Co-authorship prediction in academic social network

William Takahiro Maruyama[1], **Luciano Antonio Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP)
Av. Arlindo Béttio, Ermelino Matarazzo – 03828-000 – São Paulo – SP – Brasil

Abstract. *The prediction of relationships in a social network is a complex and extremely useful task to enhance or maximize collaborations by indicating the most promising partnerships. In academic social networks, prediction of relationships is typically used to try to identify potential partners in the development of a project and/or co-authors for publishing papers. This paper presents an approach to predict coauthorships combining artificial intelligence techniques with the state-of-the-art metrics for link predicting in social networks.*

1. Introduction

The growth of the Web has enabled the creation of many tools for users' interaction, such as online social networks (Facebook, Twitter, LinkedIn, etc.). They are increasingly present in the day lives, and are an abstraction of a (real) social network and can directly or indirectly reflect the interactions of people in the real world [Viswanath et al. 2009]. This happens because these networks provide information exchange and communication between users virtually. Due to its popularity and consequently the enormous amount of data produced, the analysis of social networks has attracted considerable attention from the scientific community, in order to better understanding the behavior of human interactions. In a social network, people or groups of people are represented by the vertices (nodes), and the relationships, interactions or links between them, the edges of a graph [Wasserman and Faust 1994, Newman 2001, Hasan and Zaki 2011, Newman 2010]. These links between people in a social network may be intangible and can have many meanings as friendships, family, professional, interaction, among others.

One type of social network is the scientific collaboration network or academic network. In these networks, the vertices represent researchers, and the edges scientific collaboration. Therefore, two researchers are connected if they are coauthors in one or more publications. In this context, you can explicitly build an academic social network from information about publications [Newman 2010, Newman 2001, Barabási et al. 2002]. According to Newman (2010), the scientific collaboration network is genuinely a network of people, because it has a significant amount of people and the collaborations are a direct reflection of the documentation of the authors collaborations.

In the social network analysis field, the data mining techniques that are focused on links is called Link Mining [Getoor and Diehl 2005, Maruyama and Digiampietri 2016]. One of its tasks is the link prediction. This task aims to predict the formation of links between network entities. The prediction of relationships within a social network is a task that has gained attention in recent years, because it may helps from finding friends who were not connected in an online social network [Vasuki et al. 2010, Tian et al. 2010, Perez et al. 2012, Fire et al. 2011, Zhong et al. 2013, Quercia and Capra 2009], to enhance the performance of work in companies or in universities [Hsieh et al. 2013,

Dong et al. 2012, de Sa and Prudencio 2011]. In the academic social network context, the link prediction is mainly used for predicting coauthorships, whose purpose is to indicate a pair of researchers that potentially will collaborate in the publication of papers [Guo and Guo 2010, Makrehchi 2011, Dong et al. 2011, Gao et al. 2012, Lin et al. 2012, Digiampietri et al. 2013, Digiampietri et al. 2015].

Scientific collaboration can improves the production of scientific works [Pavlov 2007], aggregating different expertise and perspectives, and with a possible division of efforts. However, finding suitable researchers for each team is not a quick and easy task. Thus, link prediction can enhance team building, by recommending researchers to work together. One way of performing the link prediction is addressing it as a classification problem, i.e., using a supervised learning strategy. Therefore, finding a set of characteristics is a very important step [Hasan et al. 2006].

This paper presents a link prediction system which combines domain specific attributes with network structural ones, in order to perform a binary classification. The classification will indicate if a pair of researchers will or not became collaborators. Two types of information are considered: the ones extracted from the application domain; and the ones from the network topology.

The rest of this paper is organized as follows. Section 2 summarizes the related work. Section 3 presents the methodology. Section 4 contains the results, and, finally, Section 5 presents the conclusions.

2. Related Work

A supervised learning strategy to predict links was introduced by Liben-Nowell and Kleinberg [Liben-Nowell and Kleinberg 2003]. The authors derive a set of similarity measures from the graph topology. Five sections of the Physics arXiv e-Print were used to build five co-authorship networks. They defined a three-year time window. The period from 1994 to 1996 was used for training and 1997 to 1999 for testing. They compared their results with the prediction using a random method. The results obtained with *Katz* and its variants showed good performance in most of the data sets. Moreover, according to the authors, *Common Neighbors* and *Adamic-Adar* attributes did not achieved bad results.

Bartal et al. [Bartal et al. 2009] combined social network analysis with text mining to predict co-authorships in academic journals using thirteen attributes (one related to text mining and the others related to social network analysis) and achieved a hit rate of 91%, using data from DBLP[1].

Hasan and Zaki [Hasan et al. 2006] conducted an empirical research on link prediction in a co-authorship network, using data from BIO-BASE[2] and DBLP. For BIOBASE a time window from 1998 to 2002 was used, where the first 4 years were used as training and the last one as test. For DBLP, the years from 1990 to 2004 were used, the first 11 years used as training and the last 4 for testing. They extracted nine attributes considering aggregating functions, proximity and topology. The SVM classifier with RBF kernel obtained the best results, with a square error of 0.0945 and 0.1760

[1]Digital Bibliography & Library Project: http://dblp.uni-trier.de/xml/

[2]Bibliographic Database, Elsevier BIOBASE - Current Awareness in Biological Sciences (CABS)

in BIOBASE and DBLP, respectively. According to the authors, the best attributes for linking prediction were: shortest path in BIOBASE and keywords in common in DBLP.

In order to predict new links and treating this problem as a time series, Soares et al. [da Silva Soares and Bastos Cavalcante Prudencio 2012] used two subsets from arXiv: theoretical high energy physics (data from 1991 to 2010) and high energy physics (1993-2010). The basic idea is to construct time series for each pair of nodes not connected by using a similarity score calculated using a topological metric. A predictive model is used to predict the next value in the series. This value will be used for the link prediction methods, tested using both supervised and unsupervised approaches. According to the authors, the supervised approach was better in all predictions models.

Lu et al. [Lu et al. 2010] developed a different approach, but also used arXiv data (from high energy physics, 1992-2003), with data from CiteSeer (1993-2003) and from the Society of Industrial and Applied Mathematics Publications (1999-2004). They proposed an approach to predict coauthoring relationships, citations and references. They used a supervised approach, multiple data sources and historical network observations. According to the authors, the experimental results confirm the accuracy of the approach.

Sun et al. [Sun et al. 2012] developed an algorithm that besides trying to predict links, it also tries to identify when these new links will occur. In another study, Sun et al. [Sun et al. 2011] focused on the prediction of links in bibliographic networks using different structural metrics. In tests conducted using DBLP data, the maximum hit rate was about 75%.

In their review, Hasan et al. [Hasan et al. 2006] identified three different models for supervised link prediction: binary classification, probabilistic model and linear algebraic model. The binary classification uses a set of features that can be extracted from the graph topology and from properties of the vertices and edges. The topological attributes can be subdivided in: node based and neighborhood based. Lu and Zhou [Lü and Zhou 2010] performed a review, in which link prediction methods where classified in three groups: similarity based, maximum likelihood based, and probabilistic models. The first group uses network structural information (topological), the features calculated can be subdivided into local and global.

3. Methodology

The methodology was composed of four activities: data gathering; features extraction/calculation; classification; and analysis of the results. Figure 1 outlines a schematic representation of activities performed.

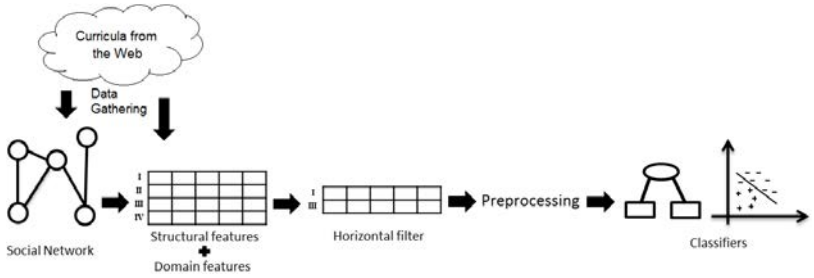

Figure 1. Schematic representation of the work process

3.1. Data gathering

All data used is public available on the Lattes Platform. This is a system created and maintained by CNPq (The Brazilian National Council for Scientific and Technological Development), which records the science curricula of the researchers, mostly Brazilians. It is possible to access information about research projects, publications, academic formation, areas of interest, and advisoring.

The sample selected correspond to all the 657 researchers that were professors in Brazilian Computer Science Graduate Programs during the period 2004-2009. We collected data in a time window from 1971 to 2015. For the training step, data from 1971 to 2000 was considered *past*; from 2001 to 2005 *present*; and from 2006 to 2010 was considered *future* (i.e., the coauthorships the system should predict). For the test step, data from 1976 to 2005 was considered *past*; from 2006 to 2010 *present* and the system tried to predict the coauthorships that occurred from 2011 to 2015.

The researchers curricula were downloaded from the Internet in XML format directly from the Lattes Platform. From these XML files a relational database was created using the methodology presented in [Digiampietri et al. 2012a, Digiampietri et al. 2012b], among the activities performed, we highlight the publication disambiguation (in order to detect collaborative publications), and researchers' names disambiguation (in order to identify the advisor-advise relationships). An academic social network was built based on the coauthorship relationships.

3.2. Features extraction/calculation

From the curricula information 32 attributes/features were extracted or calculated (Table 1). The first one is the class (which indicates if the pair of researchers will or not become coauthors); there are 16 domain specific attributes, and 15 structural (topological) attributes. The structural attributes were selected from the related literature (see Section 2). These attributes were calculated for all the pairs of researchers in the sample.

3.3. Classification

All the features extracted (or calculated) from the data were used as input data for the classifiers. In this work, several classifiers from Weka project[3] were used.

Two types of tests were performed: with and without the balancing of the training data. The balancing strategy used was the oversampling of the minority class.

The solution of two different link prediction problems were tested. The new link prediction (i.e., the prediction of links that do not exist in the *present*), and the general link prediction problem (the prediction of link independently of their existence in the *present*).

The combinations of the 657 researchers in pairs produces 215.496 pairs, most of them will not become coauthors in the future. Before the execution of the classifiers, an horizontal filter was performed to exclude the pairs with low potential to become coauthors. The criteria used was to exclude all the pairs, which the ten first domain specific features had values equal to zero. This process excluded more than 200.000 pairs (see Section 4), which were pre-classified as "will not become coauthors". Only one pair was

[3]http://www.cs.waikato.ac.nz/ml/weka/

Table 1. Attributes used in the proposed system

Feature	Description
class	This attribute assumes the value *true* if the pair of researchers will become coauthors, and *false* otherwise.
past journal papers	Number of jointly journal papers published by the pair of researchers in the *past*.
past conference papers	Number of jointly conference papers published by the pair of researchers in the *past*.
present journal papers	Number of jointly journal papers published by the pair of researchers in the *present*.
present conference papers	Number of jointly conference papers published by the pair of researchers in the *present*.
past advisoring	This attribute assumes the value 1 if the one of the researchers in the pair was the advisor of the other in the *past*.
present advisoring	This attribute assumes the value 1 if the one of the researchers in the pair was the advisor of the other in the *present*.
on going advisoring	This attribute assumes the value 1 if the one of the researchers in the pair was the advisor of the other in an ongoing advisoring.
common advisors	Number of common advisors that the pair of researchers shared.
common advisees	Number of common advisees that the pair of researchers shared.
common graduate program	This attribute indicates if the pair of researchers work in the same graduate program.
journal papers1	Number of journal papers published in the present by the first researcher in the pair.
conference papers 1	Number of conference papers published in the present by the first researcher in the pair.
journal papers 2	Number of journal papers published in the present by the second researcher in the pair.
conference papers 2	Number of conference papers published in the present by the second researcher in the pair.
geographical distance	Geographical distance between the professional address of the two researchers in the pair.
common subareas	Number of common interesting research areas.
past and present CN	Number of common neighbors in the social academic network built using data from *past* and *present*.
present CN	Number of common neighbors in the social academic network built using data only from the *present*.
SAL	Salton Index - measures the co-occurrence of two elements divided by the square root of the multiplication of the occurrence of each element. On social networks can be used to measure the relationship between the number of neighbors that two people have in common divided by the square root of the number of neighbors multiplication of each.
JAC	Jaccard's coefficient - measures the similarity between two sets by dividing the number of elements of the intersection of the sets by the number of union members (e.g. number of common neighbors divided by the union of the neighbors of two people).
AA	Adamic-Adar - index that assigns weight in the relationship of two people favoring the relationships between people who have few relationships.
RA	Resource Allocation - index that assigns weight in the relationship of two people favoring the relationships between people who have few relationships.
SOR	Sorensen Index - index calculated as being twice the intersection between two sets divided by the sum of each set of elements (e.g., number of common neighbors divided by the the number of people in the union of the neighbors).
HPI	Hub Promoted Index - index calculated by dividing the number of intersection of two sets of elements by the minimum number of elements of the sets (e.g., number of neighbors in common of two persons divided by the minimum number of neighbors of these persons).
HDI	Depressed hub Index - index calculated by dividing the number of intersection of two sets of elements divided by the maximum number of elements of these two sets (e.g., number of common neighbors of two people divided by the maximum number of neighbors of these people).
LHM	Leicht-Newman-Holme Index - index calculated by dividing the number of elements of intersection of two sets by the multiplication of the number of elements of each set (e.g., number of common neighbors divided by multiplication of the number of neighbors of each person).
PA	Preferential Attachment - index calculated multiplying the number of elements from two sets (for example, multiplying the number of neighbors).
KATZ0.05 KATZ0.005 KATZ0.0005	Katz is an index calculated iteratively to estimate the influence of a pair of nodes in a network considering the existing paths between nodes. For this calculation, it is necessary to define a constant *Beta*. The values used were: 0.05; 0.005; and 0.0005.
SP	Shortest Path - graph shortest path between the two researchers.

incorrectly classified using this approach. The results presented in Section 4 will only deal with the remaining of the pairs.

3.4. Analysis of the results

The results were analyzed considering the metrics: accuracy, recall and area under curve (AUC). These results will be presented in the next section.

4. Results

This section presents the results for the link prediction general problem (Subsection 4.1) and prediction of new links (Subsection 4.2). In each subsection, we first present the results considering the unbalanced training set and, after these results, the ones using a balanced training set.

4.1. Link prediction general problem

The dataset (after the horizontal filtering) for the experiments presented in this section is composed of 26,979 pairs. From these, 11,833 pairs belong to the training set (10,955 negative instances and 878 positive ones) and 15,146 pairs belong to the test set (14,425 negative instances and 721 positive ones). It is worth to mentioning that a classifier which classify all the instance in the test set as negative would have an accuracy of 95.24%. This value will be considered the baseline for the accuracy.

Table 2 presents the top 3 accuracy results. The best result was achieved by the *AttributeSelectedClassifier* (96.091%), it corresponds to an improvement lower then 1% when compared with the baseline, but for the link prediction problem it is a relevant improvement. Moreover, it was verified that no individual feature was able to achieve a better improvement. The *AttributeSelectedClassifier* was able to correctly classify 41.3% of the positive instances (recall).

Table 2. Top 3 accuracy results

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
Attribute Selected Classifier	F	0.778	0.98	0.971	0.988	0.587	**96.091**
	T	0.778	0.502	0.638	0.413	0.012	
	Avg	0.778	0.957	0.955	0.961	0.559	
BFTree	F	0.865	0.98	0.965	0.995	0.73	**96.065**
	T	0.865	0.396	0.736	0.27	0.005	
	Avg	0.865	0.952	0.954	0.961	0.695	
ADTree	F	0.782	0.98	0.964	0.996	0.742	**96.045**
	T	0.782	0.383	0.744	0.258	0.004	
	Avg	0.782	0.951	0.954	0.96	0.707	

Table 3 presents the top 3 recall results. The *VFI* classifier was able to correctly classify 88.2% of the positive instances, but achieved a general accuracy of only 54.998%. In other words, in order to achieve high values of recall, this classifier incorrectly classified almost half of the test instances.

Table 3. Top 3 recall results

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
VFI	F	0.829	0.693	0.989	0.533	0.118	54.998
	T	0.829	0.157	0.086	**0.882**	0.467	
	Avg	0.829	0.668	0.946	0.55	0.134	
BayesNet	F	0.881	0.921	0.986	0.864	0.24	85.924
	T	0.881	0.34	0.219	**0.76**	0.136	
	Avg	0.881	0.894	0.95	0.859	0.235	
Naive Bayes Updateable	F	0.874	0.943	0.985	0.905	0.28	89.601
	T	0.87	0.397	0.274	**0.72**	0.095	
	Avg	0.873	0.917	0.951	0.896	0.271	

Table 4 presents the top 3 AUC results. This metric is typically used because presents a good tradeoff between the accuracy and recall. DMNBtext achieved the best result (0.886), with a accuracy of 95.444% and a recall of 45.5%.

Table 4. Top 3 AUC results

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
DMNBtext	F	0.886	0.976	0.973	0.979	0.545	95.444
	T	0.886	0.487	0.525	0.455	0.021	
	Avg	**0.886**	0.953	0.952	0.954	0.52	
Bayes Net	F	0.881	0.921	0.986	0.864	0.24	85.924
	T	0.881	0.34	0.219	0.76	0.136	
	Avg	**0.881**	0.894	0.95	0.859	0.235	
Logit Boost	F	0.88	0.979	0.971	0.986	0.589	95.88
	T	0.88	0.487	0.598	0.411	0.014	
	Avg	**0.88**	0.955	0.953	0.959	0.562	

For the next results presented in this section, the training set was balanced using the *oversampling* technique.

Table 5 presents the top 3 accuracy results. The best result (95.629%), achieved by the *RandomCommittee* classifier is worse than the one achieved without the training set balancing (96.091%), what was expected for this type of problem. But, even the recall was also worse (39.5%).

Table 5. Top 3 accuracy results - balanced training set

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
Random Committee	F	0.822	0.977	0.97	0.984	0.605	**95.629**
	T	0.822	0.463	0.558	0.395	0.016	
	Avg	0.822	0.953	0.951	0.956	0.577	
Rotation Forest	F	0.829	0.975	0.973	0.977	0.534	**95.286**
	T	0.829	0.485	0.505	0.466	0.023	
	Avg	0.829	0.952	0.951	0.953	0.51	
ZeroR	F	0.5	0.976	0.952	1	1	**95.24**
	T	0.5	0	0	0	0	
	Avg	0.5	0.929	0.907	0.952	0.952	

Table 6 presents the top 3 recall results. The two best results for this metric corresponds to classifiers which classified all the instances as positive (achieving an accuracy of 4.76%). The *ClassificationViaClustering* was able to recall 98.1% of the positive instances, but with an accuracy of 9.97%.

Table 6. Top 3 recall results - balanced training set

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
StackingC	F	0.5	0	0	0	0	4.76
	T	0.5	0.091	0.048	**1**	1	
	Avg	0.5	0.004	0.002	0.048	0.048	
DMNBtext	F	0.744	0	0	0	0	4.76
	T	0.744	0.091	0.048	**1**	1	
	Avg	0.744	0.004	0.002	0.048	0.048	
Classification Via Clustering	F	0.518	0.105	0.983	0.056	0.019	9.97
	T	0.518	0.094	0.049	**0.981**	0.944	
	Avg	0.518	0.105	0.938	0.1	0.063	

Table 7 presents the top 3 AUC results. The *ADTree* classifier was able to achieve an AUC of 0.88, with an accuracy of 88.34% and a recall of 74.1% of the positive instances. This result was one of the best tradeoff achieved between accuracy and recall (an useful result for the users that needs a high recall value).

4.2. Prediction of new links

The dataset for the experiments presented in this subsection is composed of 25,088 pairs, and corresponds to the same datase used in the previous subsection, excluding the pairs

Table 7. Top 3 AUC results - balanced training set

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
ADTree	F	0.88	0.936	0.986	0.891	0.259	88.34
	T	0.88	0.377	0.253	0.741	0.109	
	Avg	**0.88**	0.909	0.951	0.883	0.252	
Naive Bayes Simple	F	0.876	0.941	0.985	0.901	0.279	89.245
	T	0.871	0.39	0.267	0.721	0.099	
	Avg	**0.876**	0.915	0.951	0.892	0.27	
Threshold Selector	F	0.874	0.927	0.986	0.875	0.251	86.927
	T	0.874	0.353	0.231	0.749	0.125	
	Avg	**0.874**	0.9	0.95	0.869	0.245	

that are coauthors in the *present*. From these, 10,976 pairs belong to the training set (10,537 negative instances and 439 positive ones) and 14,112 pairs belong to the test set (13,838 negative instances and 274 positive ones). A classifier which classifies all the instance in the test set as negative would have an accuracy of 98.05839%. This value will be considered the baseline for the accuracy.

Table 8 presents the top 3 accuracy results. *BayesianLogisticRegression* presented the best result of accuracy (98.065%), which was slightly higher than the baseline. But, it presented poor recall (0.4%). The other accuracy results presented in this table correspond to classifiers that classified all the instances as negative.

Table 8. New links - top 3 accuracy results

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
Bayesian Logistic Regression	F	0.502	0.99	0.981	1	0.996	**98.065**
	T	0.502	0.007	1	0.004	0	
	Avg	0.502	0.971	0.981	0.981	0.977	
NBTree	F	0.622	0.99	0.981	1	1	**98.058**
	T	0.622	0	0	0	0	
	Avg	0.622	0.971	0.962	0.981	0.981	
Decision Stump	F	0.572	0.99	0.981	1	1	**98.058**
	T	0.572	0	0	0	0	
	Avg	0.572	0.971	0.962	0.981	0.981	

Table 9 presents the top 3 recall results. The *VFI* classifier was again the best classifier regarding recall. It was able to correctly classify 85.8% of the positive instances, but achieved a general accuracy of only 37.791%. The *ClassificationViaClustering* achieved the second best recall (46.7%) with an accuracy of 81.094%.

Table 9. New links - top 3 recall results

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
VFI	F	0.71	0.537	0.992	0.368	0.142	37.791
	T	0.71	0.051	0.026	**0.858**	0.632	
	Avg	0.71	0.528	0.974	0.378	0.152	
Classification Via Clustering	F	0.642	0.895	0.987	0.818	0.533	81.094
	T	0.642	0.088	0.048	**0.467**	0.182	
	Avg	0.642	0.879	0.969	0.811	0.526	
Bayes Net	F	0.734	0.914	0.987	0.85	0.58	84.212
	T	0.734	0.094	0.053	**0.42**	0.15	
	Avg	0.734	0.898	0.969	0.842	0.572	

Table 10 presents the top 3 AUC results. The best result was achieved by *ClassificationViaRegression* (0.752), but with poor recall (0.007). We can highlight the results from the DMNBtext, which presented approximately same AUC, but with accuracy and recall of the positive class slightly higher than the first classifier.

Table 10. New links - top 3 AUC results

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
Classification Via Regression	F	0.752	0.99	0.981	1	0.993	98.037
	T	0.752	0.014	0.286	0.007	0	
	Avg	**0.752**	0.971	0.967	0.98	0.973	
DMNBtext	F	0.751	0.99	0.981	1	0.985	98.058
	T	0.751	0.028	0.5	0.015	0	
	Avg	**0.751**	0.972	0.972	0.981	0.966	
Naive Bayes Updateable	F	0.742	0.949	0.986	0.914	0.675	90.271
	T	0.742	0.115	0.07	0.325	0.086	
	Avg	**0.742**	0.932	0.968	0.903	0.664	

For the next results presented in this section, the training set was balanced using the *oversampling* technique. Table 11 presents the top 3 accuracy results. As occurred with the unbalanced, the best results are the ones that classified all instances as negative.

Table 11. New links - top 3 accuracy results - balanced training set

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
ZeroR	F	0.5	0.99	0.981	1	1	**98.058**
	T	0.5	0	0	0	0	
	Avg	0.5	0.971	0.962	0.981	0.981	
Vote	F	0.5	0.99	0.981	1	1	**98.058**
	T	0.5	0	0	0	0	
	Avg	0.5	0.971	0.962	0.981	0.981	
Stacking	F	0.5	0.99	0.981	1	1	**98.058**
	T	0.5	0	0	0	0	
	Avg	0.5	0.971	0.962	0.981	0.981	

Table 12 presents the top 3 recall results. The two best results for this metric corresponds to classifiers which classified all the instances as positive (achieving an accuracy of 1.942%). *ConjunctiveRule* achieved a recall of 87.2% with an accuracy of 38.91%.

Table 12. New links - top 3 recall results - balanced training set

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
StackingC	F	0.5	0	0	0	0	1.942
	T	0.5	0.038	0.019	**1**	1	
	Avg	0.5	0.001	0	0.019	0.019	
DMNBtext	F	0.449	0	0	0	0	1.942
	T	0.449	0.038	0.019	**1**	1	
	Avg	0.449	0.001	0	0.019	0.019	
Conjunctive Rule	F	0.626	0.549	0.993	0.38	0.128	38.91
	T	0.626	0.053	0.027	**0.872**	0.62	
	Avg	0.626	0.54	0.975	0.389	0.137	

Table 13 presents the top 3 AUC results. *NaiveBayesUpdateable* and *NaiveBayes* achieved the same results: an AUC of 0.742, with an accuracy of 87.252%, and a recall of 39.8%. *ThresholdSelector* achieved an AUC of 0.738, with a little lower accuracy (85.636%) and a higher recall (46.0%).

5. Conclusions

This paper presented a study about the prediction of links in academic social networks. The problem of link prediction was treated as an classification problem, and 32 attributes/features (domain specific and structural) were extracted from the curricula data to be used by different classifiers. Among these attributes are the ones that achieved the best results in the related literature.

Table 13. New links - top 3 AUC results - balanced training set

Classifier		AUC	F-Measure	Precision	Recall	FP rate	%Accuracy
Naive Bayes Updateable	F	0.742	0.931	0.987	0.882	0.602	
	T	0.742	0.108	0.063	0.398	0.118	87.252
	Avg	**0.742**	0.915	0.969	0.873	0.593	
Naive Bayes	F	0.742	0.931	0.987	0.882	0.602	
	T	0.742	0.108	0.063	0.398	0.118	87.252
	Avg	**0.742**	0.915	0.969	0.873	0.593	
Threshold Selector	F	0.738	0.922	0.988	0.864	0.54	
	T	0.738	0.111	0.063	0.46	0.136	85.636
	Avg	**0.738**	0.906	0.97	0.856	0.532	

In order to deal with the link prediction problem in academic social networks it is necessary to analyze the tradeoffs between the recall of the positive class and the general accuracy. Tests considering different metrics were performed and the oversampling balancing strategy was used. Most of the tested classifiers was not able to produce results more accurate than the ones produced by a simple classification strategy which considers all the instances as negative ones. Some promising results were achieved in the general link prediction problem. But no classifier was able to achieve a satisfactory accuracy for the prediction of new links (links between researchers that were not related).

Acknowledgments

The work presented in this paper was funded by CAPES and CNPq (processes 306046/2013-0 and 477246/2013-3).

References

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614.

Bartal, A., Sasson, E., and Ravid, G. (2009). Predicting links in social networks using text mining and sna. In *Social Network Analysis and Mining, 2009. ASONAM '09. Int. Conf. on Advances in*, pages 131–136.

da Silva Soares, P. and Bastos Cavalcante Prudencio, R. (2012). Time series based link prediction. In *The 2012 Int. Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

de Sa, H. and Prudencio, R. (2011). Supervised link prediction in weighted networks. In *The 2011 Int. Joint Conference on Neural Networks (IJCNN)*, pages 2281–2288.

Digiampietri, L., Mena-Chalco, J., de Jésus Pérez-Alcázar, J., Tuesta, E. F., Delgado, K., and Mugnaini, R. (2012a). Minerando e caracterizando dados de currículos lattes. In *CSBC 2012 - BraSNAM*.

Digiampietri, L., Mena-Chalco, J., Silva, G. S., Oliveira, L., Malheiro, A., and Meira, D. (2012b). Dinâmica das relações de coautoria nos programas de pós-graduação em computação no brasil. In *CSBC 2012 - BraSNAM*.

Digiampietri, L., Santiago, C., and Alves, C. (2013). Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em ciência da computação. In *CSBC 2013 - BraSNAM*.

Digiampietri, L. A., Maruyama, W. T., Santiago, C. R. N., and da Silva Lima, J. J. (2015). Um sistema de predição de relacionamentos em redes sociais. In *XI Simpósio Brasileiro de Sistemas de Informação (SBSI 2015)*, pages 139–146.

Dong, Y., Ke, Q., Rao, J., Wang, B., and Wu, B. (2011). Random walk based resource allocation: Predicting and recommending links in cross-operator mobile communication networks. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th Int. Conf. on*, pages 358–365.

Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N., Rao, J., and Cao, H. (2012). Link prediction and recommendation across heterogeneous social networks. In *Data Mining (ICDM), 2012 IEEE 12th Int. Conf. on*, pages 181–190.

Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., and Elovici, Y. (2011). Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third Int. conference on and 2011 ieee third Int. conference on social computing (socialcom)*, pages 73–80.

Gao, S., Denoyer, L., and Gallinari, P. (2012). Link prediction via latent factor blockmodel. In *Proceedings of the 21st Int. Conf. Companion on World Wide Web*, WWW '12 Companion, pages 507–508, New York, NY, USA. ACM.

Getoor, L. and Diehl, C. P. (2005). Link mining: A survey. *SIGKDD Explor. Newsl.*, 7(2):3–12.

Guo, J. and Guo, H. (2010). Multi-features link prediction based on matrix. In *Computer Design and Applications (ICCDA), 2010 Int. Conf. on*, volume 1, pages V1–357–V1–361.

Hasan, M. and Zaki, M. (2011). A survey of link prediction in social networks. In Aggarwal, C. C., editor, *Social Network Data Analytics*, pages 243–275. Springer US.

Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

Hsieh, C.-J., Tiwari, M., Agarwal, D., Huang, X. L., and Shah, S. (2013). Organizational overlap on social networks and its applications. In *Proceedings of the 22Nd Int. Conf. on World Wide Web*, WWW '13, pages 571–582, Republic and Canton of Geneva, Switzerland. Int. World Wide Web Conferences Steering Committee.

Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the Twelfth Int. Conf. on Information and Knowledge Management*, CIKM '03, pages 556–559, New York, NY, USA. ACM.

Lin, Z., Yun, X., and Zhu, Y. (2012). Link prediction using benefitranks in weighted networks. In *Proceedings of the The 2012 IEEE/WIC/ACM Int. Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 423–430, Washington, DC, USA. IEEE Computer Society.

Lü, L. and Zhou, T. (2010). Link prediction in complex networks: A survey. *Physica A*, abs/1010.0725(6):1150–1170.

Lu, Z., Savas, B., Tang, W., and Dhillon, I. (2010). Supervised link prediction using multiple sources. In *Data Mining (ICDM), 2010 IEEE 10th Int. Conf. on*, pages 923–928.

Makrehchi, M. (2011). Social link recommendation by learning hidden topics. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 189–196, New York, NY, USA. ACM.

Maruyama, W. and Digiampietri, L. (2016). Predição de relacionamentos em redes sociais, uma revisão sistemática. In *CSBC2016 - BraSNAM*.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–9.

Pavlov, M. (2007). Finding experts by link prediction in co-authorship networks. *CEUR Workshop Proceedings*, 290:42–55.

Perez, C., Birregah, B., and Lemercier, M. (2012). The multi-layer imbrication for data leakage prevention from mobile devices. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th Int. Conf. on*, pages 813–819.

Quercia, D. and Capra, L. (2009). Friendsensing: Recommending friends using mobile phones. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 273–276, New York, NY, USA. ACM.

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., and Han, J. (2011). Co-author relationship prediction in heterogeneous bibliographic networks. In *IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining*, pages 121–128.

Sun, Y., Han, J., Aggarwal, C. C., and Chawla, N. V. (2012). When will it happen?: relationship prediction in heterogeneous information networks. In *Proceedings of the fifth ACM Int. Conf. on Web Search and Data Mining*, WSDM '12, pages 663–672, New York, NY, USA. ACM.

Tian, Y., He, Q., Zhao, Q., Liu, X., and Lee, W. (2010). Boosting social network connectivity with link revival. In *Proceedings of the 19th ACM Int. Conf. on Information and Knowledge Management*, CIKM '10, pages 589–598, New York, NY, USA. ACM.

Vasuki, V., Natarajan, N., Lu, Z., and Dhillon, I. S. (2010). Affiliation recommendation using auxiliary networks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 103–110, New York, NY, USA. ACM.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the evolution of user interaction in facebook. *Proceedings of the 2nd ACM workshop on Online social networks - WOSN '09*, page 37.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Zhong, E., Fan, W., Zhu, Y., and Yang, Q. (2013). Modeling the dynamics of composite social networks. In *Proceedings of the 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD '13, pages 937–945, New York, NY, USA. ACM.