 Overall Procedures for Our Experiments

Table 1. Performance of Link Predictions for QABB

Categories	CN	CNw	AA	AAw	PA	PAw	RD
Yahoo!	29.5	32.0	29.9	32.2	24.5	24.7	2.8
News	23.5	25.2	23.8	25.4	25.2	25.9	3.1
Health	15.7	17.4	16.0	16.9	16.6	17.1	1.3
Children	20.5	22.9	22.3	23.0	19.4	22.0	2.4
Manners	29.2	30.2	29.4	30.3	27.5	27.6	5.3
Sports	23.2	25.4	24.8	25.6	16.2	15.9	2.1
Entertainments	15.2	16.1	15.3	16.1	14.4	14.6	1.6
Life	18.2	18.7	18.3	19.2	18.7	18.9	1.5
Science	15.8	15.9	16.1	16.4	12.6	12.3	1.4
Travel	20.1	22.0	20.5	22.0	16.0	15.2	2.3
Business	26.3	26.3	26.9	27.6	19.6	19.0	3.6
Internet	18.6	18.9	19.2	19.4	17.5	17.9	1.5
Jobs	14.5	14.9	16.9	16.9	16.6	15.0	2.2
Average	20.8	22.0	21.5	22.4	18.9	18.9	2.4

Figure 2 shows the overall procedures for our experiments. QABB data are divided into training period and testing period according to the timestamp of posting to QABB. Social networks are generated based on the former data. As described in section 3, a node in the social networks represents a user and an edge represents an encounter between users on a question-answering. Graph proximity measures for each pairs of nodes are calculated. Links that will appear in testing period are predicted in decreasing order of the graph proximity measures. Accuracies of the predictions are evaluated by comparing them with new links that actually appeared in the testing period.

Table 1 shows the percentages of the accuracies of link predication by original and weighted proximity measures of common neighbor, Adamic/Adar, and preferential attachment as well as random prediction. In the Table, CN, AA, PA, and RD indicate common neighbors, Adamic/Adar, preferential attachment, and random respectively. CNw, AAw, and PAw are weighted proximity measures of CN, AA, and PA, respectively.

6. Discussion

6.1 Link prediction for open and dynamic online social networks

In Lieben-Nowell's experiments, the numbers of core edges are from 486 to 1790, the numbers of |Eold| (the number of edges during training period) are from 519 to 6654, and the numbers of |Enew| (the number of newly generated edges in testing period) are from 400 to 5751. The numbers of nodes in our experiments are about ten times of those of Lieben-Nowell's experiments. Social networks of Yahoo! Chiebukuro are open to public and their numbers of users are much larger. Table 1 shows that link prediction based on graph proximity measures is effective for open and dynamic online social networks.

It is often reported that users of online social networks often misrepresent their personal attributes such as age, gender, job and so on. Our approach use structural properties of social networks only; it does not use any information about node (user) attributes. Link prediction based on graph proximity measures is thus promising for analyzing online social networks.

6.2 Performance improvements of graph proximity measures

6.2.1 Link prediction based on original graph proximity measures

- Link prediction based on graph proximity measures perform better for denser graphs

Performances of link prediction are quite different among categories. We focus on "Health", "Entertainments", "Internet", and "Jobs" that are relatively worse performance among all categories. Analysis of the degree distributions shows that the percentages of high-degree nodes in these social networks are small. Let us suppose that 70% of the maximum number of degree in each social network as the threshold for high-degree nodes. The numbers of nodes of high-degree nodes for the above categories are 4, 42, 6, and 10 respectively (less than 3% of overall nodes). On the other hand, social networks of the categories of "Manners" and "Business" contain more high-degree nodes (8%-14% of overall nodes). Based on the result, we can assume that the percentages of high-degree nodes of social networks affect the performance of link prediction. If a social

network is sparse and is composed of low-degree nodes, many of the nodes will be disconnected from others, and differences among the values of graph proximity measures become obscure.

- Adamic/Adar performs better than common neighbors

As you can observe from Table 1, Adamic/Adar is the best and stable graph proximity measures for link prediction. Common neighbor is the second-best performance. This is the same as the results of Lieben-Nowell's experiments.

- Preferential attachment performs worse for networks whose degree distributions are almost uniform

Performance of referential attachment is the worst among the three graph proximity measures. Preferential attachment is based on the idea that high-degree nodes will have more chances of getting more edges. If the degree distribution is almost uniform, this "rich get richer" strategy is not appropriate.

6.2.2 Link prediction based on weighted graph proximity measures

You can see from Table 1 that our weighted Adamic/Adar outperforms the original Adamic/Adar further. This shows that the number of encounters (weights) on QABB is an important factor for measuring proximities among users. In Table 1, categories are sorted in decreasing order of average number of answers for each bulletin board. In general, better predictions can be made for denser social networks (for upper categories in the table) by our weighted graph proximity measures.

Weighted common neighbors also outperform original common neighbors for almost all categories. Weighted preferential attachment is slightly better than original preferential attachment only when social networks are relatively dense. This is because weighted preferential attachment takes low-degree nodes that are connected with high-weight edges too seriously in the process of calculating score(x,y), which is against the idea of "rich get richer" strategy.

7. Conclusion

This paper shows that link prediction based on graph proximity measures is suitable for open and dynamic online social networks. We propose new weighted graph proximity measures for link prediction of social networks. By taking weights of links into consideration, the performances of link predictions are improved rather than previous proximity measures. We can expect that further improvements can be made by treating more recent links as more important, which is left for our future work.

8. Acknowledgements

We would like to express our thanks to Mr. Makoto Okamoto (Yahoo! Japan), Prof. Kikuo Maekawa (The National Institute for Japanese Language), and Prof. Sadaoki Furui (Tokyo Institute of Technology) for allowing us to use the data of Yahoo! Chiebukuro.

9. References

[1] Adamic, L. A., Adar, E., Friends and Neighbors on the Web, Social Networks, Vol.25, No.3, pp.211-230, 2003.

[2] Barabasi, A. L., Linked – The New Science of Networks, Perseus, 2002.

[3] Getoor, L., Diehl, C. P., Link Mining: A Survey. SIGKDD Explorations, Vol.7, No.2, pp.3-12, 2005.

[4] Hasan, M. A., Chaoji, V., Salem, S., Zaki, M. Link Prediction using Supervised Learning, in workshop on link discovery; issues, approaches and applications, 2005.

[5] Huang, Z., Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient, in Proceedings of LinkKDD, 2006.

[6] Kautz, H., Selman, B., Shah, M. The Hidden Web. AI Magazine, Vol. 18, No. 2, 27-36, 1997.

[7] Liben-Nowell, D., Kleinberg, J. The Link Prediction Problem for Social Networks, in Proceedings of CIKM, pp.556-559, 2003

[8] Newman, M. E., Clustering and Preferential Attachment in Growing Networks, Physical Review Letters E, Vol.64 (025102), 2001.

[9] O'Madadhaim, J., Hutchins, J., Smyth, P., Prediction and ranking algorithms for event-based network data, SIGKDD Explorations, Vol.7, No.2, pp.23-30, 2005.

[10] Popescul, A., Ungar, L. H. Statistical relational learning for link prediction, in IJCAI workshop on learning statistical models from relational data, 2003.