Link Prediction for Question-Answering Bulletin Boards

Tsuyoshi Murata
Tokyo Institute of Technology
W8-59 2-12-1 Ookayama, Meguro
Tokyo 152-8552 Japan
+81-3-5734-2684
murata@cs.titech.ac.jp

Sakiko Moriyasu
Tokyo Institute of Technology
W8-59 2-12-1 Ookayama, Meguro
Tokyo 152-8552 Japan
sakiko@ai.cs.titech.ac.jp

ABSTRACT
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. Communications on QABB connect users, and the overall connections can be regarded as a social network. If the evolution of the network can be predicted, it is quite useful for encouraging communications among users. This paper describes a method for predicting links based on the structure of social network. The method is based on an assumption that new links can be predicted better using both graph proximity measures and the weights of existing links in a social network. The data of Yahoo! Chiebukuro (Japanese Yahoo! Answers) are used for our experiments. The results show that our method is better than previous approaches, especially when target social network is sufficiently dense.

Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.*

General Terms
Algorithms, Experimentation, Human Factors.

Keywords
Link prediction, social network, question-answering bulletin board.

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 [3]. Communications on QABB connect users, and the connections of users as a whole can be regarded as a social network. If the evolution of such social network can be predicted, it is quite useful for encouraging communications among users. Link prediction is one of the challenging research topics of link mining [1].

This paper proposes a new graph proximity measure, which is called weighted graph proximity measure, on social networks in order to improve link prediction. The measure is based on an assumption that new links can be predicted better 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 is better than previous approaches, especially when target social network is sufficiently dense.

2. Link Prediction for QABB
Based on the taxonomy of common link mining tasks described by Getoor [1], tasks of link mining are broadly categorized as the following: 1) object(node)-related tasks (object ranking, object classification, object clustering, and object identification), 2) link(edge)-related tasks (link prediction), and 3) graph-related tasks (subgraph discovery, graph classification, and generative models for graphs). Link prediction belongs to 2), and it is the problem of predicting the existence of an edge between two nodes based on attributes of the nodes and other observed edges. Examples of link prediction include predicting links among actors in social networks (such as predicting friendship), and predicting interactions among proteins of metabolic networks in the field of bioinformatics.

Approaches of link prediction can be broadly divided into the following: a) link prediction based on node attributes, b) link prediction based on graph structures, and c) link prediction based on both node attributes and graph structures. Approach a) is taken by [5] and [6]. Approach b) is taken by [4]. Approach c) is taken by [2]. Approach a) can be regarded as naïve extension of binary classification problem. We take approach b) for the link prediction of evolving online social network. This is because node attributes mean personal information and most of them are not open to public. Liben-Nowell presents a survey of predictors based on several graph proximity measures and compares their performance using academic co-authorship networks of physics [4]. In general, online communities of question-answering bulletin boards are relatively “open” rather than academic co-authorship networks, and the speeds for evolution are quite different.

We would like to investigate 1) whether the same predictors based on graph proximity measures are appropriate for predicting new links of open online social networks and 2) whether the predictors can be improved by taking weights into consideration. If better predictors can be developed, they can be used for encouraging communications among users. For example, suitable questions can be recommended to potential answerers based on the structure of previous communications. Another example is to predict future “hot” questions that attract many users.

3. WEIGHTED GRAPH PROXIMITY

Link prediction based on graph proximity measure is an approach to make prediction entirely based on structural properties of given network. A connection weight *score(x,y)* is assigned to pairs 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 the social network, and the weight of links between x and y respectively. The followings are the original *score(x,y)* of common neighborhoods, Adamic/Adar, and preferential attachment [4].

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

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

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

The followings are the new *score(x,y)* that we propose in this paper: weighted common neighbors, weighted Adamic/Adar, and weighted preferential attachment.

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

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

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

Original Adamic/Adar refines common neighbor (simple counting of intermediate nodes) by weighting nodes of fewer outlinks heavily. Weighted Adamic/Adar refines the Adamic/Adar further by taking weights of links into consideration. Figure 1 is for the explanation of weighted Adamic/Adar. 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. In figure 1, *score(x,y)* of original Adamic/Adar is 1/log4 + 1/log3, while our weighted Adamic/Adar is 1.5/log5 + 1/log4. The idea of putting weights to links is based on an assumption that nodes x and y are closely related when 1) there are more intermediate nodes between them, 2) intermediate nodes have fewer outlinks, and 3) links between x (or y) and intermediate nodes have more weights, which means more encounters between x (or y) and intermediates on QABB.

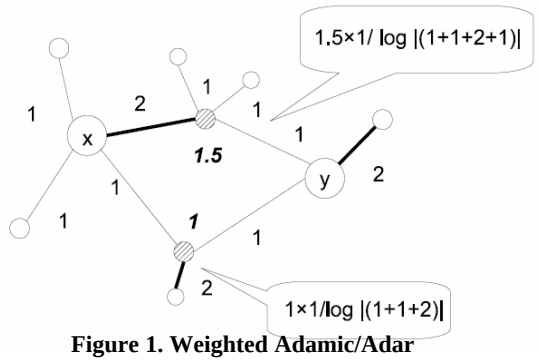

Figure 1. Weighted Adamic/Adar

4. QABB DATA

The service of Yahoo! Chiebukuro (http://chiebukuro.yahoo.co.jp) started on April 2004, and it is one of the most popular question and answering sites in Japan. QABB services are popular especially in Japan and Korea. English version (Yahoo! Answers) started on December 2005. 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 and time, and so on. We have used encrypted user ID, categories, date and time in order to generate a social network. A social network is generated by connecting links to all the pairs of the questioner and 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 considered. To avoid computational explosion, the above QABB data are divided by the categories, and social networks are generated for every categories. This is because more than 1/3 of users submit questions or answers to only one category.

5. EXPEIRMENTS

Based on the above graph proximity measures, experiments of link prediction for QABB social network are performed. Table 1 shows the number of nodes, answers, and edges during training period as well as the number of edges, $|E_{old}|$(the number of edges during training period), and $|E_{new}|$ (the number of newly generated edges in test period) between the nodes that exist in both training period and test period.

Table 1. Sizes of Social Networks for All Categories.

	Learning period			Core		
	nodes	answers	edges	edges	$\|E_{old}\|$	$\|E_{new}\|$
Yahoo!	5820	39546	94449	1290	43525	44963
News	4992	34433	76070	862	26889	22055
Health	9991	59931	149230	2351	66522	67970
Childern	5804	26752	76197	1133	29243	30572
Manners	2782	10003	29535	414	8183	7733
Sprots	4789	25660	49937	1102	21459	22699
Entertainments	7454	29734	68538	1411	22977	29966
Life	5409	21529	40026	985	13859	14736
Science	4568	17048	28486	813	8442	8963
Travel	3109	10321	17327	470	4562	4575
Business	2103	6256	10533	278	2198	2658
Internet	3198	14887	13573	575	5111	5106
Jobs	2179	5001	7811	206	807	893

Table 2. Percentages of Link Predictions for QABB Networks

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

Table 2 shows the results 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.

Table 3. Maximum Degrees for Each Category

	Max Deg	Max Ans	Ave Ans
Yahoo!	1070	563	4.57
News	1764	1096	4.35
Health	4356	2782	4.23
Children	986	273	4.61
Manners	591	154	4.79
Sports	483	226	3.62
Entertainments	749	293	3.26
Life	1195	625	3.12
Science	409	194	2.81
Travel	454	181	2.82
Business	254	178	2.81
Internet	932	1231	2.16
Jobs	322	126	2.74

Table 3 shows the results of the maximum number of degrees, the maximum number of answers, and the average number of answers for each category. Categories are sorted in decreasing order of average answers, which roughly corresponds the average degrees of social networks. This table is for analyzing the relation between densities of social networks and their prediction performances.

6. DISCUSSIONS

In Lieben-Nowell's experiments, the numbers of core edges are 486-1790, the numbers of $|E_{old}|$ are 519-6654, and the numbers of $|E_{new}|$ are 400-5751. Table 1 shows that 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 the numbers of users among them are much larger. Table 2 shows that link prediction based on graph proximity measures is effective for such different network structures.

It is often reported that users of online communities often tell lies about their personal attributes such as age, gender and occupation. Our approach does not use any information about node (user) attributes. An approach of link prediction based on graph proximity measures is thus useful for online social networks.

6.1 Link prediction based on non-weighted proximity measures

- Three graph proximity measures perform better for denser graphs

Performances of link prediction are quite different among categories. We will 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, all the values of graph proximity measures are small and differences among the values become obscure.

- Adamic/Adar performs better than common neighbors

As you can observe from the Table 2, 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 Link prediction based on weighted proximity measures

You can see from Table 2 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 the table, 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 online social networks. We propose new graph proximity measures for link prediction of social networks. By taking weights of links into consideration, our proximity measures perform better than previous ones. Further improvements can be made by treating more recent links as more important than older ones, which is left for our future work.

8. ACKNOWLEDGMENTS

We would like to express our thanks to Dr. Makoto Okamoto (Yahoo! Japan Corporation), 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] Getoor, L., Diehl, C. P. Link Mining: A Survey. SIGKDD Explorations, Vol.7, No.2, 3-12, 2005.

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

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

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

[5] 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.

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