2009 Advances in Social Network Analysis and Mining

Predicting Links in Social Networks using Text Mining and SNA

Alon Bartal, Elan Sasson and Gilad Ravid
Industrial and Management Engineering
Ben-Gurion University
Beer-Sheva, 84105 Israel
{bartala, elans, rgilad}@bgu.ac.il

***Abstract*- Lately there is great progress in business organizations perception towards social aspects. Competitive organizations need to create innovation and segregate in the market. Business interactions help reaching those goals but finding the effective interactions is a chalange.**

We propose a prediction method, based on Social Networks Analysis (SNA) and text data mining (TDM), for predicting which nodes in a social network will be linked next. The network which is used to demonstrate the proposed prediction method is composed of academic co-authors who collaborated on writing articles. Without loss of generality, the academic co-authoring network demonstrates the proposed prediction procedure due to its similarity to other networks, such as business co-operation networks. The results show that the best prediction is achieved by incorporating TDM with SNA.

***Keywords*- Social network; Prediction Social network analysis; styling**

I. Introduction

The challenge of predicting changes in a social network is called the link prediction problem [14]. Liben-Nowell and Kleinberg, [8] explain it as: Given a snapshot of a social network at time t, we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t+1.

Social networks are highly dynamic objects which grow and change rapidly over time. The mechanisms by which they evolve is a fundamental question that is still not well understood, thus it forms the motivation for this work. For example, two people who are not connected and are close in a friendships network will have friends in common. Our approach used to solve the link prediction problem in an academic co-authoring network is being addressed by combining TDM methods to evaluate and represent authors' interest topics by extracting key concepts from common articles titles and SNA methods. Two TDM methods are compared (NLP-Natural Language Processing and VSM- vector space model) and the relevance of this knowledge to the prediction accuracy is examined. The goal of this research is to find which measures of the network can lead to the most accurate link predictions and to examine whether TDM methods can contribute to the overall prediction accuracy.

II. Background and Research Model

According to Wasserman and Faust, [18] "a social network consists of a finite set or sets of actors and the relation or relations defined on them". It is a complex system that characterized by a high number of dynamically interconnected entities [3] and connects entities in any type of link that implies a peer-to-peer relationship. Social networks are typically affiliation networks in which participants collaborate in groups and links are established by common group membership [17]. Examples of a relationship between people include friendship, email exchange etc. [13].

Potgieter el al., [3] indicate that SNA is a research area aimed at understanding social complexity by representing and analyzing social networks using mathematical graphs. A graph G is a structure consisting of a set of vertices V and a set of links E. Social networks can represent many different types of collaboration links, such as marriages, classmates, friendships, the Hollywood-actor collaboration graph [16], [4] and the academic collaboration graph [1], where two people (i.e. vertices) who collaborated together are connected by an edge (i.e. link). In this research we will use the academic co-author collaboration network. These networks follow a power law distribution of degree also known as scale free networks where vertices that are well connected grow more connections than those who are less connected [1].

The co-authorships networks are an important class of collaboration networks and have been used extensively to determine the academic structure of scientific collaborations and the status of individual researchers [20]. Collaboration networks can be represented by an undirected graph where edges exist between two vertices if they collaborate together. This representation allows researchers to apply SNA methods [18] to evaluate the properties of the network. In an authors network the vertices represents authors and the links are the papers, two authors are considered connected if they have authored at least one paper together.

978-0-7695-3689-7/09 $25.00 © 2009 IEEE
DOI 10.1109/ASONAM.2009.12

Liben-Nowell and Kleinberg [8] have tested the success of how well proximity metrics can predict new links. Results obtained from their work suggest that success rate of the best model does not exceed 16% in co-author’s network when compared to a random prediction's accuracy of less than one percent. One of their hypotheses was that link prediction can be performed from graph topology alone. Liben-Nowell indicate that the method of common neighbors “perform surprisingly well"[7], however this method is limited to predict links between vertices with distance smaller than three edges, since vertices with a common neighbor can have a maximum two edges distance.

A. Data and Experimental Setup

A graph G is a structure consisting of a set of vertices V and a set of links (edges) *E* and can be indicated as *G=(V, E).* A link $e \in E$ represents a connection between two vertices from the set *V.*

The problem of link prediction involves finding relationships between network features, indicating whether two vertices will form a new connection. In other words, trying to find classification quantitative rules that classify dyads into two classes, *C1* and *C2* where *C1* includes vertices which will not build new links and *C2* contains vertices which will be part of a new link, based only on the values of their features. Fitting models to a data set is called data mining, part of the broader field of statistics [11].

Due to the small ratio of actual forming links in the network to potential forming links, the model will not predict actual forming links using a system trained on a random sample of a social network.

This means that the number of positive instances (actual forming links) in the training set is not representative of the number of positive instances that one can find in the whole social network.

The link prediction problem divided to two sub-predictions. In the first prediction, active nodes (authors who will co-author a paper) will be discovered by classifying them into two classes- *C3* and, *C4* where *C3* is a class which indicates that an author will not be active and *C4* indicates that the author will be active, in order to minimize the potential forming links group and then a second prediction will be preformed aiming to predict links between authors (classifying them to *C1* and *C2*).

We partition the range of publication year into two non-overlapping sub-ranges representing training and testing data sets. Each pair of authors either represents a positive example (a link will form) or a negative example (a link will not form), depending on whether those authors published at least one paper. Co-authoring a paper in test years by a pair or group of authors, establishes a link between every possible pair of authors in this group. The DBLP data base which includes information about different research publications in the field of computer science was used. More specifically the data set is from 1970 to 1984 where the first 4 years were used as a training set (See Fig. 1) and the last 10 years as a testing data. Also feature vector for each pair of authors was constructed. A detailed description of the features is given in the following sub-section.

B. Features for Predicting Active Authors

A list of features and definitions is detailed in table 1. Few modifications have been made in order to fit the features to the first prediction procedure, those are discussed below.

VSM - The vector space model provides a way to assess vector similarities based on concept word matches. The procedure is comprised of word indexing, term weighting, and calculating similarity coefficients by distance in a vector space [10]. Word indexing provides the keywords that represent an author as a vector. Term weighting is providing weights to the indexed terms which indicate their importance. Term frequency (*tf)* factor [10] is a normalized measure of a term importance to an author’s vocabulary-where its frequency (*fi*), divided by the maximum frequency of any term in the vocabulary *tf=fi/max* (*fi*). Inverse document frequency *(idf)* - importance of a term is inversely proportional to frequency of occurrence id*f= log(1+N/ni),*where N is the number of authors and n is the number of authors which used a term i. At the end of this stage we have a vector that represents each author$^{idf \times tf}$. Similarity between any two vectors is determined by the distance between vectors in a high-dimensional space such as cosine coefficient which measures the angle between two vectors, *cos(Θ)=* $\frac{V1 \cdot V2}{\|V1\| \cdot \|V2\|}$ [19]. That is, the inner product of the two vectors divided by the products of the vector length. In this research the cosines measure is summed over all the co-authors one author has.

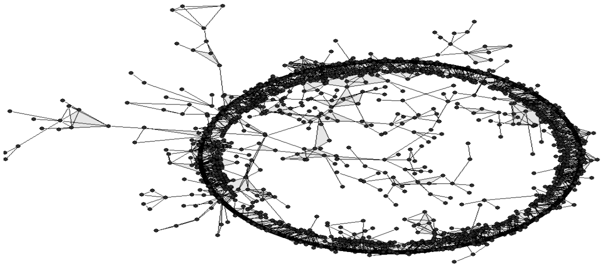

Figure 1. Learning network (1970-1974)

Short_sum - is the shortest distance paths between a vertex and all reachable vertexes, which is significant in link prediction. Kleinberg, [12] discovered that in social network most of the nodes are connected with a very short distance.

TABLE I- METRIC DEFINITIONS

Name	Definition	Description
Text Mining -VSM	$\cos(\alpha) = \frac{A \times B}{\lvert A\rvert \cdot \lvert B\rvert}$	Vector space model (see section 2)
Link's Weight(vi,vj)	Weight(vi,vj)	Weight is defined as the number of papers two authors have co authored together.
Common neighbors 21]	$\lvert (\Gamma(Vi) \cap \Gamma(Vj) \rvert$	The number of Vertices linked to both examined Vertices (i.e. mutual friends).
Jaccard's coefficient [21]	$\frac{\lvert \Gamma(Vi) \cap \Gamma(Vi) \rvert}{\lvert \Gamma(Vi) \cup \Gamma(Vi) \rvert}$	Number of common neighbors of the examined Vertices divided by the number of Vertices that are neighbors of either examined Vertex.
Adamic\ Adar similarity [13]	$\sum_{Vz \in \Gamma(v_i) \cap \Gamma(v_j)} \frac{1}{\log(\lvert \Gamma(v_z) \rvert)}$	The number of features shared by the Vertices, divided by the log of the frequency of the features. This metric rates rarer features more heavily.
Preferential attachment [2]	$\lvert \Gamma(Vi) \rvert \cdot \lvert (\Gamma(Vi) \rvert$	The product of the number of edges incident to the two Vertices.
Cc1	$C_i = \frac{2\lvert\{e_{j,k}\}\rvert}{k_i(k_i-1)} : v_{j,} v_k \in V, e_{jk} \in E$	Clustering Coefficients considering only 1-neighborhood
Cc2	$C_i = \frac{\lvert\{e_1\}\rvert}{\lvert\{e_2\}\rvert}$	Clustering Coefficients considering 2-neighborhood
Degree	$\lvert\Gamma(Vi)\rvert$	The number of links from vertex v i to any other vertex
Normalized degree	$\frac{\lvert\Gamma(Vi)\rvert}{\max(\Gamma(Vi))}$	A standardized degree that ranges from 0 to 1 to ensure that the values fall within a standard range.
Shortest path	$\mathrm{Min}\{dist(v_i, v_j) : v_i \in V\}$	The length of the shortest path between two vertices
Closeness [19]	$\sum_{v_i \in V, v_i \neq v_j} \frac{1}{dist(v_i, v_j)}$	How close the vertex is to all other vertices. Ranges from 0 (far away) to $1/\lvert v\rvert - 1$ (very central).
Betweenness [19]	$\sum_{\substack{v1 \neq v2 \neq vi \\ v1 \neq vi}} \frac{short_{v1,v2}(v_i)}{K - short(v1, v2)}$	Sum of all shortest paths between all vertices that contain vi as a percentage of all shortest paths between all vertices. $short_{v1,v2}(v_i)$ is the number of shortest paths from v1 to v2 that pass through a vertex vi

K-Short_Sum, K-Short_Max are the sum and max, respectively, of k (integer) -shortest distance paths where k is the number of paths with the same length to the destination vertex. The importance of the feature gradually decreases as k increases. The shortest distance is weighted and each edge has an actual weight. For any pair of nodes, the weight on the edge indicates the number of papers the authors have co-authored.

SumInput- This metric is calculated by summing the number of links (papers) that two vertices (authors) published in the training years.

MinInput_Sum -sum of the minimum edge connected to vertex *v1* and the minimum edge connected to *v2*.

MaxInput_Sum-sum of the maximum edge connected to vertex *v1* and the maximum edge connected to *v2*.

CC1 - Clustering Coefficients considering only 1-neighborhood. It is the fraction of pairs of vertex's collaborators who have also collaborated with one another [15].

Many initiatives within social network research [9] indicated clustering index as an important feature of a node in a social network.

CC2 - Clustering Coefficients considering 2-neighborhood. A vertex connected to a highly connected vertex, has a better chance to cooperate with a distant vertex through the later since the number of secondary neighbors in social network usually grows exponentially.

Norm_input- number of publication a vertex has, divided by the maximum degree in the network.

Preferential_Max - the maximum number of edges product incident to two vertices.

Adamic_Sum - sum of features shared by two vertices, divided by the log frequency of the features.

Jaccard_Sum - the sum of common neighbors of the examined vertices divided by the number of vertices that are neighbors of either examined vertex.

CommonNabor_Sum - the sum of the vertices linked to both examined vertices (i.e. mutual friends).

CommonNabor_Max - the maximum number of vertices linked to both examined vertices.

C. Classification Algorithm

Neural Networks (NN) and Decision Trees (DT) were considered in order to deal with the classification problem of trying to classify authors into two classes: C3 and C4 (see sub section 2.1).

The NN models are different from each other because they have different topology. SPSS Clementine was utilized to implement, build and validating the NN and DT models. In the first stage the data was divided into two groups: the training data set which was used to build the model and testing dataset which was used to evaluate the model results. The standard partitioning is choosing randomly 70% of the data into training set and 30% of the data into the testing set [5].

The first NN model was build based on the training set and contained one hidden layer and cross validation on 30% of the data. The model contained three neurons in the hidden layer and was able to reach 73.6% of accuracy over the training set. The second NN model contained two hidden layers, in order to improve the prediction accuracy, the rest of the parameters were not changed. The results of the second model were 74% of accuracy over the training set which is a slight improvement. The third NN model had two hidden layers- the cross validation ratio was changed to 50%. This model achieved accuracy of 75.1% over the training set. Other parameters were used in order to improve the prediction using a NN model with no success of improving the accuracy of 75.1%. The most effective features for the NN model are Jaccard, VSM, cc2 and suminput.

The DT model was built on the same training set as the NN model. Several parameters where used to build the DT models using the C 5.0 algorithm. The forth model included 10 times Cross validation and generated a DT with 17 levels and accuracy level of 86% over the training set. In the fifth DT model the parameters were not changed but the partition of the training and testing sets were changed to 20% of the data for the testing set. The model that was created contained a DT with 23 levels and 87.9% accuracy over the training set. Other partitions did not achieve better performance of accuracy. The sixth and best DT model was created using the Boosting option for 10 times, Boosting causes the DT algorithm to iteratively build a new DT while concentrating on the missed matches occurred in the previous step. Pruning severity was set to 30 affecting the decision of how much pruning to make in order to prevent over fitting.

The DT created using those parameters has 31 levels and 88.5% prediction accuracy over the training set. The most influence features on classifying are: Betweenness, maxinput, kshort and VSM

D. Experiment Results

There exist two types of mistakes: false negative (FN) and false positive (FP). FN is much more concerning than FP, any potential active authors being predicted by a model, even if they have a very low probability to be active, is more interesting and the concern is of missing authors which will be active.

The first NN model achieved accuracy of 71.5% over the testing set. The second two layer NN model did not achieve significant improvement over the testing set. The third NN model also did not achieve better results over the tasting set. The forth DT model achieved accuracy of 88% over the testing set similarly to the accuracy over the training set. The fifth DT model, which was build with a partition on 80% of the data for training and 20% of the data as testing set, achieved 87.6% accuracy over the testing data set, similar to the accuracy achieved over the training set. The sixth and best DT model achieved the highest accuracy level of 90.99% over the testing set. Fig. 2 visualizes the network created by the training data set. The gray vertices represent authors which the model predicted will be active and the black vertices are authors which the model predicted to be non active.

Fig. 3 visualizes the same network as in Fig. 2 but the size of the vertices is proportional to their input degree. The experiments discovered that prediction based on graph metrics could aid link prediction by finding new active authors. Dyadic common neighbor-based local metrics can also be used to predict the active authors, distance-based metrics can be calculated quickly and contribute to the prediction. The sixth DT model has achieved improvement in the FN mistakes but has more FP mistakes compared to the fifth DT model. In this research we are more sensitive to the FN mistakes as discussed earlier. The same models were produced with the same parameters on the same training set and were tested on the same testing data set but with one difference- we excluded the VSM feature. The results were significantly worse for all the models. The most critical features for predicting active authors are the Cos_Sum, Betweenness and the K-Short distance

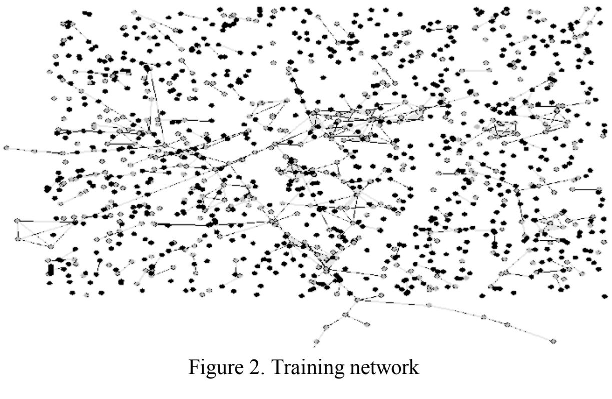

Figure 2. Training network

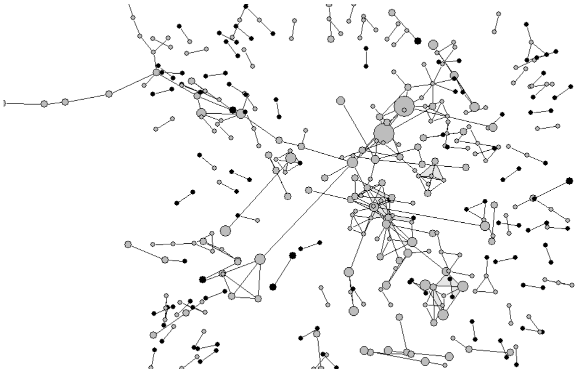

Figure 3. Network predicted by the best DT model- Gray vertices were predicted as “active”

When the classification procedure was engaged without the Cos_Sum feature the prediction results were less accurate showing only 80.93% of accuracy using the best model, but when the Cos_Sum feature was engaged, the prediction accuracy achieved approximately 91% accuracy. Another experiment was carried out in order to improve the prediction accuracy using NLP (Natural Language Processing) stemmer, java tool, instead of the statistics measures as discussed above. This method enables distinguishing words with different grammar inflections, Linguistics-related text mining known as Natural Language Processing. NLP is a research field which deals with formal theories about generating and understanding spoken language. Linguistics can be considered as an automatic processing of natural language, since the main task of computational linguistics is the construction of algorithm to process words and texts in natural language.

Thus, NLP process intelligently text document using the general knowledge about natural language and semantic information [6]. The NLP approach applies the principles of computer-assisted analysis of human languages, to the analysis of words, phrases, and the syntax, or structure, of text. A system that incorporates NLP can intelligently extract terms as part of the TDM process. Using a DT model in the same procedure but instead of the Statistic VSM feature we engaged an NLP feature, the results showed slight improvement achieving 91.86% of accuracy over the testing set.

E. Features for Link Prediction

The second prediction stage is based on the results achieved in the first prediction procedure (described above) which enable us to focus on the "active" authors trying to predict new links between authors. Instead of looking at every possible pair of authors, which in this case is $\binom{4446}{2}$ -about 10 million couples, we can look at 993 different authors which are only $\binom{993}{2}$-about 500,000 couples. In this stage we created a feature set more compatible for co-authors prediction, which are detailed in table 1. NN and DT models were considered in order to deal with the classification problem, this time we tried to classify pairs of authors into classes: C1 and C2. The first NN model was build based on the training set (70% of the data) and contained two hidden layer and cross validation on 50% of the data. The model contained two neurons in the first and second hidden layers. The second NN model contained one hidden layer with 3 neurons. It can be deduced that the most effective features are preferential, VSM, CC2 and MinInput. The DT model was built on the same training set as the NN model. The model included 10 times Cross validation and generated a DT with 28 levels and accuracy level of 97.1% over the training set.

III. Results and Conclusions

This research presented a view into various aspects of the prediction problem in social networks. It described experiments conducted to distinguish between prediction using only SNA methods and prediction using TDM and SNA methods. The aim was to prove the accuracy of prediction in large networks using metrics that contained TDM are more accurate compared to metrics that contained only SNA. The first NN model achieved accuracy of 92.23% over the testing data set. The second NN model achieved 93.24% of accuracy. The DT model achieved accuracy of 96.99% over the testing set.

The DT model has achieved improvement in the FN and FP mistakes compared to the other models. The same models were produced with exactly the same parameters but with one difference, we excluded the statistic VSM feature and engaged the NLP feature. The results were slightly improved and achieved 97.73% of accuracy over the testing set and those are the best achievable results which enable total accuracy of 89.78% in predicting new links in the co-author social network (which is the product of the first prediction and the second prediction).

The experiments discovered that prediction based on graph metrics could aid link prediction by finding new active authors. Dyadic common neighbor-based local metrics can also be used to predict the "active authors".

The experiment found that there is a difference in prediction accuracy between the two methods. Although all metrics are significantly very different no individual metric alone is useful for classifying the two classes. The most critical features for classifying and thus predicting new co-operations between authors are VSM, Wight (link's strength) and Jaccard’s coefficient which indicates the similarity between two vertices group of friends.

The contribution of this work emphasizes the difference between predicting links in a graph using SNA alone and predicting links using a combination of different methods of TDM and SNA. It is shown through empirical testing that the two predictions, with TDM and without

TDM, have different results in favor of using TDM in the prediction algorithm.

References

[1] A.L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, “Evolution of the social network of scientific collaboration”, Physica A 311, 2002, 590-614

[2] A.L. Barabasi, and R. Albert, “Emergence of scaling in random networks”, Science, AAAS, 286, 1999, 509-512

[3] Potgieter, K. A. April, R. J. E. Cooke, and I. O. Osunmakinde, "Temporality in Link Prediction: Understanding Social Complexity", Respectively Department of Computer Science, University of Cape Town, Republic of South Africa, 2007

[4] Aleman-Meza, M. Nagarajan, C.Ramakrishnan, A. Sheth, I. Arpinar, L. Ding, P. Kolari, A. Joshi, and T. Finin, “Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection”. Proceedings of the 15th international conference on World Wide Web, ACM New York, 2006, 407-416

[5] M. Bishop, “Neural Networks for Pattern Recognition”, Clarendon Press, Oxford, 1995

[6] Landau, R. Feldman, Y. Aumann, M. Fresko, Y. Lindell, O. Liphstat, and O. Zamir, “Textvis: An integrated visual environment for text mining In Principles of Data Mining and Knowledge Discovery”, The Pennsylvania Stae University, 1998, 56-64

[7] Liben-Nowell, "An Algorithmic Approach to Social Networks", PhD thesis at MIT Computer Science and Artificial Intelligence Laboratory, 2005

[8] [8] D. Liben-Nowell, and J. Kleinberg, “The link prediction problem for social networks”, Proceedings of the twelfth international conference on information and knowledge management, 2003, 556-559

[9] D. Liben-Nowell and J. Kleinberg, “The Link Prediction Problem for Social Networks”, American society for information science and technology, John Wiley, 58(7),2007,1019

[10] D.L. Lee, H. Chuang, and K. Seamons, “Document ranking and the vector space model”, IEEE, 14 (2), Wiley Blackwell Publishing, 1997, 67-75

[11] I. Witten, and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", 2nd ed. Elsevier, Morgan Kaufmann, San Francisco, USA,2005, 525

[12] J. Kleinberg, “The Small-World Phenomenon: An Algorithmic Perspective”, Proceedings of the 32nd ACM Symposium on Theory of Computing, Cornell University Press, 2000, 163- 170

[13] L.A. Adamic., and E. Adar, “Friends and Neighbors on the Web”, Social Networks, Elsevier, 25(3), 2003, 211-230

[14] L. Getoor, and C. Diehl, “Link Mining: A Survey”, ACM SIGKDD Explorations Newsletter, 7(2),2005, 3-12

[15] M. E. J. Newman, "The structure of scientific collaboration networks", Proceedings of the National Academy of Science, USA, 2001, 404-409

[16] M. E. J. Newman, "Assortative Mixing in Networks", Physical Review Letters, 89(5), APS, Santa Fe, New Mexico, 2002, 1120

[17] M. E. J. Newman, "The structure and function of complex networks", SIAM Review 45,Ann Arbor, Department of Physics and Center for the Study of Complex Systems, University of Michigan, 2003, 167-256

[18] S. Wasserman, and K. Faust, "Social Network Analysis: Methods and Applications", Structural Analysis in the Social Sciences, 8, Cambridge University Press, Cambridge, England, 1994

[19] V.V. Raghavan, and S. K. M. A. Wongcritical, "A critical analysis of vector space model for information retrieval", Journal of the American Society for Information Science, 37(5), Computer Science Department, University of Regina, Regina, Saskatchewan, Canada , 1986, 279-287

[20] X. Liu, J. Bollen, M.L. Nelson, and H. Van de Sompel, “Co-Authorship Networks in the Digital Library Research Community”, Information Processing and Management, 41(6), 2005, 1462-1480

[21] Z. Huang, X. Li, and H. Chen, “Link Prediction Approach to Collaborative Filtering”, Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, IEEE, 2005, 7-11