A Link Prediction Approach for Drug Recommendation in Disease-Drug Bipartite Network

Esra Gündoğan[1] and Buket Kaya[2]
[1]Department of Computer Engineering, Fırat University, Elazığ, Turkey
[2] Elazığ EOSB Maden Vocational School, Fırat University, Elazığ, Turkey
{egundogan, bkaya}@firat.edu.tr

***Abstract*— Social networks we have encountered in different areas and in different forms have a dynamic structure because the relationships they define constantly change. Link prediction is an important and effective solution to understand this dynamic nature of networks and to identify future relations. It estimates of possible future connections between nodes in the network taking advantage of network's current state. In this study, a method for link prediction in the disease-drug network is proposed. Sofar, the most of studies done is usually based on connection prediction in single mode networks. This method has been applied on a bipartite such as disease-drug network, as apart from single mode networks. To compare performance of the proposed method, four of similarity based link prediction methods has been also applied to the network. The results obtained from experiments show that the proposed method has a good percentage of success than the other similarity based link prediction methods.**

Keywords*— *Link prediction; Recommendation systems; Social network analysis;

I. INTRODUCTION

Social networks are one of the structures used to model relationships between the objects, individuals, organizations and communities in our environment [1]. Many complex systems that we face in different areas and in different ways can be visualized with social networks. Social network is a useful tool for understanding the structure, development, and relationships of complex systems. Social network analysis provides valuable information about complex systems. This information can be used for many purposes.

Link prediction from research subjects related to the analysis of the future state of the network is the problem of predicting a link between two nodes by the characteristics of the nodes in the network and other observed connections [2]. The analysis of the future state of a social network and the determination of the changes that may occur in the network are made by using the past states of the network.

Over time, the relationships among individuals, communities, and organizations change. Social networks that model these relationships in the real world constantly change and evolve due to these changes. In the dynamic nature of social networks, new links, nodes can be added to the network, or existing links, nodes may disappear from the network. Link prediction tries to determine dynamic relationships between objects in a social network. Link prediction has many important applications in different areas. It can be used to find co-authors in academic social networks [3,4], recommend product to customers in online shopping [5], find protein-protein interactions, estimate future citations [6], determine and categorize people's activities [7], determine relationships between diseases [8-10].

The first goal of this study is to create disease-drug network by using data that is about which drugs can be written for which diseases that are obtained from www.drugs.com [11]. After this network is constructed, it is discovered that a given drug can be used for which diseases with link prediction. Internal links method is proposed to predict link in the disease-drug network. To compare performance of the proposed method, four of similarity based link prediction methods are also applied to the network. The results of internal links method are compared with those obtained from other similarity based link prediction methods. The results of experiments show that the proposed method gives better results in terms of precision, recall and F-measure criteria.

The rest of the paper is organized as follows. Section II defines the link prediction problem. Section III gives the related work. Section IV presents the constructed bipartite network and the proposed method. Section V shows experimental results and finally Section VI includes the conclusions.

II. LINK PREDICTION PROBLEM

The link prediction problem in social networks makes predictions about the future structure of the network. The network is defined as a graph. The data in the network are represented with nodes, and the relations are represented with links. The future of non-connected links between node pairs is predicted. A score is calculated for each pair of nodes that are likely to be connected. The higher the score calculated between a pair of nodes, the higher the probability of a future link between that pair of nodes. Node pairs are sorted from big to small according to their score. The links between the pairs at the top are the most likely future links. The accuracy of predictions that are made up changes by depending on algorithms used.

Bipartite social networks are a special form of social networks. In bipartite networks, nodes are located in two different clusters. Links are only among nodes in different clusters. There are no links between nodes in the same cluster.

978-1-5386-1880-6/17/$31.00 ©2017 IEEE

Many social networks in our environment have bipartite network structure [12-15]. Most link prediction methods are for single mode networks. Therefore, traditional link prediction methods cannot be directly applied to bipartite networks. For link prediction in bipartite social networks, such networks are often transformed into a single mode social network.

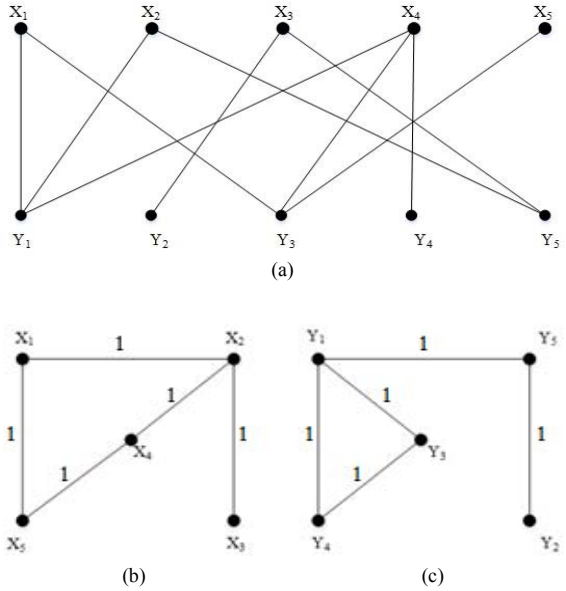

Fig. 1. (a) An example of bipartite graph, (b) X-projection, (c) Y-projection

In the given bipartite graph representation in Fig. 1(a), single mode graph where only *X* nodes are contained is expressed as X-projection while single mode graph where only *Y* nodes are contained is expressed as Y-projection. If any two *X* nodes in the bipartite graph establish neighborhood relation with same *Y* node, the connection is established between the two nodes chosen in single-node X-projection. For example; since X_1 and X_2 nodes are in connection with Y_1 node in Fig. 1(a). Therefore, a link between X_1 and X_2 nodes is constructed in X-projection shown in Fig. 1(b) [16].

The easiest way to convert bipartite social networks into single mode networks is unweighted networks created without considering the frequency of collaboration. However, because single mode networks contain less useful information than bipartite social networks, bipartite social network shown in Fig. 1 is transformed into weighted single mode network. The easiest way to give weight to links is to determine the number of relations [17]. For example; in Fig. 1, the weight between X_1 and X_2 nodes in X-projection has calculated as 1. The reason of this is that there are one common nodes Y_1 and Y_2 connected with X_1 and X_2 in bipartite social network. In this way weight finding is called weight function [18] and is calculated with Eq. 1:

$$\omega(u, v) = \left|N(u) \cap N(v)\right| \tag{1}$$

III. Related Work

Many approaches have been proposed for link prediction until today. These approaches are generally based on similarity [19]. The basic idea of similarity-based approaches is that the greater the intersection of the neighbors of the *x* and *y* nodes, the higher the likelihood that they will connect in the future. *x* and *y* are pair of nodes, $\Gamma(x)$ and $\Gamma(y)$ are cluster of neighbors of *x* and *y* nodes in the network. Four of the best-known similarity-based link estimation algorithms are as follows.

- **Common Neighbors (CN):** It is a measure based on the assumption that connected of two nodes in the future are proportional to the number of their neighbors [20]. The more common nodes two nodes have, the higher the probability of a link being formed between them. It is one of the most widely used measure in link prediction problem due to its simplicity. Mathematical equation is given below.

$$CN(x, y) = \left|\Gamma(x) \cap \Gamma(y)\right| \tag{2}$$

- **Jaccard Coefficient (JC):** The Jaccard coefficient is normalized form of common neighborhood measure [21]. It demonstrates the probability that a selection made randomly from the set of neighbors of the pair of *x* and *y* nodes is one of the common neighbors of the pair of *x* and *y* nodes. The greater the number of common neighbors, the greater this measure. Equation is given below.

$$JC(x, y) = \frac{\left|\Gamma(x) \cap \Gamma(y)\right|}{\left|\Gamma(x) \cup \Gamma(y)\right|} \tag{3}$$

- **Adamic/Adar (AA):** This measure was originally proposed by Adamic and Adar to calculate similarity between two web pages [22]. Later, it is customized by selecting common neighbors as features for link prediction [23]. The new measure is calculated based on the weight of the partner neighbors. The weight ratios of the rare ones are increased. Its definition is as follows.

$$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log\left|\Gamma(z)\right|} \tag{4}$$

- **Preferential Attachment (PA):** The probability that one of the end points of a new connection to be formed in the network is *x* is proportional to the number of neighbors of *x* node. In other words, nodes with many neighbors are more likely to create new connections. According to Newman [18], the probability of co-operation between *x* and *y* in a cooperation network is proportional to the number of cooperations of *x* and *y*. The mathematical equation is given below.

$$PA(x, y) = \left|\Gamma(x)\right| \cdot \left|\Gamma(y)\right| \tag{5}$$

IV. The Proposed Method

In this study, the data required to construct the drug-disease network was obtained from [11]. This site includes information such as which drugs are written for diseases, the percentage of users who prefer these drugs, class of drugs, category of pregnancy. The nodes on the drug-disease network represent drugs and diseases. Connection has created between the three most popular drugs for a disease. Given that percentage of other drugs used is lower, these links are not included. Table I shows a certain section of the data received from the site.

TABLE I. A Section From the Used Data

Disease	Drug	Rate
Asthma	Flunisolide	10
	Levalbuterol	10
	Xopenex	9
Bronchiectasis	Doryx	9
	Doxycycline	7
	Ventolin HFA	8
Gastritis/Duodenitis	Omeprazole	6
	Pantoprazole	5
Sinusitis	Augmentin XR	10
	Deconamine SR	8
	Cedax	9
Migraine	Imitrex	8
	Sumatriptan	8
	Amerge	9

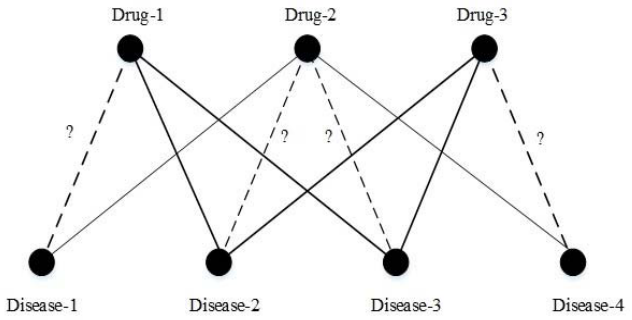

Fig. 2. Example of drug-disease network

A bipartite network whose node clusters are diseases and drugs was established according to the disease-drug relationships in Table 1. Links have been constructed between diseases and first three drugs with the highest rate used for each disease. The representation of such a network is given in Fig. 2. The purpose of this study is to predict which drugs can be written to diseases using existing relationships.

Unlike traditional methods, several methods have been proposed to predict possible links in a bipartite network [24-26]. In this work, internal links method [24] has been proposed for link prediction in the constructed disease-drug bipartite network. In this method, a private kind of links called internal links are defined and link prediction is made based on these links. After bipartite network has been converted to a weighted single-mode network, links that may be internal links in the network are identified. In a bipartite network, a node pair with no connection is selected. If there is no change in the projection graph when a link is created between this node pair, the link between the selected nodes is called as internal link [24].

Links that have no connection in bipartite network and do not change projection of bipartite network when they have a connection are found. In this work, all links found are not considered internal links. A threshold value is determined to ensure that important links are not lost and unnecessary links are not taken into consideration. Links whose weight is greater than this threshold value are considered internal links.

V. Experimental Results

In this section, the proposed method has been evaluated on the constructed drug-disease network. The data at the network were obtained from www.drugs.com web site. In this network, the number of edges is 500. Randomly selected 100 links between diseases and drugs are removed to test our approach. Thus, a training network with 400 links was constructed. Then, the proposed method was applied to disease-drug bipartite network. Links that could be an internal link between nodes that are not connected in the network were found. The links with higher than the determined threshold value was identified as internal links. The found internal links were sorted in decreasing order with respect to their weight. The top-100 high weighted links were selected. Four of the similarity-based link prediction methods were applied to disease-drug bipartite network to evaluate the prediction success of the proposed method. A *score(x,y)* value is calculated for each node pair that is not connected in the training network. As mentioned before, this value measures the similarity between node x and y. A larger score means a higher probability that the related link missing at current time will appear in the future. The *score(x,y)* function was calculated respectively for *Common Neighbors, Jaccard Coefficient, Preferential Attachment* and *Adamic/Adar* methods. All the predicted links were sorted in decreasing order with respect to their scores. The top-100 high scored links are selected. The predicted links with the proposed method *Internal Links(IL), Common Neighbors(CN), Jaccard Coefficient(JC), Preferential Attachment(PA)* and *Adamic / Adar(AA)* methods were compared with the extracted links. The results of comparison were used for computing three performance evaluation criteria precision, recall and F-measure. Fig. 3 shows the performance comparison of five different methods in terms of precision, recall and F-measure. The proposed method has the highest values for the three performance measures. The removed links are obtained at a higher rate than the used similarity-based methods by the proposed method. According to the results obtained, the proposed method with high prediction success can be used to recommend which drugs can be used for treatment of a disease.

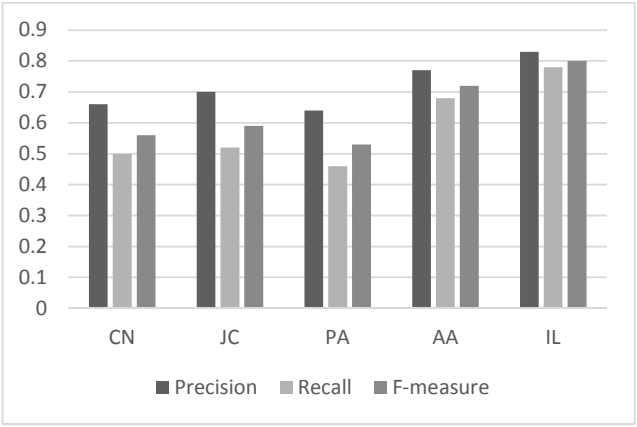

Fig. 3. Performance comparison of five different methods

VI. CONCLUSION

Today, link prediction in bipartite networks has become a focal point for many science areas. Researchers from different fields have increased their work on this subject. In this study, link prediction in bipartite networks taking attention in recent years was performed on constructed disease-drug network. Firstly, a drug-disease network was created with data drawn from www.drugs.com site. This network is a bipartite graph. Link prediction in bipartite networks is a more difficult and complex task than link prediction in single mode networks. Link prediction methods for single mode networks cannot be applied directly to bipartite networks. Therefore, bipartite network has been converted to single mode network for implementation of link prediction methods. The internal links method, which achieves high performance success in the studies used, has been applied on the constructed disease-drug network. Also, Common Neighbor, Jaccard Coefficient, Preferential Attachment and Adamic / Adar similarity-based link prediction methods have been applied to bipartite network. When success criteria of the five different approaches are considered, the predictability of that which drug can be used in treatment of which disease has been achieved with a high level of accuracy.

ACKNOWLEDGMENT

This study was partially supported Fırat University, Scientific Research Projects Office under Grant No MF.17.04.

REFERENCES

[1] Wasserman, S. and Faust, K., 1994. Social Network Analysis: Methods and Applications, Cambridge University Press, England.

[2] Liben Nowell, D., and Kleinberg, J., 2007. The link prediction problem for social networks. journal of the Association for Information Science and Technology, 58(7), 1019-1031.

[3] Pavlov, M., and Ichise, R., 2007. Finding experts by link prediction in co-authorship networks. In Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics-Volume 290 (pp. 42-55). CEUR-WS. org.

[4] Wohlfarth, T., and Ichise, R., 2008. Semantic and Event-Based Approach for Link Prediction. In PAKM (pp. 50-61).

[5] Akcora, C. G., Carminati, B., and Ferrari, E., 2011. Network and profile based measures for user similarities on social networks. In Information Reuse and Integration (IRI), 2011 IEEE International Conference on (pp. 292-298). IEEE.

[6] Jawed, M., Kaya, M., and Alhajj, R., "Time frame based link prediction in directed citation networks." Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference on. IEEE, 2015.

[7] Sid Ahmed, H., Mohamed Faouzi, B., and Caelen, J., 2013. Detection and classification of the behavior of people in an intelligent building by camera. International Journal on Smart Sensing & Intelligent Systems, 6(4).

[8] Gül, S., Kaya, M. and Kaya, B. 2016. Predicting links in weighted disease networks. In Computer and Information Sciences (ICCOINS), 2016 3rd International Conference on (pp. 77-81). IEEE.

[9] Kaya, B., and Poyraz, M., Age-series based link prediction in evolving disease networks. Computers in biology and medicine, 63, 1-10, 2015.

[10] Kaya, B. and Poyraz, M., 2015. Finding relations between diseases by age-series based supervised link prediction, 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1097-1103, Paris.

[11] https://www.drugs.com/medical_conditions.html.

[12] Savic, M., Ivanovic, M., Radovanovic, M., Ognjanovic, Z., Pejovic, A., and Krueger, T. J., 2014. The structure and evolution of scientific collaboration in Serbian mathematical journals, Scientometrics, 101(3), 1805-1830.

[13] Gündoğan, E. and Kaya, B., 2017. A Recommendation Method Based on Link Prediction in Drug-Disease Bipartite Network, IEEE 2nd International Conference on Advanced Information and Communication Technologies, 4-7 July, Lviv, Ukraine.

[14] Ferreri, L., Venturino, E., and Giacobini, M., 2011. Do diseases spreading on bipartite networks have some evolutionary advantage? In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (pp. 141-146). Springer, Berlin, Heidelberg.

[15] Shahinpour, S., and Butenko, S., 2013. Distance-based clique relaxations in networks: s-clique and s-club. In Models, algorithms, and technologies for network analysis (pp. 149-174). Springer, New York, NY.

[16] Zhou, T., Ren,J., Medo, M., Zhang, Y., 2007. "Bipartite Network Projection and Personal Recommendation", Physical Review, pp:1-7.

[17] Ramasco, J.J., Morris, S.A., 2006. "Social Inertia and Diversity in Collaboration Networks", Physical Review, pp: 47-50.

[18] Newman, M.E.J., 2001. Clustering and preferential attachment in growing networks, Phys. Rev. E 64, 025102.

[19] Xiang, E.W., 2008. A Survey on Link Prediction Models for Networked Data, PhD Qualifying Exam, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology.

[20] Kossinets, G., 2006. Effects of missing data in social networks. Social networks, 28(3), 247-268.

[21] Tan, P.-N. and Kumar, V., 2005. Introduction to Data Mining, Addison Wesley

[22] Adamic, L. A., and Adar, E., 2003. Friends and neighbors on the web. Social networks, 25(3), 211-230.

[23] Liben Nowell, D., and Kleinberg, J., 2007. The link prediction problem for social networks. journal of the Association for Information Science and Technology, 58(7), 1019-1031.

[24] Allali, O., Magnien, C., and Latapy, M., 2011. Link prediction in bipartite graphs using internal links and weighted projection. In Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on(pp. 936-941). IEEE.

[25] Benchettara, N., Kanawati, R., and Rouveirol, C., 2010. Supervised machine learning applied to link prediction in bipartite social networks. In Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on (pp. 326-330). IEEE.

[26] Xia, S., Dai, B., Lim, E. P., Zhang, Y., and Xing, C., 2012. Link prediction for bipartite social networks: the Role of Structural Holes. In Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on (pp. 153-157). IEEE.