Association for Information Systems
AIS Electronic Library (AISeL)

SAIS 2014 Proceedings Southern (SAIS)

4-14-2014

A NETWORK LINK PREDICTION MODEL BASED ON OBJECT-OBJECT MATCH METHOD

Woo-Hyuk Jang
Korea Advanced Institute of Science and Technology, torajim@kaist.ac.kr

Myungjae Kwak
Middle Georgia State College, myungjae.kwak@mga.edu

Dong-Soo Han
Korea Advanced Institute of Science and Technology, dshan@kaist.ac.kr

Follow this and additional works at: http://aisel.aisnet.org/sais2014

Recommended Citation
Jang, Woo-Hyuk; Kwak, Myungjae; and Han, Dong-Soo, "A NETWORK LINK PREDICTION MODEL BASED ON OBJECT-OBJECT MATCH METHOD" (2014). *SAIS 2014 Proceedings*. 14.
http://aisel.aisnet.org/sais2014/14

This material is brought to you by the Southern (SAIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in SAIS 2014 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

A NETWORK LINK PREDICTION MODEL BASED ON OBJECT-OBJECT MATCH METHOD

Woo -Hyuk Jang
Korea Advanced Institute of Science and Technology
torajim@kaist.ac.kr

Myungjae Kwak
Middle Georgia State College
myungjae.kwak@mga.edu

Dong-Soo Han
Korea Advanced Institute of Science and Technology
dshan@kaist.ac.kr

ABSTRACT

In this paper, we proposed and evaluated a new network link prediction method that can be used to predict missing links in a social network. In the proposed model, to improve the prediction accuracy, the network link prediction problem is transformed to a general object-object match prediction problem, in which the nodes of a network are regarded as objects and the neighbors of a node are regarded as the node's associated features. Also a machine learning framework is devised for the systematic prediction. We compare the prediction accuracy of the proposed method with existing network link prediction methods using well-known network datasets such as a scientific co-authorship network, an e-mail communication network, and a product co-purchasing network. The results showed that the proposed approach made a significant improvement in all three networks. Also it reveals that considering the neighbor's neighbors are critical to improve the prediction accuracy.

Keywords

Network science; network link prediction; object-object match prediction; feature cohesion and coupling

INTRODUCTION

With the recent explosion of social networks, whose nodes represent members or entities and whose edges represent interaction between entities, many researchers have studied how networks grow and how members are connected in a social network (Liben-Nowell and Kleinberg, 2007; Adamic and Adar2003). Among those network studies, network link prediction, which is a technique to predict missing nodes or undiscovered edges in a network based on the patterns of existing topology, has been known as one of the most extensively used techniques (Liben-Nowell and Kleinberg, 2007; Adamic and Adar2003)**.** Typical applications that can utilize the link prediction are friends recommendation, terrorist or criminal network analysis, and effective grouping of a task force team in a company (Dombroski and Carley, 2002; Xu and Chen, 2005). For example, using the technique, on-line shopping mall systems can recommend new items to their customers from the purchasing pattern analysis, and social network sites can recommend potential friends based on the present connections of users. Since the accuracy of link prediction is a critical factor for the quality of such recommendations, researchers and practitioners have tried to improve the prediction accuracy.

Studies of network link prediction have focused on two main streams: neighbor-based methods and path distance-based methods (Katz, 1953; Linyuan and Zhou, 2011; Sarukkai, 2000; Zhu, Hong and Hughes, 2002; Newman, 2001 and 2003). While the neighbor-based methods suppose that the connection probability of two nodes is determined by the nodes commonly connected to the two nodes, the path-based methods assume that the connection probability is determined by how short the paths between two nodes are. It has been generally known that the neighbor-based methods could achieve more accurate prediction than the path-based methods could. However, their overall prediction accuracy has not been good enough for practical use (Jaccard, 1912; Salton and McGill, 1983; Sorensen, 1948). Moreover, these methods would have difficulties in providing more detailed information such as which node plays a key role for the neighboring nodes.

We developed and evaluated a network link prediction algorithm based on the object-object match prediction method. The typical examples of object-object match prediction are marriage, protein-protein interaction, and diagnosing a patient with a set of symptoms. In our model, we converted the link prediction problem to an object-object match problem and the network nodes are treated as objects, and the neighboring nodes are considered as the object's associated features. We believe the proposed model is more comprehensive and complete than other conventional link prediction methods in that it considers the effect of the neighbor's neighbor nodes as well as the neighbor nodes in link prediction. Also our approach can be extended for any network link prediction problems.

We evaluated our model by using a well-known open network called the Enron e-mail network (Klimmt and Yang, 2004). The result showed that the proposed model achieved higher accuracy than other conventional methods.

RELATED WORK

The conventional link prediction methods can be categorized into two major streams by the data they use: 1) neighbor-based methods and 2) path distance-based methods (see figure 1). The path distance-based estimations assume that the shorter the path distance between two nodes is, the greater the probability of the two nodes to be linked together. For example, Katz's index can be calculated by giving heavier weight to the shorter paths and by summing up every path between two nodes. The weights exponentially decrease as the path length increases (Liben-Nowell and Kleinberg, 2007). In the random work or hitting time approaches, the link connectivity is computed by the average number of steps from one node to another and also in the opposite direction (Liben-Nowell and Kleinberg, 2007; Linyuan and Zhou, 2011). Sometimes, the Hidden Markov Model was incorporated into these approaches by modeling a node as a state and an edge as a state transition probability (Sarukkai, 2000; Zhu, Hong and Hughes, 2002).

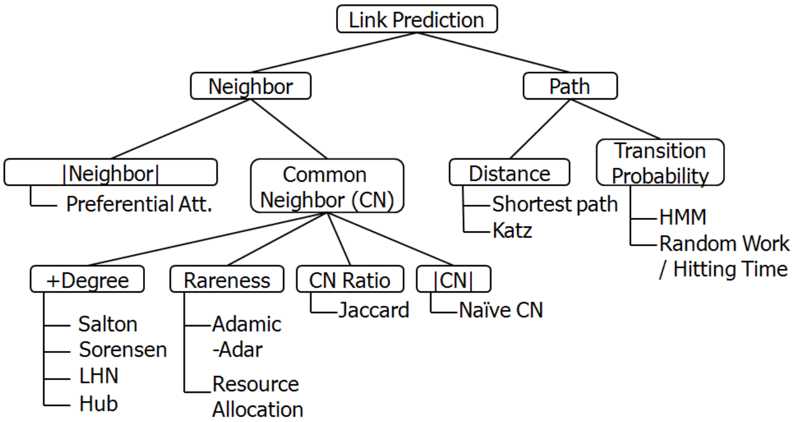

Figure 1. Link prediction methods

The neighbor-based methods assume that the more neighbors shared by any two nodes, the greater probability of the nodes to be linked together. A number of researches devised prediction formula based on the number of common neighbors (Newman, 2001 and 2003), which can be measured by using the ratio of common neighbors to all neighbors (Jaccard, 1912). Some researchers incorporated the degree (the number of edges incident to) of nodes as well as the common neighbors into their formula (Salton and McGill, 1983; Sorensen, 1948; Leicht, Holme and Newman, 2006). Among the neighbor-based approaches, some studies showed that the methods considering the rareness of common neighbors outperformed other methods (Adamic and Adar2003; Ou, et al., 2007)**.** After the scale-free network analysis was introduced by Barabasi (Barabasi and Albert, 1999), several attempts have been made to incorporate the properties of the hub node into link prediction. For instance, Ravasz assumed that nodes with many edges are more likely to have new edges (Linyuan and Zhou, 2011; Ravasz, et al., 2002). Generally, the neighbor-based link prediction methods achieved about twice more accurate prediction than the path distance-based approaches did (Adamic and Adar2003).

The scoring functions of the neighbor-based methods are summarized in Table I. CN denotes the common neighbor(s), and $\Gamma(x)$ represents a set of neighbors of a node x. The link connectivity between node x and y is decided by $score(x, y)$. As shown in Table 1, most neighbor-based link prediction methods assign a relatively high weight to common neighbors, and give a penalty to the nodes connected with many neighbors. Adamic-Adar and resource allocation methods are known to have achieved the best performance among them. This implies that in the link prediction of two nodes, considering the degree of common neighbors is more effective than counting the degree of the two nodes. Despite the long research history of the methods, only a little progress has been made because most of the methods focused mainly on the analysis of neighbor nodes and relatively little attention was paid to the effect of neighbor's neighbor nodes.

Methods	Score Function	Remarks
Naïve common neighbors	$score(N_1, N_2) = \vert CN \vert.$	Fit to social network
Jaccard coefficient	$score(N_1, N_2) = \frac{\vert CN \vert}{\vert \Gamma(N_1) \bigcup \Gamma(N_2) \vert}.$	Ratio of *CN*
Preferential attachment	$score(N_1, N_2) = \vert \Gamma(N_1) \vert \times \vert \Gamma(N_2) \vert.$	Scale-free network
Salton	$score(N_1, N_2) = \frac{\vert CN \vert}{\sqrt{\vert \Gamma(N_1) \vert \times \vert \Gamma(N_2) \vert}}.$	Opposite to preferential attachment
Sorensen	$score(N_1, N_2) = \frac{2 \times \vert CN \vert}{\vert \Gamma(N_1) \vert + \vert \Gamma(N_2) \vert}.$	
Leicht-Holme-Newman	$score(N_1, N_2) = \frac{\vert CN \vert}{\vert \Gamma(N_1) \vert \times \vert \Gamma(N_2) \vert}.$	
Hub related	$score(N_1, N_2) = \frac{\vert CN \vert}{(\max \vert \min)\{\vert \Gamma(N_1) \vert, \vert \Gamma(N_2) \vert\}}.$	
Adamic-Adar & Resource allocation	$score(N_1, N_2) = \sum_{x \in CN} \frac{1}{\log \vert \Gamma(x) \vert}.$ $score(N_1, N_2) = \sum_{x \in CN} \frac{1}{\vert \Gamma(x) \vert}.$	Best performance, sum of the inversed degree of *CN*

Table 1. Neighbor-based link prediction methods. Most of the methods used similar factors for link prediction: 1) the number of common neighbors and 2) the number of neighbors of the two nodes to be predicted.

LINK PREDICTION AND OBJECT-OBJECT MATCH

What makes our method distinguished from conventional network link prediction methods the most is that our method finds tightly coupled groups in neighbor nodes. If the neighbor nodes of two target nodes are tightly coupled, then we assume that the target nodes will have a higher probability to be linked together. To find tightly coupled groups in the neighbor nodes, we analyze the topology of both the neighbor and their neighboring nodes.

To solve the link prediction problem more systematically, we develop a machine learning framework to measure the similarity between objects associated with features. In the framework, we extract matched feature pairs out of matched objects and then use the information to predict a match for objects. We treat the feature set as a basic unit of the feature pair so that we can accommodate the group of the neighbor nodes into the framework.

Figure 2 illustrates a network where the link between node N_1 and N_2 is to be predicted. The shaded area is the boundary of the neighbors, and we find or identify the tightly coupled nodes in the boundary by computing the degree of coupling and cohesion of nodes. Unlike the conventional neighbor-based methods, which usually ignore the effect of nodes not shared by both N_1 and N_2, our method includes N_3 and N_4 as well as CN_1 and CN_2 in the neighbors for the analysis.

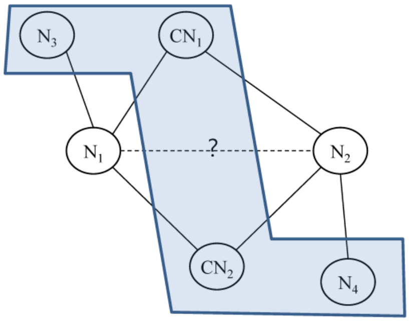

Figure 2. Group detection among the neighbors. We supposed that the link connectivity between two nodes, N1 and N2 can be measured by the probability of the neighbors to form a group. While the conventional neighbor-based approaches only utilized the common neighbors, CN1 and CN2, in our approach, the inter-relationship is measured not only between common neighbors but also between non-overlapping neighbors.

To transform the link prediction problem into the object-object match prediction problem, we propose a new data structure (see figure 3). Each node is represented with a set of its neighbors. And the connected nodes are expressed with a pair of nodes such as <A, B>. Since we only consider an undirected edge, the node pair <A, B> and <B, A> are treated as equivalent. Once a network is represented by the data structure, the link prediction problem is identical to the object-object match prediction problem. The nodes of a network are regarded as objects, and neighbor nodes are regarded as the object's associated features. If two arbitrary objects are predicted to have a match or an interaction, then we predict that the corresponding nodes should be or will be linked together.

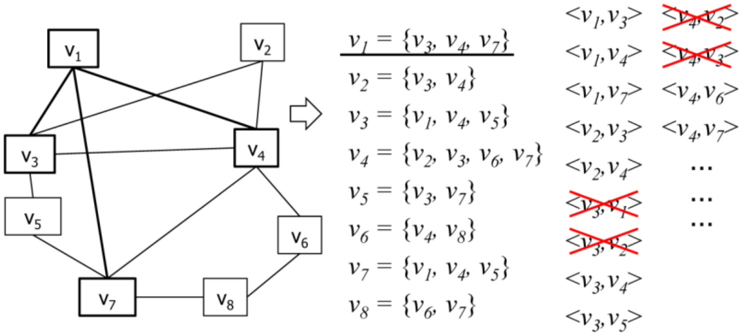

Figure 3. Data structure preparation from a network for the object-object match prediction. Each node can be considered as a set of neighboring nodes, and we can extract a pair of nodes connected to each other. Note that the pair <A, B> and <B, A> are identical because we assumed the edge is undirected. So <v_3, v_1>, <v_3, v_2>, <v_4, v_2>, and <v_4, v_3> were crossed-off.

EVALUATION

To evaluate our model, we compared the prediction accuracy of the proposed method with the existing link prediction methods by using three network datasets: a scientific co-authorship network (CORA) (McCallum, 2000), a communication network (Enron) (Klimmt and Yang, 2004), and a product co-purchasing network (Amazon) (Leskovec, Adamic and Adamic, 2007). Its details are described in Table 2. The source of the dataset, the total number of nodes and edges, average degree of nodes, and the ratio of the reserved test data size to the learning data size are summarized in the table. CORA is a network of authors who wrote papers together. In the network, nodes represent authors, and edges represent collaboration relationships among the authors. CORA has no direction, and among the three datasets, CORA had the highest average degree of nodes. As a communication network, we used the e-mail exchange logs from Enron. The dataset contained communications among the workers only when both the sender and receiver had Enron e-mail addresses. The product co-purchasing network dataset was built on the product sales information in Amazon.com. If one product is frequently co-purchased with another product, the directed edge is connected from one product to a co-purchased product. In order to prepare an undirected network, we selected node pairs with only bi-directional connections.

Source	Type (original)	Nodes	Edges	Avg. Degree	Learning/Test set
CORA	Undirected	15,496	23,006	2.97	19,640 / 3,366
ENRON Intra-mail	Undirected	16,566	18,623	2.25	15,486 / 3,137
Amazon	Undirected	31,056	30,000	1.93	27,620 / 2,380

Table 2. Summary of co-authorship network (CORA)

	0	1	2	3	Total
CORA	5,387 (13.3%)	20,604 (51.0%)	10,102 (25.0%)	4,309 (10.7%)	40,402 (100.0%)
ENRON Intra-mail	4,024 (16.2%)	15,509 (62.2%)	3,920 (15.7%)	1,468 (5.9%)	24,921 (100.0%)
Amazon	3,464 (11.7%)	18,624 (62.7%)	3,411 (11.5%)	4,183 (14.1%)	29,682 (100.0%)

Table 3. The number of neighbor pairs according to the shortest path distance

To make computation more practical, vertices having eleven or more neighbors were removed because using the vertices with more than eleven neighbors doesn't add much value considering the computational cost; that is, we considered the vertices only when the number of total neighbors was less than or equal to 10 for predicting the link of two nodes. Since we filtered all edges incoming from or outgoing to high-degree nodes, there might have been some loss of paths. We selected test pairs only when at least one link of the pairs was observed in the learning set. A negative test set was generated from the neighbors of connected nodes, and we removed node pairs observed in the positive test set from the negative test set.

Also to show clearly the difference between the proposed method and conventional methods, we devise a simplified formula to simulate the proposed method using conventional methods. The formula is developed using the existing link predictors; the Jaccard coefficient (Jaccard, 1912) and the Katz predictor (Katz, 1953). The Jaccard coefficient is a formula developed with a similar concept to the cohesion, and the Katz predictor is a formula developed with a similar concept to the coupling. The Jaccard coefficient is the ratio of the common neighbors to all the neighbors of the two nodes.

We compared the proposed method with naïve common neighbor, Jaccard coefficient, Adamic-Adar, preferential attachment, and the simplified version of the proposed method. The accuracy of the methods was depicted using receiver-operator characteristic (ROC) curves (figure 4). In the figure, the accuracy is measured by the area either under or above the ROC curve. The closer the ROC curve of a method is to the base line (i.e. its area size is 0.5), the worse the method is. The result showed that the proposed method outperformed other methods.

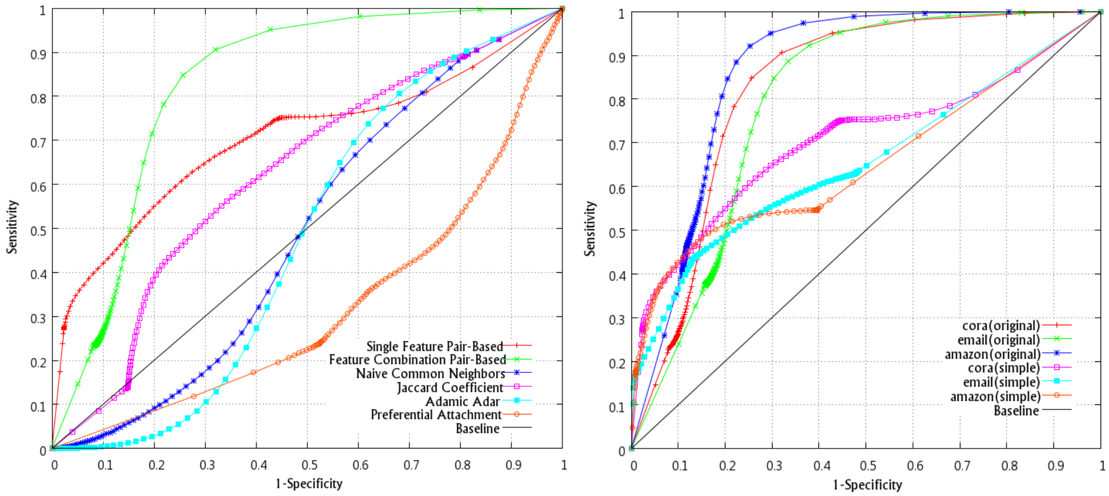

Figure 4. The ROC curve of scientific co-authorship network (CORA) and the ROC curve of the proposed method.

The results strongly support that considering the effect of the neighbor's neighbors improves the overall accuracy in link prediction. The improvement was more apparent when the target nodes were connected by paths with relatively longer path distance.

CONCLUSION

In this paper, we proposed and evaluated a network link prediction model based on an object-object match prediction method and a machine learning framework to predict missing links in a network. The results showed that the proposed method is better than other conventional methods and considering the neighbor's neighbors is critical to improve the link prediction accuracy.

The proposed method produces good prediction results in the networks where the paths between nodes are long regardless of the average degree of the nodes. In addition, using the neighbor's neighbor effect is highly believed to improve prediction accuracy. To better understand the effect of the path distance and of the degree of a node, it is needed to perform additional experiments with the conventional methods. If similar prediction accuracy improvements are observed with the conventional methods with the addition of the neighbor's neighbor effect, then it could be concluded that the addition is an essential step for link prediction.

REFERENCES

1. Liben-Nowell, D. and Kleinberg, J. (2007) The link prediction problem for social networks, *J. of. the Amercan Society for Information Science and Technology*, 58, 7, 1019-1031.
2. Adamic, L. and Adar, E. (2003) Friends and neighbors on the web, *Soc. Networks*, 25, 3.

3. Dombroski, M. J. and Carley, K. M. (2002) NETEST: Estimating a Terrorist Network's Structure, *Computational & Mathematical Organization Theory*, 8, 235-241.
4. Xu, J. J. and Chen, H. (2005) CrimeNet Explorer: A Framework for Criminal Network Knowledge discovery, *ACM Transactions on Information Systems*, 23, 2, 201-226.
5. Katz, L. (1953) A new status index derived from sociometric analysis, *Psychometrika*, 18, 1, 39-43.
6. Linyuan L. and Zhou, T (2011) Link prediction in complex networks: a survey, *Physica A*, 390, 6, 1150-1170.
7. Sarukkai, R. R. (2000) Link prediction and path analysis using Markov chains, *Computer Networks*, 33, 1-6, 377-386.
8. Zhu, J., Hong, J. and Hughes, J. G. (2002) Using Markov chains for link prediction in adaptive web sites, *Lecture Notes in Computer Science*, 2311/2002, 55-66.
9. Newman, M. (2001) Clustering and preferential attachment in growing networks, *Phys. Rev. E.*, 64(025102).
10. Newman, M. (2003) The structure and function of complex networks, *SIAM review*, 45, 167-256.
11. Jaccard, P. (1912) The distribution of the flora in the alpine zone, *New Phytologist*, 11, 2, 37-50.
12. Salton, G. and McGill, M. (1983) Introduction to modern information retrieval, McGraw-Hill.
13. Sorensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skr., 5, 1-34.
14. McCallum, et al. (2000) Automating the construction of internet portals with machine learning, *Information Retrieval*, 3, 2, 127-163.
15. Klimmt, B. and Yang, Y. (2004) Introducing the Enron corpus, *CEAS conference*.
16. Leskovec, J., Adamic, L. and Adamic, B. (2007) The dynamics of viral marketing, *ACM Tran.on the Web (ACM TWEB)*, 1, 1.
17. Leicht, E. A., Holme, P. and Newman, M. (2006) Vertex similarity in networks, *Phys. Rev.*, E 73, 026120.
18. Ou, et al. (2007) Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, *Phy. Rev.*, E 75, 021102.
19. Barabasi, A. L. and Albert, R. (1999) Emergence of scaling in random networks, *Science*, 286, 5439, 509-512.
20. Ravasz, et al. (2002) Hierarchical organization of modularity in metabolic network, *Science*, 297, 5586, 1551-1555.