Combined Proceedings of the Social Network Analysis and Norms for MAS Symposium - SN-MAS2010 Section,
Giulia Andrighetto, Guido Boella, Ugo Pagallo, and Serena Villata (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

A Process Based on the Fuzzy Set Theory for Evaluation of Link Prediction Methods

Elisandra Aparecida Alves da Silva[1], Marco Túlio Carvalho de Andrade[1]

Abstract. Link Prediction is an important research line in the Social Network Analysis context, as predicting the evolution of such nets is a useful mechanism to improve and encourage communication among users. This paper describes a process based on the Fuzzy Set Theory for evaluation of Link Prediction methods. In the data selection step, the use of fuzzy sensors is considered, the determination of new links is based on fuzzy compositions and the fuzzy ROC curve is used to analyse the results.

1 INTRODUCTION

Nowadays many databases are described as a linked collection of interrelated objects. The networks formed by such objects can be homogeneous, in which there is a single object type and link type or heterogeneous networks in which objects and links may be of multiple types. An example of homogeneous network is the WWW (World Wide Web), and examples of heterogeneous networks include co-authorship networks, which are used in this project.

The main goal of traditional data mining algorithms is to find patterns in a dataset characterized by a collection of independent instances of a single relation. However, the application of traditional statistical inference procedures which use independent instances can lead to inappropriate conclusions about data [11]. According to Liben-Nowell [14], a challenge in the Data Mining area is to deal with richly structured, heterogeneous data. This way, linking features among objects need to be used to improve the accuracy of predictive models [7]. Some of these features are mentioned by Davidsen [4]: the correlation between attributes of interconnected objects and the existence of links between objects which have similarity.

Link Mining refers to Data Mining techniques that explicitly consider the links in the development of predictive or descriptive models of interconnected data. The Link Mining tasks, according to the taxonomy presented by Getoor [7], are: object-related tasks (Ranking, Classification, Clustering and Object Identification), link-related tasks (Link Prediction) and graph-related tasks (Subgraph Detection, Graph Classification and Generative Models for Graphs). Link Mining is an emergent area that represents the intersection of different areas: Link Analysis, Web and Hypertext Mining, Relational Learning and Inductive Logic Programming, and Graph Mining.

The main goal of Link Prediction is to determine the existence of a link between two entities using object or link attributes. Link Prediction is useful in different application fields, such as recommendation systems, detection of links not observed in terrorist networks, protein interaction networks, prediction of collaborations between scientists and Web hyperlinks prediction [23].

In related works ad hoc strategies are adopted to evaluate the results of predictions. This paper presents a systematic process to evaluate Link Prediction methods that is based on non dichotomic metrics for data selection, determination of new links and evaluation of results. The use of non dichotomic metrics makes it possible to use the specialist's knowledge and to adopt a perspective more similar to the human perception of the problem. The ROC curve is used to evaluate the results, but this step is under construction.

The paper is organized as follows: Section 2 presents an overview of the main methods of Link Prediction and Section 3 presents the proposed process. Finally, the application of the process and the results are presented.

2 LINK PREDICTION

The main goal of Link Prediction (LP) is to predict the existence of a link between two entities using features of objects and other observed links.

The basic approach of Link Prediction methods is the classification of all node pairs based on graph proximity measure. The link weight called $score(x,y)$ is assigned to each node pair x and y, and then a list is generated in decreasing order of score. Considering node x, $\Gamma(x)$ denotes a neighbor set of x in G_{collab}. Neighbors of x are the nodes which are directly connected with x.

Thus, these methods can be seen as the computation of proximity measure or similarity between nodes x and y, related to the topology of the network. In general, the methods derive from Graph Theory and Social Network Analysis and are designed to measure similarity between nodes. According to Liben-Nowell [14], these methods need to be modified for applications in different contexts.

Many approaches are based on the idea that the greater the number of common neighbors between two objects, the greater the chance of a link between x and y. Davidsen [4] and Jin [12] have proposed abstract models for network growth using this idea. These authors present the most direct idea of the application of Common Neighbors to Link Prediction. Newman [17] has used this measure in the context of collaborations network.

$$\mathrm{score}(x, y) = |\Gamma(x) \cap \Gamma(y)| \qquad (1)$$

[1] Dept. of Computer Engineering and Digital Systems, Univ. of São Paulo, Av. Prof. Luciano Gualberto, travessa 3, 158, SP, Brazil. Email: {**elisandra.silva**, **marco.andrade**}@**poli.usp.br**.

Adamic [1] used the proximity idea to verify the similarity between personal web pages. They consider common neighbors with lower degrees as more relevant, as follows:

$$\text{score}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(|\Gamma(z)|)} \quad (2)$$

Another approach, called Preferential Attachment, assumes that the probability of a new link involving x and y is proportional to the number of their neighbor's links. This measure is given as follows [2]:

$$\text{score}(x,y) = |\Gamma(x)| \times |\Gamma(y)| \quad (3)$$

The Link Prediction methods presented above are based on structural properties of networks and do not consider the connection weights between users. Murata [16] proposed some adjustments based on proximity measures to be used in online social networks. As user's personal information is not generally available in these networks, only the structural properties have been used. Additionally, the connection weight between x and y, $w(x,y)$, was defined as the number of encounters between x and y. A simple adaptation of Common Neighbors including link weights is presented in Murata [16]:

$$\text{score}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x,z) + w(y,z)}{2} \quad (4)$$

In this context, there are also approaches based on path analysis [13], [8] and on graph structure of networks [10]. Different approaches consider the application of probabilistic models [23] and similarity measurements between two objects [20], [22]. The main problems related to these approaches are the high complexity of the probabilistic models and the utilization of node attributes in similarity measurements that sometimes are not available in networks. Additionally, the link's information is not considered in these approaches.

The next section deals with the fuzzy approaches. As Murata [16], we assume the existence of link weights between users in a social network, but considering attributes of co-authorship networks.

3 PROPOSED APPROACH

The determined process involves the steps presented in Figure 1. In the data selection step, the use of fuzzy sensors is considered, the determination of new links is based on fuzzy compositions and the fuzzy ROC curve is used to analyse the results. This way, we have a process based on non dichotomic metrics in order to evaluate the methods of Link Prediction, what makes it possible to use the specialist's knowledge and to adopt a perspective more similar to the human perception of the problem.

Figure 1. Evaluation Process

The following sections present the approaches used in the evaluation process.

3.1 Fuzzy Sensor

According to Mauris [15], the goal of a sensor is to generate a symbolic linguistic representation from numerical measurements, i.e., a numeric-linguistic conversion considering the subjectivity of the problem.

Thus, the sensor creates a symbolic qualitative description in two stages: (1) numeric measurement and (2) numeric-linguistic conversion. A numeric measure, generally obtained by an electronic processing, provides an objective quantitative description of objects. The measure language, usually obtained from the interrogation of users provides a qualitative description of subjective objects. The conversion should provide a very accurate description as the one performed directly by a human. Therefore, to implement a symbolic sensor, the symbolism of the adopted language should be considered in order to artificially reproduce the human perception of the measure.

The fuzzy sensors used for data selection includes two input variables: NumberOfPapers and NumberOfCoauthors and an output variable that determines the choice of the node. Input variables are named so because they represent authors of a co-authorship network. However, they can be obtained in different areas. The NumberOfPapers is the number of encounters with other users and the NumberOfCoauthors represents the number of neighbors.

3.2 New Link Determination

This work views the link weight between two users x and y as the "relation quality". This measure is obtained by the application of approaches that use features from co-authorship social networks, which can be directly used in other domains.

Two approaches based on fuzzy theory are presented here. Both consider the use of fuzzy compositions to determine new links between two authors and employ the relation quality to determine the link weight.

The approaches consider that the quality of the relation between two authors is higher in the following situations:

- When two authors have a large number of papers, mainly in recent years.
- When the average of co-authors of the authors in the relation is low, but the common co-authors are not considered as they influence the relation in a positive way.

The next sections present the technique used for new link prediction.

3.2.1 Fuzzy Compositions

In this paper, a new link prediction approach is proposed using fuzzy compositions. Supposing that $R(X,Y)$ and $R(Y,Z)$ are two fuzzy relations[2], the composition $C(X,Z)$ between $R(X,Y)$ and

[2] A fuzzy relation establishes associations of different truth degrees between related elements, which are similar to the Fuzzy Set membership degrees [24]. A fuzzy relation example is given by "physical similarity between members from x and y".

S(*Y*,*Z*) is a fuzzy relation now between *X* and *Z*, using *Y* as a bridge (transitivity) [24]. It is given by:

$$C(X,Z) = R(X,Y) \circ S(Y,Z) \tag{5}$$

Figure 2 shows some user relations in a specific context as an example. It presents the link weights and their membership degrees.

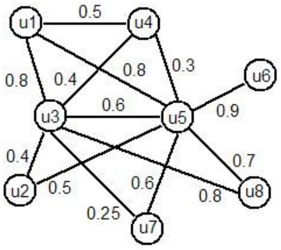

Figure 2. Users relations in a social network

Therefore, using relation compositions makes it possible to predict new link weights connecting users that have not been connected yet. The result of the composition to the previous network is presented in Table 1. The values in the example are obtained by using the Max-product. In the approaches, the Max-Product showed better results than the Max-Min.

Table 1. Relation Composition

	u3	u5
u2	0.4	0.5
u8	0.8	0.7

O

	u7	u4	u1
u3	0.25	0.4	0.8
u5	0.6	0.3	0.8

	u7	u4	u1
u2	0.3	0.16	0.4
u8	0.42	0.32	0.64

3.2.2 Relation Quality

This measure represents the quality of the relation between two users. We propose two approaches to obtain this value. In the first approach this measure is the output variable of the fuzzy model and in the second, this is obtained by using a metric. The next subsections detail these approaches.

Using Fuzzy Model (Fuzzy1)

The input variables used in this approach are NumberOfPapers, CoauthorsAverage and RelationTime.

NumberOfPapers is the number of papers coauthored by *A* and *B*, considering different weights according to the paper's year. Recent papers are more important. The value a paper adds to the total number of papers is obtained as follows:

$$PaperYear - InitialYearOfTraining \tag{6}$$

Thus, the first year of the training period is not relevant to this measure and the total sum is the number of papers coauthored by *A* and *B*.

CoauthorsAverage is the average of coauthors of *A* and *B*, but the common coauthors are not considered. $\Gamma(A)$ is the number of coauthors of *A* and $\Gamma(B)$ is the number of coauthors of *B*. This value is obtained as follows:

$$C_o = \frac{\Gamma(A) + \Gamma(B)}{2} - (\Gamma(A) \cap \Gamma(B)) \tag{7}$$

RelationTime is the difference between the last year of training and the year the oldest paper was published.

The rules used are presented below:

if CoauthorsAverage is low AND NumberOfPapers is low AND RelationTime is low THEN RelationQuality is regular
if CoauthorsAverage is low AND NumberOfPapers is low AND RelationTime is high THEN RelationQuality is low
if CoauthorsAverage is low AND NumberOfPapers is high THEN RelationQuality is high
if CoauthorsAverage is high AND NumberOfPapers is low THEN RelationQuality is low
if CoauthorsAverage is high AND NumberOfPapers is high THEN RelationQuality is regular

RelationTime is important in case of low coauthors average and low number of papers. In this situation, RelationTime is used to determine if quality is low or regular.

Figure 3 shows an example according to the fuzzy model. In this example, u_4 and u_5 have a relation quality of *4.68*. This value is the output of the fuzzy model considering the input variables and rules presented.

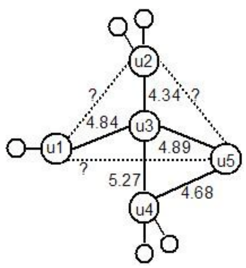

Figure 3. Relation quality using fuzzy model (Fuzzy1)

Using metric (Fuzzy2)

The second approach uses the following metric to determine the RelationQuality. The value a paper adds to the total number of papers is obtained as follows:

Combined Proceedings of the Social Network Analysis and Norms for MAS Symposium - SN-MAS2010 Section,
Giulia Andrighetto, Guido Boella, Ugo Pagallo, and Serena Villata (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

$$\text{PaperYear} - \text{InitialYearOfTraining} \tag{8}$$

$$C_o = \frac{\Gamma(A) + \Gamma(B)}{2} - (\Gamma(A) \cap \Gamma(B)) \tag{9}$$

$$\text{RelationQuality} = \frac{p(A,B)}{C_o} \tag{10}$$

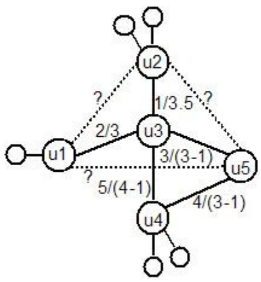

Figure 4. Relation quality using metric (Fuzzy2)

p(A,B) is the number of papers co-authored by *A* and *B*, considering different weights according to the paper's year. Recent papers are more important. Figure 4 shows an example according to the used metric, where users u_4 and u_5 have value *4* to co-authored papers. This value indicates that they produced *1* paper in the last year of training in an interval of *5* years. Value *3* is the average of co-authors of users u_4 and u_5 and *1* is the number of common co-authors.

The next section presents the approach proposed to analyse the results of Link Prediction methods. This step is under construction.

3.3 ROC Curve

According to Prati [19], the ROC analysis is a graphical method to evaluate diagnostic and prediction systems. The ROC graphs were initially proposed to analyze the quality of signal transmission [6]. Nowadays they are used as a powerful tool to evaluate classifiers in Machine Learning and Data Mining areas [3], [21].

The ROC curve is obtained from the rate of false positives and true positives. This way, it is possible to compare these values in various cutoff points, not just considering a single threshold. A measure often used to evaluate classifiers is the Area Under the Curve, which can range from 0 to 1, and the greater the value the better is their performance.

In the context of Link Prediction the main goal is to determine the existence of a link between two entities. In order to do so, the link weight is assigned to each pair of nodes *x* and *y*, and then a list is generated in decreasing order of *score*. This value can represent the membership degree of the link to the fuzzy set Positive. Thus, given that the methods provide a value representing the weight of the link and not only its existence, you can take a fuzzy approach to generate the ROC curve and evaluate the Link Prediction methods.

The methodology adopted in the experiment to evaluate the process defined is presented in the next section.

4 EXPERIMENT METHODOLOGY

The adopted methodology is similar to the one presented in Liben-Nowell [14]. Suppose we have a social net $G = \langle V, E \rangle$ in which edge $e = \langle u, v \rangle \in E$ represents the total of interactions (co-authored papers) between u and v at different times. Having times t_0, t_0', t_1, t_1' and assuming that $t_0 < t_0' < t_1 < t_1'$. $[t_0, t_0']$ the training interval and $[t_1, t_1']$ the test interval are considered. Let $G[t, t']$ consist of all edges in t and t'. Thus, an algorithm has access to the network $G[t_0, t_0']$, generates a list of links that are not present in $G[t_0, t_0']$ which needs to be verified in $G[t_1, t_1']$.

Liben-Nowell [14] observed that the evaluation of Link Prediction methods use parameters k_{training} and k_{test} and assumed that the Core set are nodes that belong to at least k_{training} links in $G[t_0, t_0']$ and at least k_{test} links in $G[t_1, t_1']$. In our work, we consider Core set as nodes selected by the use of Fuzzy Sensors which consider the variables NumberOfPapers and NumberOfCoauthors.

The training interval is denoted as $G = \langle A, E_{old} \rangle$ and E_{new} is used to denote the link set $\langle u, v \rangle$, u and $v \in A$. Let u and v co-author a paper during the test interval but not in the training interval ($E_{\text{new}} = A \times A - E_{\text{old}}$). These are the new interactions sought to be predicted.

Each link predictor p produces a list L_P of pairs $A \times A - E_{\text{old}}$ in decreasing order of score. For the evaluation, we focus on the Core set, thus we denote $E^*_{\text{new}} = E_{\text{new}} \cap \text{Core} \times \text{Core}$ and $n := |E^*_{\text{new}}|$. The performance measure for predictor p is determined as follows:

Considering the first n pairs in the list L_P in $\text{Core} \times \text{Core}$ and determining the size of the intersection of this set of pairs with the E^*_{new} set.

5 DATASET

The experiments were performed according to the methodology presented in the previous section. DBLP (Digital Bibliography & Library Project) is the dataset used in the experiment. This dataset contains data of Computer Science publications and has been used in different works [9], [23].

The DBLP Computer Science Bibliography from the University of Trier contains more than 1.15 million records. The DBLP contains details from publications of conferences proceedings related to Data Mining, Databases, Machine Learning and other areas. The dataset is public and is in XML format [5].

6 RESULTS AND EVALUATION

The methods based on proximity measures and the fuzzy approaches proposed in this paper must be evaluated by comparing their performance measure.

Table 2. Training and Test Periods

Period	Training	Test	$\lvert E_{old} \rvert$	$\lvert E_{new} \rvert$	$\lvert E^{*}_{new} \rvert$
1	1999-2004	2005-2007	663530	745629	52145
2	2000-2006	2007-2009	1057817	523844	42394

Table 3. Number of Authors and Papers in Periods

Period	Authors in Training	Papers in Training	Authors in Test	Papers in Test
1	303615	377143	330324	338712
2	416808	551713	277499	245252

Table 2 shows the periods used in the evaluation and the information about dataset in periods. E^{*}_{new} represents the link number obtained by the use of Fuzzy Sensors. Table 3 presents the additional information about dataset in periods. Two periods were used to evaluate the methods considering different numbers of authors and publications. Period 2 has a greater number of authors and publications compared to period 1.

The results obtained are presented in Table 4. Both approaches use fuzzy compositions to determine new link weights. Fuzzy1 considers the fuzzy model to determine RelationQuality and the Fuzzy2 uses a metric.

The results show that the use of fuzzy composition represents an adequate strategy to tackle the Link Prediction problem. The use of a fuzzy model to determine the RelationQuality is interesting because it allows the use of specialist knowledge and can be adequate to different domains, considering additional or other variables.

The use of the metric presented is very simple and considers that the quality of the relation between two authors is higher in the following situations:

- When two authors have a large number of papers, mainly in recent years.
- When the average of co-authors of the authors in the relation is low, but the common co-authors are not considered, as they influence the relation in a positive way.

Table 4. Results obtained in different periods

Period	Common Neighbors	Preferential Attachment	Adamic/ Adar	Fuzzy1	Fuzzy2
1	3.87	0.33	4.56	5.83	5.68
2	3.83	0.16	4.58	5.71	5.58

Table 4 shows that the Preferential Attachment method presents worse performance in both periods. According to Murata [16], the method is not appropriate for networks which have degree distributions that are almost uniform. Adamic/Adar performs better than Common Neighbor in both periods. The metric used in Fuzzy2 performs better than Fuzzy1 in both periods. The Fuzzy approach (Fuzzy1) used to obtain the RelationQuality is very intuitive and has presented better results than traditional methods.

7 CONCLUSION

The results show that the application of a process based on non dichotomic metrics in order to evaluate the methods of Link Prediction makes it possible to use the specialist's knowledge and to adopt a perspective more similar to the human perception of the problem.

The use of Fuzzy Sensors in data selection step is very simple and intuitive and can be adequate to different domains considering different variables. Fuzzy compositions represent a satisfactory strategy for the problem of Prediction Links. The use of a fuzzy model to determine the RelationQuality is interesting because it allows the specialist knowledge in the field to be exploited in the definition of variables, since some features are peculiar to the type of social network. ROC curve can be used to analyse the results, but this step is under construction.

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.; Jeong, H.; Eda, Z. N.; Ravasz, E.; Schubert, A.; Vicsek, T. Evolution of the social network of scientific collaboration. Physica A, v. 311(3-4), p. 590-614, 2002.

[3] Bradley, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition, v. 30, n. 7, p. 1145-1159, 1997.

[4] Davidsen, J., Holger, E., Bornholdt, S., "Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks," Physical Review Letters, 88(12), p. 128701, 2002.

[5] http://dblp.uni-trier.de/xml/

[6] Egan, J. P. Signal detection theory and ROC analysis. New York, USA: Academic Press, 1975.

[7] Getoor, L., Diehl, C. P., "Link Mining: A Survey," SIGKDD Explorations Special Issue on Link Mining, 7(2), pp. 3-12, 2005.

[8] Glen, J., "Simrank: a measure of structural-context similarity" Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538-543, 2002.

[9] Hasan, M. A., Chaoji, V., Salem, S., Zaki, M., "Link Prediction using Supervised Learning," Workshop on Link Analysis, Counter-terrorism and Security (with SIAM Data Mining Conference), 2006.

[10] Huang, Z., "Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient," Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (LinkKDD2006), 2006.

[11] Jensen, D., "Statistical Challenges to Inductive Inference in Linked Data," Seventh International Workshop on Artificial Intelligence and Statistics, 1999.

[12] Jin, E. M., Girvan, M., Newman, M. E., "The structure of growing social networks," Physical Review E, 64(4), 2001.

[13] Katz, L., "A new status index derived from sociometric analysis," Psychometrika, 18(1), pp.39-43, 1953.

[14] Liben-Nowell, D., Kleinberg, J., "The Link-Prediction Problem for Social Networks," Journal of the American Society for Information Science and Technology, 58(7),pp.1019-1031, 2007.

[15] Mauris, G.; Benoit, E.; Foulloy, L. Fuzzy symbolic sensors - from concept to applications. Measurement, v. 12, n. 4, p. 357-384, 1994. ISSN 0263-2241. Disponível em: <http://www.sciencedirect.com/science/article/B6V42-47YRV6C-6/2/bba07d339e347b07a06c370afb0e8f21>.

[16] Murata, T., Moriyasu, S., "Link Prediction based on Structural Properties of Online Social Networks," New Generation Computing, 26(3),pp.245-257, 2008.

[17] Newman, M. E., "The structure of scientific collaboration networks," Proceedings of the National Academy of Sciences USA, 98(2), pp.404-409, 2001.

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

[19] Prati, R. C., Batista, G. E. A. P. A., Monard, M. C. Curvas ROC para a Avaliação de Classificadores. Revista IEEE América Latina, v. 6, p. 1-8, 2008.

[20] Popescul, R., Ungar, L. H., "Statistical Relational Learning for Link Prediction", 2003.

[21] Spackman, K. A. Signal detection theory: Valuable tools for evaluating inductive learning. Proceedings of the 6th Int. Workshop on Machine Learning, p. 160-163, 1989.

[22] Taskar, B., Wong, M.-F., Abbeel, P., Koller, D., "Link Prediction in Relational Data," Neural Information Processing Systems Conf., 2003.

[23] Wang, C., Satuluri, V., Parthasarathy, S., "Local Probabilistic Models for Link Prediction" Data Mining, IEEE International Conference on, pp.322-331, 2007.

[24] Zadeh, L. A., "Fuzzy Sets," Information and Control, 8(3), pp.338-353, 1965.