New Generation Computing, 26(2008)245-257
Ohmsha, Ltd. and Springer

Link Prediction based on Structural Properties of Online Social Networks

Tsuyoshi MURATA, Sakiko MORIYASU
Department of Computer Science, Tokyo Institute of Technology
W8-59 2-12-1 Ookayama, Meguro, Tokyo 152-8552 JAPAN
murata@cs.titech.ac.jp

Received 31 August 2007
Revised manuscript received 10 January 2008

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 social networks can be predicted, it is quite useful for encouraging communications among users. Link prediction on QABB can be used for recommendation to potential answerers.

Previous approaches for link prediction based on structural properties do not take weights of links into account. 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.

Keywords: Web Usage Mining, Link Prediction, Question-Answering Bulletin Boards.

§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.[7] 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[9] 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

Based on the taxonomy of common link mining tasks described by Getoor,[3] tasks of link mining are broadly categorized as the followings:

1. object(node)-related tasks (object ranking, object classification, object clustering, and object identification)
2. link(edge)-related tasks (link prediction)
3. graph-related tasks (subgraph discovery, graph classification, and generative models for graphs)

Link prediction belongs to the second category, and it is the problem of predicting the existence of an edge between two nodes based on attributes of the nodes and/or 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.

There are several related works for link prediction. However, many of them are not applicable for predicting links of online social networks, which is our target in this paper. The followings are the reasons that previous works are not appropriate for our target.

Different network properties Sarukkai's work[12)] for link prediction is about Web access patterns using Markov Chains. He proposes Web sequence modeling, which is useful only for navigational purposes. Kashima[6)] proposes a parameterized probabilistic model of network evolution and efficient prediction algorithm for the model. The model is based on copy and paste mechanism between edges. Although it is appropriate for biological networks, it is not appropriate for online social networks.

No node attributes O'Madadhain[10)] proposes prediction algorithms for event-based network data. His approach combines feature vectors of node attributes and structural information of edges. In the case of online social networks, node attributes mean personal information and they are not available. Popescul[11)] introduces a structured logistic regression model that can make use of relational features to predict the existence of links. The model depends on node attributes, which are not available in our experimental setting. Hasan[4)] proposes a method for identifying a set of attributes which are key to the performance of link prediction under supervised learning setup. The method also depends on node attributes.

Computationally infeasible Taskar[13)] proposes probabilistic model for the whole networks instead of predicting each link separately. Although the model have a universal application, it is computationally costy. Calculation of parameters using maximum likelihood estimation needs approximation or sampling. Huang[5)] proposes a new link prediction method based on generalized clustering coefficient. However, his cycle formation model and parameter estimation require much computational time.

As the approaches of link prediction based solely on structural information, Liben-Nowell[8)] presents a survey of predictors based on several graph proximity measures and compares their performance using academic co-authorship networks of physics. In general, online communities of question-answering bulletin boards are more "open" rather 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 Proximity

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. Neighbors of x mean the nodes that are directly connected to x with an edge.

Several definitions of $score(x, y)$ are proposed. Common neighbors[9] define $score(x, y)$ as the number of neighbors that x and y have in common:

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

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:

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

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].

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

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 Fig. 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

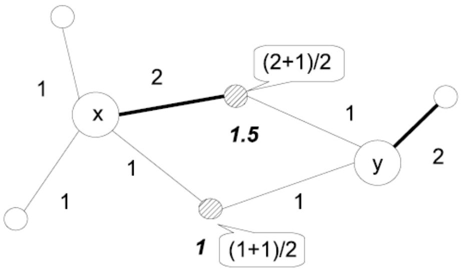


Fig. 1 Weighted Common Neighbors

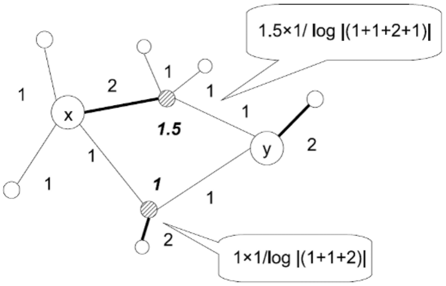


Fig. 2 Weighted Adamic/Adar

because of the weight of the link between x and upper intermediate node. The score of weighted common neighbor is 2.5.

Figure 2 shows an example of weighted Adamic/Adar. Definition of its score is given as follows:

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

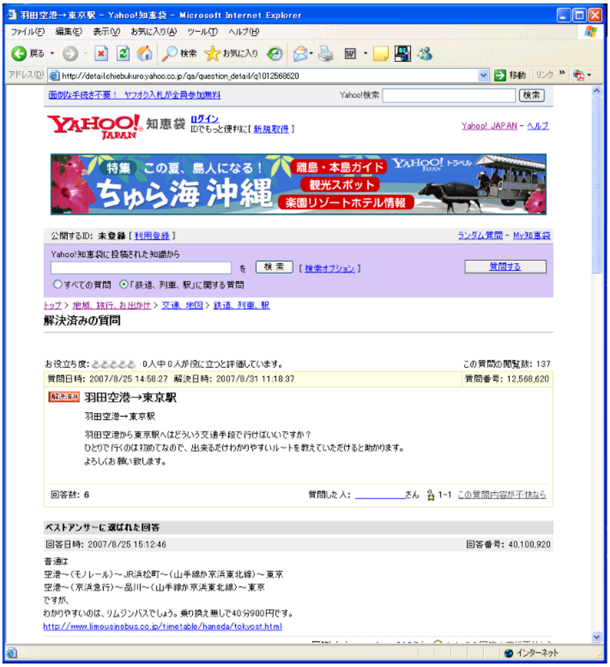
Original Adamic/Adar refines common neighbor (simple counting of intermediate nodes) by weighting nodes of fewer outlinks heavily. Weighted Adamic/Adar refines the Adamic/Adar further by taking weights of links into consideration. In Fig. 2, $score(x, y)$ of original Adamic/Adar is 1/log4 + 1/log3, while 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.

Weighted preferential attachment is introduced in the same manner. Its definition is given as follows:

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

§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. It is often reported that QABB services are popular especially in Japan and Korea. The English version of Yahoo! Answers started on December 2005. In either service, a bulletin board is generated for each submitted question, and answers to the questions follow on the board. Figure 3 shows an example of the Question-Answering on Yahoo! Chiebukuro. In this example, a questioner asks the easiest way from Haneda Airport to Tokyo


羽田空港→東京駅 - Yahoo!知恵袋 - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り
アドレス(D) http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q1012568620
移動 リンク
面倒な手続き不要！ ヤフオク入札が全員参加無料
Yahoo!検索 検索
YAHOO! JAPAN 知恵袋 ログイン
IDでもっと便利に[新規取得]
Yahoo! JAPAN - ヘルプ
特集 この夏、島人になる！
ちゅら海沖縄
離島・本島ガイド
観光スポット
楽園リゾートホテル情報
YAHOO! JAPAN トラベル
公開するID: 未登録 [利用登録]
ランダム質問 - My知恵袋
Yahoo!知恵袋に投稿された知識から
を 検索 [検索オプション]
質問する
○すべての質問 ●「鉄道、列車、駅」に関する質問
トップ > 地域、旅行、お出かけ > 交通、地図 > 鉄道、列車、駅
インターネット

解決済みの質問

お役立ち度: 0人中0人が役に立つと評価しています。 この質問の閲覧数: 137

質問日時: 2007/8/25 14:58:27 解決日時: 2007/8/31 11:18:37 質問番号: 12,568,620

解決済み 羽田空港→東京駅

羽田空港→東京駅

羽田空港から東京駅へはどういう交通手段で行けばいいですか？
ひとりで行くのは初めてなので、出来るだけわかりやすいルートを教えていただけると助かります。
よろしくお願い致します。

回答数: 6 質問した人: さん 1-1 この質問内容が不快なら

ベストアンサーに選ばれた回答

回答日時: 2007/8/25 15:12:46 回答番号: 40,100,920

普通は
空港~(モノレール)~JR浜松町~(山手線か京浜東北線)~東京
空港~(京浜急行)~品川~(山手線か京浜東北線)~東京
ですが、
わかりやすいのは、リムジンバスでしょう。乗り換え無しで40分900円です。
http://www.limousinebus.co.jp/timetable/haneda/tokyost.html

Fig. 3 Yahoo! Chiebukuro

Station. Many other users post answers to the question, and one of them is selected as the best answer.

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 pre-

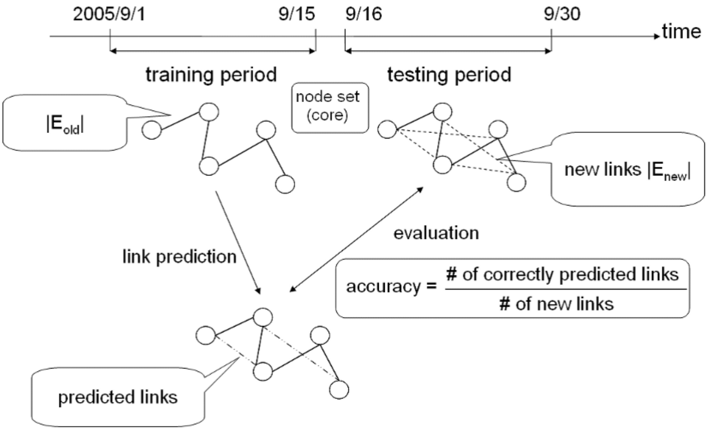


Fig. 4 Overall Procedures for Our Experiments

diction for QABB social networks are performed. Figure 4 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 encounter(s) between connecting users on question-answering bulletin board(s). Each edge is weighted by the number of encounters of its connecting users. Graph proximity measures for every pair of nodes are then calculated. Link candidates that connect the pairs of nodes are sorted in decreasing order of the graph proximity measures. Then $|Enew|$ high-ranking link candidates are predicted, where $|Enew|$ is the number of newly generated edges in testing period. Accuracies of the predictions are calculated as the number of correctly predicted links divided by $|Enew|$.

Table 1 shows the number of nodes, answers, and edges during training period as well as the number of edges, $|Eold|$ (the number of edges in training

Table 1 Sizes of Social Networks for All Categories

Categories	Training period			Core		
	nodes	answers	edges	edges	$\|E_{old}\|$	$\|E_{new}\|$
Yahoo! JAPAN	5820	39546	94449	1290	43525	44963
News	4992	34433	76070	862	26889	22055
Health	9991	59931	149230	2351	66522	67970
Children	5804	26752	76197	1133	29243	30572
Manner	2782	10003	29535	414	8183	7733
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 the Performance of Link Predictions for QABB Networks

Categories	CN	CNw	AA	AAw	PA	PAw	RD
Yahoo! JAPAN	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
Manner	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

period), and $|Enew|$ (the number of newly generated edges in testing period) between the nodes that exist in both training period and testing period.

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. Computational time for performing an experiment for one category is about 10-300 minutes, which depends on the size of target social network.

Figure 5 shows a bar chart of the performance of Yahoo! JAPAN category. Y-axis represents the percentage of correct link predictions, and the bars correspond to weighted common neighbors, original common neighbors, weighted Adamic/Adar, original Adamic/Adar, weighted preferential attachment, original preferential attachment, and random respectively. Figure 6 shows a bar chart of the performance of the average of all categories.

In order to evaluate the performance of proposed methods in detail, ROC

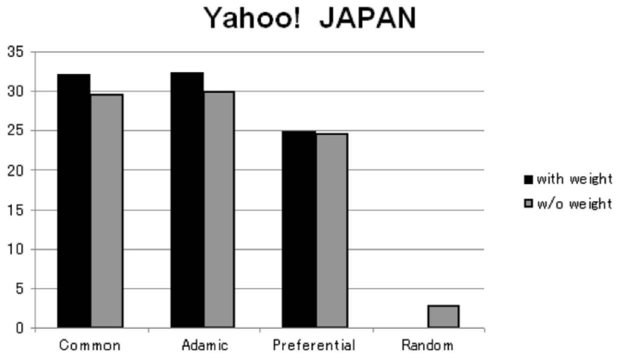


Fig. 5 Performance of Link Prediction for Yahoo! JAPAN Category

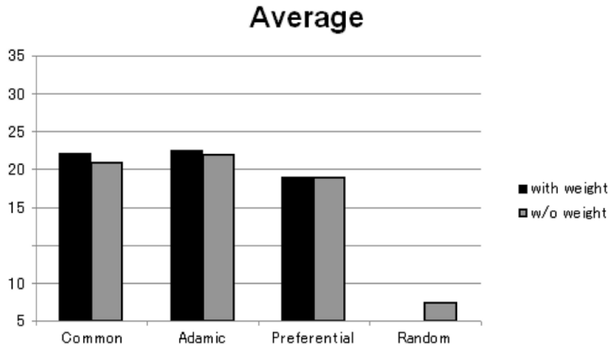


Fig. 6 Performance of Link Prediction for the Average of All Categories

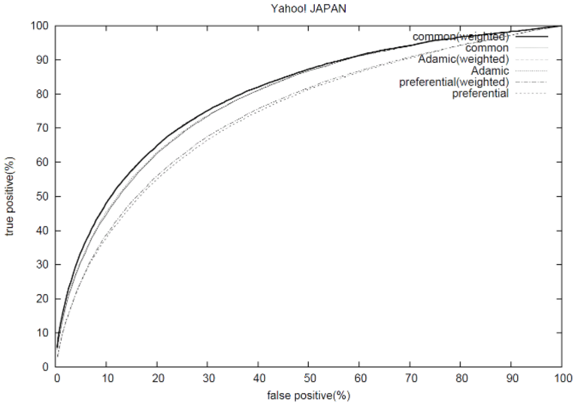


Fig. 7 ROC Curves for Yahoo! JAPAN Category

curves for the experiments are plotted. A ROC curve is a graphical plot of the fraction of true positives on the vertical axis against the fraction of false positives on the horizontal axis. The performance of plotted algorithm is better if its ROC curve is nearer the upper left-hand corner. The curve is often used as the evaluation of link prediction tasks.[5)] Figure 7 shows a ROC curve of the experiment of Yahoo! JAPAN. Focused one is also shown in Fig. 8. These curves show that the proposed weighted graph proximities always outperform original graph proximities.

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.

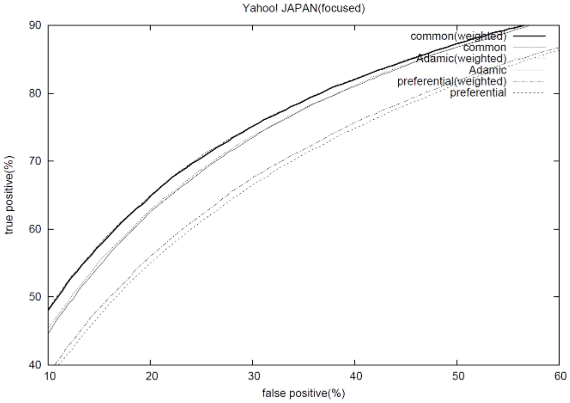


Fig. 8 Focused ROC Curves for Yahoo! JAPAN Category

Table 3 Maximum Degrees for Each Category

	Max Deg	Max Ans	Ave Ans
Yahoo! JAPAN	1070	563	4.57
News	1764	1096	4.35
Health	4356	2782	4.23
Childern	986	273	4.61
Manner	591	154	4.79
Sprots	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

§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 $|Eold|$ are from 519 to 6654, and the numbers of $|Enew|$ are from 400 to 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 their numbers of users are much larger. Table 2, Fig. 5 and Fig. 6 show 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

[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 “Manner” 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 Fig. 5 and Fig. 6, 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.

[2] Link prediction based on weighted graph proximity measures

You can see from Table 2, Fig. 5 and Fig. 6 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 2, 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 outperforms 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.

Acknowledgements

We would like to express our thanks to Mr. 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.

References

1) Adamic, L. A., Adar, E., "Friends and Neighbors on the Web," *Social Networks*, 25(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, 7(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 (LinkKDD-2005)*, 2005.

5) Huang, Z., "Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient," in *Workshop on Link Analysis; Dynamics and Static of Large Networks, (LinkKDD-2006)*, 2006.

6) Kashima, H., Abe, N., "A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction," in *Proc. of the Sixth IEEE Int. Conf. on Data Mining(ICDM'06)*, 2006.

7) Kautz, H., Selman, B., Shah, M., "The Hidden Web," *AI Magazine, 18(2)*, pp. 27-36, 1997.

8) Liben-Nowell, D., Kleinberg, J., "The Link Prediction Problem for Social Networks," in *Proc. of the Twelfth Int. Conf. on Information and Knowledge Management (CIKM)*, pp.556-559, 2003.

9) Newman, M. E., "Clustering and Preferential Attachment in Growing Networks," *Physical Review Letters E, 64 (025102)*, 2001.

10) O'Madadhaim, J., Hutchins, J., Smyth, P., "Prediction and ranking algorithms for event-based network data," *SIGKDD Explorations, 7(2)*, pp.23-30, 2005.

11) Popescul, A., Ungar, L. H., "Statistical relational learning for link prediction," in *IJCAI Workshop on Learning Statistical Models from Relational Data*, 2003.

12) Sarukkai, R. R., "Link Prediction and Path Analysis Using Markov Chains," in *Proc. of the Ninth Int. World Wide Web Conf. (WWW9)*, 2000.

13) Taskar, B., Wong, M.-F., Abbeel, P., Koller, D., "Link Prediction in Relational Data," in *Proc. of Neural Information Processing Systems Conf. (NIPS)*, 2003.

Tsuyoshi Murata, Dr.: He is an associate professor in the Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology. He obtained his doctor's degree in Computer Science at Tokyo Institute of Technology in 1997, on the topic of Machine Discovery of Geometrical Theorems. At Tokyo Institute of Technology, he conducts research on Web mining, diagrammatic reasoning and machine discovery.

Sakiko Moriyasu: She is a master course student of the Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology.