Learning Links in MeSH Co-occurrence Network: Preliminary Results

Andrej Kastrin
Faculty of Information Studies
Novo mesto, Slovenia
Email: andrej.kastrin@guest.arnes.si

Dimitar Hristovski
Institute of Biostatistics and Medical Informatics
Faculty of Medicine, University of Ljubljana
Ljubljana, Slovenia
Email: dimitar.hristovski@gmail.com

***Abstract*—Literature-based discovery (LBD) is focusing on automatically generating scientific hypotheses by uncovering hidden, previously unknown relations between existing knowledge. Co-occurrences between biomedical concepts can be represented by a network that consists of a set of nodes representing concepts and a set of edges representing their relationships. In this work we propose a method for link prediction of implicit connections between Medical Subject Headings (MeSH®) descriptors. Our approach is complementary to standard LBD. Link prediction was performed using Jaccard and Adamic-Adar similarity measures. Preliminary results showed high prediction performance with area under the ROC curve of 0.78 and 0.82 for Jaccard and Adamic-Adar coefficient, respectively.**

***Index Terms*—network analysis, link prediction, literature-based discovery**

I. Introduction

Retrieval and linking different chunks of scientific information into understandable and interpretable knowledge becomes a challenging task. Text mining technologies complement manual information retrieval from biomedical sources [1]. Common text mining tasks in biomedicine include the recognition of explicit facts from the literature, document summarization, question answering and literature-based discovery (LBD).

LBD is a methodology for automatically generating hypotheses for scientific research by uncovering hidden, previously unknown relationships between existing knowledge [2]. The LBD methodology was pioneered by Swanson [3], who proposed that dietary fish oils might be used to treat Raynauds disease because they lower blood viscosity, reduce platelet aggregation and inhibit vascular reactivity. The basic assumption of Swansons approach is that there exists two scientific domains that do not communicate. A segment of knowledge in one domain may be related to knowledge in the other domain, but this relationship is unknown. The methodology of LBD relies on the idea of concepts relevant to three literature domains: X, Y, and Z. For example, suppose a researcher has found relationship between disease X and a gene Y. Further suppose that a separate researcher has studied the effects of substance Z on gene Y. The use of LBD may suggest an XZ relationship, indicating that substance Z may potentially treat disease X.

Associations between literature entities based on co-occurrence of biomedical terms, such as diseases or genes constitute an important part of knowledge representation. A co-occurrence approach is built on the assumption that biomedical concepts occurring together in the same title or abstract are in some way biologically related [4], [5]. Biomedical knowledge can be thus viewed as a set of concepts along with the relations between them. Interactions between concepts can be described in terms of a graph, consisting of nodes and edges, where the former represent concepts and the latter represent their relationships. Knowledge network is not static. It is a dynamic structure that evolves over time either by addition of new nodes or by new links that form between nodes.

Link prediction is a newly emerging research field that is at the intersection of the network analysis and machine learning. Understanding the mechanisms of link formation in complex networks is a long standing challenge for network analysis. Link prediction refers to the discovery of future links between nodes that are not directly connected in the current snapshot of a given network [6]. Seen in this way, the link prediction problem is similar to LBD. In the literature several link prediction techniques have been proposed. These techniques can be used to predict new link formation by estimating the likelihood of link formation between two nodes on the basis of the observed network topology.

In this work we propose a method for link prediction in biomedical domain, i.e. for prediction and evaluation of implicit or previously unknown connections between biomedical concepts. Our approach is complementary to the traditional LBD. To evaluate the link prediction techniques for LBD, here we investigate the performance of link prediction techniques for networks obtained from Medical Subject Headings (MeSH®) [7] co-occurrence data.

II. Methods

A. Basic Terminology

A network is represented by a graph $G(V, E)$ that consists of a set of nodes V representing concepts and a set of edges E representing relationships between the nodes [8]. The number of edges of a node i is denoted by its degree k_i.

The link prediction problem can be formally represented as follows. Suppose we have a network $G[t_1, t_2]$ which contains all interactions among nodes that take place in the time interval $[t_1, t_2]$. Further suppose that $[t_3, t_4]$ is a time interval occurring after $[t_1, t_2]$. The task of link prediction is to provide a list of edges that are present in $G[t_3, t_4]$ but absent in $G[t_1, t_2]$. We

 ISBN: 978-1-61208-334-6

refer to $G[t_1, t_2]$ as the train network and $G[t_3, t_4]$ as the test network (Figure 1). In this work the prediction was performed on a core subnetwork which consists of nodes which have at least $k = 3$ neighbors.

B. Data Collection and Network Construction

The train and test networks were constructed from the Unified Medical Language System (UMLS®) [9] co-occurrence table (MRCOC). MRCOC table includes statistical aggregations of co-occurrences of biomedical concepts in different data sources. Two overall frequencies of MEDLINE® co-occurrence are provided: one for recent MEDLINE data (MED) and one for MEDLINE data from a preceding block of years (MBD). MRCOC provides the frequency of co-occurrence of two concepts in the same indexed articles from MEDLINE (i.e., the number of articles discussing both concepts during a given time period). We select only those co-occurrence pairs that refer to MeSH descriptors.

The constructed networks were post-processed to remove all non-useful edges. We applied the Pearsons chi-square (χ^2) test for independence for each co-occurrence pair to obtain a statistic, which indicates whether a particular pair of MeSH descriptors occurs together more often than by chance [10]. If χ^2 is greater than the critical value of 3.84 ($p \leq 0.05$), we can be 95% confident that a particular MeSH relation occurs more often than by chance.

C. Experimental Setup

The link prediction framework we use follows the procedure first introduced by Liben-Nowell and Kleinberg [11]. We perform link prediction using proximity measures. Proximity measures are used to find similarity between a pair of nodes. For each node pair (u, v), a link prediction method gives score $s(u, v)$, an estimate of the likelihood of link formation between nodes u and v. Among various proximity measures proposed in the literature we use Jaccard and Adamic-Adar coefficients. Jaccard coefficient measures the probability that a neighbor of u or v is a neighbor of both u and v [12]. Jaccard coefficient simply divides the number of common neighbors by the number of total neighbors. Adamic-Adar coefficient measures neighborhood overlap between nodes u and v, weighting the overlap of smaller neighborhoods more heavily [13].

We examine how accurately we can predict which node pairs will connect between times t_3 and t_4 despite not having any co-occurrences before time t_3. The major challenge in prediction evaluation is the huge number of possible node pairs which can greatly increase computational time. To cope with this issue we use bootstrap resampling approach [14]. We use 100 bootstrap samples. In each bootstrap step we draw a random sample of 1000 nodes and create appropriate train and test subgraphs. Next, we compute the link prediction score $s(u, v)$ for each node pair (u, v) that is not associated with any interaction before time t_3 by using one of the link prediction techniques introduced in the previous paragraph. We assign class label 'positive' to this node pair if the connection occurs in test network and 'negative otherwise. The prediction performance was evaluated using the receiver operating characteristic (ROC) curves. A ROC graph depicts relative tradeoffs between true positives and false positives. As a measure of prediction performance we use the area under the ROC curve (*AUC*). The *AUC* is a widely used performance measure which can be interpreted as the probability that a randomly selected link is given a higher link prediction score than a randomly selected non-existent link. Final *AUC* value was averaged over 100 bootstrap samples.

TABLE I
BASIC TOPOLOGICAL CHARACTERISTICS OF THE MESH NETWORKS.

Parameter	MBD network	MED network
Density	0.01	0.01
Mean degree	274.78	298.05
Average path length	2.23	2.20
Clustering coefficient	0.27	0.26
Small-worldness index	21.57	20.70

III. RESULTS

Before evaluating the effectiveness of link prediction techniques, we describe the characteristics of the used datasets. The MBD network consists of 24,225 nodes and 4,897,380 edges. We filter out all non-useful edges using χ^2 test. After reduction the MBD network contains 3,328,288 edges. The MED network contains 25,570 nodes and 5,615,965 edges. After the filtering step, the number of edges decreased to 3,810,535. The topological properties of both networks are summarized in Table I. Both networks are very similar regarding topological properties. The networks exhibit relatively short average path length between all pairs of nodes. On average there are only about two hops from the selected node to any other node. Both networks exhibit small world property because of small average path length and relatively high clustering.

The classification performance is summarized in Figure 2. Mean *AUC* for Jaccard coefficient was $AUC = 0.78$ with $SD = 0.02$. Mean *AUC* for Adamic-Adar coefficient was $AUC = 0.82$ with $SD = 0.01$. Adamic-Adar coefficient exhibited slightly better performance than Jaccard coefficient.

IV. CONCLUSION AND FUTURE WORK

In this paper, we apply and evaluate link prediction methods on a network based on co-occurrence patterns between MeSH descriptors. We have exploited two methods for link prediction task, namely Jaccard and Adamic-Adar coefficients and demonstrated that link prediction is plausible with high prediction performance. To the best of our knowledge, this is the first work that investigates unsupervised learning for link prediction of literature-derived network in the biomedical domain.

There are many possible directions for future work. One is to consider addition similarity measures, including preferential attachment, Katz measure, SimRank, or cosinus similarity, to

 ISBN: 978-1-61208-334-6

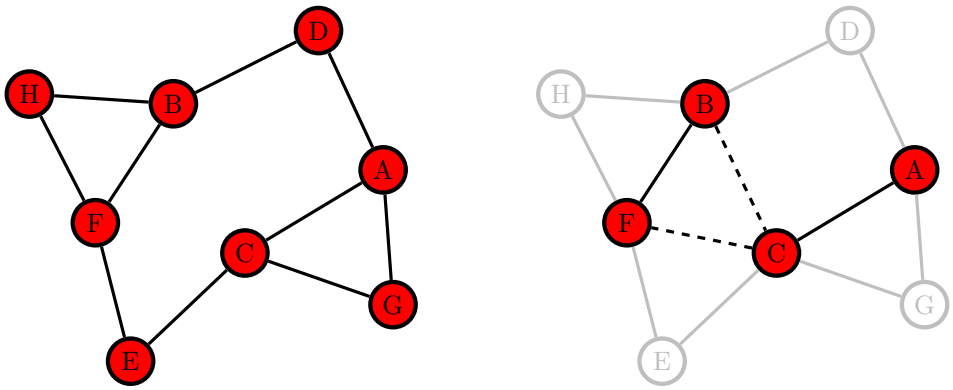

Fig. 1. Train (left) and test (right) network. New link formation is predicted from the topology of a network obtained from co-occurrences in the training period. We investigate the performance of link prediction techniques by comparing the predicted links with actual new links within the testing period. Prediction and evaluation was performed on a core subnetwork which consists of nodes which have at least $k = 3$ neighbors.

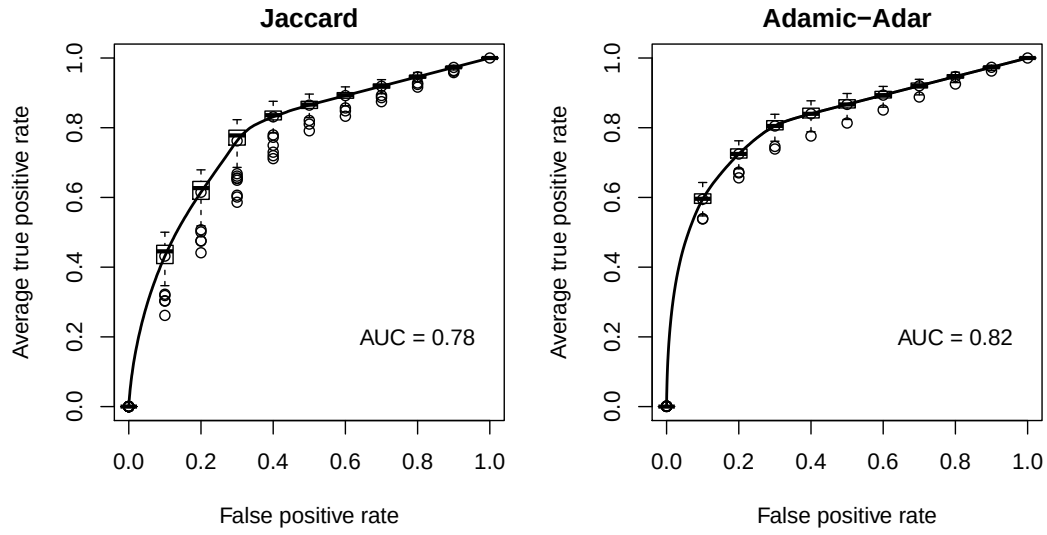

Fig. 2. ROC curves.

name just a few. Second, we should investigate prediction performance of link prediction algorithm on different networks. Biomedical science is full of interesting complex networks; for example we should consider NCBI gene or protein network, KEGG collection, or UniProt database.

Further we should investigate the connection between network attributes and prediction performance. It is well known that nodes which share similar properties tend to create links to each other. For example, persons in a social network who share similar interests are very likely to be friends [15]. In our case we could use attributes such as number of occurrences of particular MeSH descriptor, semantic type from UMLS Semantic Network, etc. We should also systematically analyze the influence of network topology on prediction performance. The process of link creation may be a result of the joint influence of several mechanisms such as small world effect and preferential attachment.

ACKNOWLEDGMENT

This work was supported by Slovenian Research Agency.

REFERENCES

[1] D. Rebholz-Schuhmann, A. Oellrich, and R. Hoehndorf, "Text-mining solutions for biomedical research: Enabling integrative biology," *Nature Reviews. Genetics*, vol. 13, no. 12, pp. 829–839, 2012.

[2] D. Hristovski, T. Rindflesch, and B. Peterlin, "Using literature-based discovery to identify novel therapeutic approaches," *Cardiovascular & Hematological Agents in Medicinal Chemistry*, vol. 11, no. 1, pp. 14–24, 2013.

[3] D. R. Swanson, "Fish oil, Raynaud's syndrome, and undiscovered public knowledge," *Perspectives in Biology and Medicine*, vol. 30, no. 1, pp. 7–18, 1986.

[4] B. Stapley and G. Benoit, "Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts," *Pacific Symposium on Biocomputing*, vol. 5, pp. 526–537, 2000.

[5] R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg, and W. Alkema, "Literature mining for the discovery of hidden connections between drugs, genes and diseases," *PLoS Computational Biology*, vol. 6, no. 9, p. e1000943, 2010.

[6] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

 ISBN: 978-1-61208-334-6

[7] M. H. Coletti and H. L. Bleich, "Medical subject headings used to search the biomedical literature." *Journal of the American Medical Informatics Association : JAMIA*, vol. 8, no. 4, pp. 317–323, 2001.

[8] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[9] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, "The Unified Medical Language System." *Methods of information in medicine*, vol. 32, no. 4, pp. 281–91, 1993.

[10] C. D. Manning and H. Schuetze, *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press, 1999.

[11] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[12] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. New York, NY: Cambridge University Press, 2011.

[13] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Social Networks*, vol. 25, no. 3, pp. 211–230, Jul. 2003.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer, 2011.

[15] Z. Liu, J.-L. He, K. Kapoor, and J. Srivastava, "Correlations between community structure and link formation in complex networks," *PloS ONE*, vol. 8, no. 9, p. e72908, 2013.

 ISBN: 978-1-61208-334-6