Predição de Coautorias em Redes Sociais Acadêmicas: Um Estudo Exploratório em Ciência da Computação

Luciano A. Digiampietri[1], **Caio R. N. Santiago**[1], **Caio M. Alves**[1]

[1]Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

{digiampietri,caio.santiago,caio.margutti.alves}@usp.br

Abstract. *The identification of future relationships in a social network is a relevant subject which is receiving, in the last years, attention from both, the scientific community and the industry, because it allows the prediction of the networking behavior or the recommendation of new relationships in order to optimize the network evolution. This paper presents a methodology and the results of the use of attribute selectors, filters, classifiers and functions to predict coauthorships in academic social networks in order to evaluate the performance of different algorithms and to identify some of the most relevant attributes for this prediction.*

Resumo. *A identificação de possíveis relacionamentos futuros em uma rede social é um assunto relevante e que tem recebido bastante atenção da comunidade científica e da indústria nos últimos anos, pois pode permitir tanto a previsão do comportamento de uma rede como possibilitar a sugestão de novos relacionamentos de forma a otimizar a evolução da rede. Este artigo apresenta uma metodologia e os resultados da aplicação de seletores de atributos, filtros, classificadores e funções preditivas para a predição de coautorias em redes sociais acadêmicas com o objetivo de avaliar o desempenho de diferentes algoritmos e de se identificar alguns dos atributos mais relevantes para esta predição.*

1. Introdução

Cada vez mais as atividades humanas estão sendo analisadas através de redes sociais de interações entre indivíduos. Esta análise pode ter diversos objetivos, por exemplo identificar possíveis interesses de um consumidor com base no padrão de consumo de seus amigos; identificar o fluxo de conhecimento dentro de uma empresa a fim de otimizar o processo produtivo; e sugerir pessoas com quem um indivíduo provavelmente gostará de se relacionar.

O problema de predição de relacionamentos é um problema complexo por requerer a identificação dos melhores conjuntos de atributos relevantes à predição (dentre as dezenas, centenas ou milhares de combinações possíveis de atributos), tratar conjuntos de dados tipicamente muito desbalanceados (isto é, dado um par arbitrário de pessoas há uma probabilidade muito grande de que elas não irão se relacionar) e por envolver diversas técnicas estatísticas ou de inteligência artificial onde cada técnica poderá apresentar melhores resultados de acordo com o domínio onde ela foi aplicada.

Recentemente, diversos estudos estão propondo técnicas para a predição de relacionamentos usando diferentes algoritmos e conjuntos de atributos de entrada [Bartal et al. 2009, Hoseini et al. 2012, Sun et al. 2011, Sun et al. 2012,

Narayanan et al. 2011]. O objetivo deste artigo é apresentar um estudo exploratório onde foram testados dezenas de algoritmos de classificação a fim de prever relações de coautoria em redes sociais acadêmicas. Além disto, este artigo contém uma metodologia para filtragem dos dados e seleção dos atributos relevantes que pode ser utilizada para diferentes sistemas de predição e que foi aplicada neste trabalho em um estudo de caso realizado sobre os dados extraídos dos currículos Lattes de docentes permanentes de programas de pós-graduação em Ciência da Computação.

2. Metodologia

O trabalho apresentado neste artigo foi dividido em oito atividades: seleção da amostra; obtenção e armazenamento dos dados; identificação das informações relevantes; seleção dos atributos; filtragem dos dados; montagem dos conjuntos de treinamento e teste; execução dos testes; e análise dos resultados.

Seleção da amostra. Como amostra, optou-se por utilizar os currículos Lattes dos pesquisadores permanentes dos programas de pós-graduação em ciência da computação com doutorado e/ou mestrado acadêmico que foram avaliados nos triênios 2004-2006 e 2007-2009 pela CAPES. Foram selecionados apenas os 657 pesquisadores que atuaram em ambos triênios. Esta escolha foi feita pelo fato de os pesquisadores permanentes tipicamente terem uma alta produtividade e por se acreditar que ocorram diversas colaborações entre os docentes deste grupo. Além disso, diversos destes pesquisadores orientaram ou foram orientadores de outros pesquisadores deste grupo de forma que este conjunto de dados possibilita a identificação de outros relacionamentos além dos de coautoria. Também fizeram parte do conjunto de dados os currículos de todos os pesquisadores relacionados a estes 657 currículos, pois uma das medidas utilizadas para a predição de relacionamentos é a quantidade de vizinhos em comum e, para isto, se fez necessária a identificação de outros 11.308 currículos como será visto adiante. Neste trabalho foram consideradas as produções e orientações de 1971 a 2010. O período de 2001 a 2005 foi considerado o período atual, os dados até o ano 2000 foram considerados dados anteriores e a predição pretende prever as relações de coautoria ocorridas de 2006 a 2010.

Obtenção e armazenamento dos dados. Os currículos foram baixados através da Web no formato HTML diretamente do site do CNPq utilizando-se a ferramenta *wget*. O conteúdo de cada arquivo HTML foi tabulado e convertido para XML utilizando *parsers* e o conjunto de arquivos XML foi utilizado para carregar um banco de dados relacional. Todo este processo utilizou as ferramentas desenvolvidas e disponibilizadas por Digiampietri e outros [Digiampietri et al. 2012a]. Dentro deste banco de dados, as produções bibliográficas já estão discriminadas por tipo (artigo completo publicado em anais; artigos publicados em periódicos, etc), assim como as orientações (mestrado, doutorado, iniciação científica, etc).

Identificação das informações relevantes. As informações relevantes para o estudo realizado neste artigo são: identificação de todos os currículos relacionados a cada orientador (incluindo coautores, orientandos, orientadores, coparticipantes em bancas e em projetos de pesquisa); identificação dos orientadores e orientandos; identificação dos artigos completos publicados em anais de congressos e em periódicos; e identificação das coautorias. Estas informações servirão de base para a execução dos algoritmos de seleção de atributos, classificação e regressão que serão utilizados para a predição de relaciona-

mentos e identificação das características mais importantes para esta predição.

Para a identificação dos currículos relacionados a cada pesquisador foram utilizadas as relações explícitas existentes em cada currículo, isto é, os *links* HTML existentes dentro dos currículos Lattes para indicar o currículo de um coautor, coparticipante de um projeto, coparticipante de uma banca, orientador ou orientando. Uma ferramenta foi desenvolvida para identificar os diferentes *links* existentes nos arquivos HTML dos currículos dos 657 pesquisadores. Este procedimento identificou 11.308 novos currículos que foram baixados utilizando o mesmo procedimento apresentado na etapa *obtenção e armazenamento dos dados*.

Seleção de atributos. Treze atributos foram considerados para a predição de relações, sendo dois relacionados à classe e os demais utilizados para o treinamento. Neste trabalho apenas duas classes serão consideradas: "serão coautores" e "não serão coautores" e estas serão aplicadas a cada par de pesquisadores da amostra selecionada. A Tabela 1 contém a descrição dos treze atributos utilizados.

Tabela 1. Atributos utilizados

	Nome do atributo	Descrição
1	coautorias a serem preditas	Quantidade de artigos completos publicados em coautoria pelo par de pesquisadores em análise em conferências ou em periódicos no período de 2006 a 2010.
2	classe	Atributo que assume o valor "serão coautores" caso o atributo "coautorias a serem preditas" seja maior que zero e, caso contrário, "não serão coautores".
3	periódicos anterior	Quantidade de artigos publicados em periódicos em coautoria pelo par de pesquisadores antes do ano 2001.
4	conferências anterior	Quantidade de artigos completos publicados em conferências em coautoria pelo par de pesquisadores antes do ano 2001.
5	periódicos atual	Quantidade de artigos publicados em periódicos em coautoria pelo par de pesquisadores entre os anos 2001 e 2005.
6	conferências atual	Quantidade de artigos completos publicados em conferências em coautoria pelo par de pesquisadores entre os anos 2001 e 2005.
7	orientação anterior	Atributo que recebe o valor 1 (um) caso um dos pesquisadores tenha sido orientador do outro antes do ano 2001, ou 0 (zero) caso contrário.
8	orientação atual	Atributo que recebe o valor 1 (um) caso um dos pesquisadores tenha sido orientador do outro entre os anos 2001 e 2005, ou 0 (zero) caso contrário.
9	orientação em andamento	Atributo que recebe o valor 1 (um) caso um dos pesquisadores seja orientador, em uma orientação em andamento, ao final do ano 2005, ou 0 (zero) caso contrário.
10	orientadores em comum	Quantidade de orientadores e coorientadores que foram orientadores dos dois pesquisadores em análise.
11	orientandos em comum	Quantidade de orientandos e coorientandos que foram orientados pelos dois pesquisadores em análise.
12	vizinhos em comum	Quantidade de vizinhos em comum entre os dois pesquisadores na rede de coautorias formada por produções bibliográficas publicadas de 2001 até 2005.
13	programas em comum	Atributo que recebe o valor 1 (um) caso os dois pesquisadores pertençam ao mesmo programa de pós-graduação, ou 0 (zero) caso contrário.

Para a verificação se dois artigos presentes em diferentes currículos correspondem a uma única publicação (que foi publicada pelos pesquisadores possuido-

res destes dois currículos) utilizou-se a metodologia de resolução de entidades proposta em [Digiampietri et al. 2012b], específica para tratar publicações cadastradas em currículos Lattes. Para verificar se dois pesquisadores possuem uma relação de orientação (um é ou foi orientador do outro) ou mesmo para se saber quantos (co)orientandos em comum dois pesquisadores possuem utilizou-se o algoritmo de normalização de nomes presentes em registros bibliográficos proposto por Mugnaini e outros [Mugnaini et al. 2012].

Filtragem (horizontal) dos dados. A montagem do conjunto de treinamento envolve combinar os 657 pesquisadores da amostra dois a dois e extrair os atributos selecionados para cada par de pesquisadores. Porém, esta combinação produzirá 215.496 pares o que corresponde a um volume de dados grande para a maioria dos classificadores. Além disso, destes pares poucos serão coautores no período analisado. Desta forma, será feita uma filtragem horizontal dos dados (isto é, serão excluídos pares de pesquisadores) antes do processo de treinamento, seguindo o critério de se excluir todos os pares cujos 11 atributos de treinamento tem valor igual a zero. Na Seção 3 são apresentados os detalhes e as consequências do uso desta filtragem.

Montagem dos conjuntos de treinamento e teste. Diferentes conjuntos de treinamento serão montados de acordo com uma filtragem vertical dos dados, isto é, da exclusão ou não de um ou mais atributos dos pares selecionados. A montagem destes conjuntos será feita com base no resultado da execução de seletores de atributos e da análise da correlação entre os 11 atributos de treinamento e o atributo classe. A execução de testes utilizando os diferentes conjuntos formados permitirá a identificação dos atributos mais importantes para a predição, bem como, indicará se algum dos atributos utilizados não está ajudando na predição.

Execução dos testes. Neste artigo, todos os testes realizados utilizaram versões já implementadas de algoritmos de classificação disponíveis através do arcabouço *Weka* [Hall et al. 2009] que contém dezenas de implementações em Java de algoritmos para a seleção de atributos, classificação, agrupamento e identificação de regras de associação. Serão testados utilizando-se os 11 atributos de treinamento todos os algoritmos de classificação disponíveis que classifiquem conjuntos de dados cujo atributo classe seja não numérico. O desempenho dos algoritmos será medido pela taxa média de acertos utilizando-se *10-fold cross-validation*. Os dez algoritmos com melhor desempenho serão utilizados para os demais testes utilizando os diferentes (sub)conjuntos de treinamento. Além disso, serão executados testes específicos para verificar a predição de coautorias entre dois pesquisadores que nunca haviam colaborado em publicações em períodos anteriores. Por fim, serão utilizados algoritmos que estimam valores de atributos para tentar prever a quantidade de artigos que será publicada por um par de pesquisadores que se enquadre na classe "serão coautores".

Análise dos resultados. Com base nos dados produzidos durante a execução dos testes será possível comparar o desempenho dos classificadores, os resultados dos seletores de atributos e da análise da correlação dos dados, e o resultado dos algoritmos para estimar a quantidade de coautorias entre dois pesquisadores. Estes resultados, bem como detalhes das filtragens horizontais e verticais serão apresentados na próxima seção.

3. Execução dos testes e análise dos resultados

Ao se combinar os 657 pesquisadores dois a dois obtém-se um total de 215.496 pares diferentes de pesquisadores. Destes, 804 publicaram um ou mais artigos completos em anais ou periódicos no período de 2006 a 2010, correspondendo a menos de 0,3731% do total de pares. Desta forma, dado um par arbitrário de pesquisadores, qualquer classificador que sempre prediga que os dois pesquisadores não possuirão um relacionamento de coautoria no período terá mais de 99,6% de chance de acertar. Esta grande diferença entre o número de pares que possui um relacionamento (a ser previsto) e os que não possuem é um desafio aos classificadores pois, tipicamente, é muito difícil obter resultados melhores do que os 99,6% de apenas se classificar todos os pares como "não serão coautores". Por outro lado, utilizar como treinamento um conjunto contendo, por exemplo, 50% de pares que não possuem relacionamento no período e 50% de pares que possuem criará um classificador com um grande viés e que não poderá ser aplicado em situações reais.

Para obter resultados aplicados a situações reais sem a necessidade de processar os 215.496 pares optou-se por fazer um filtro nos dados de treinamento, porém sem utilizar o atributo classe ("serão coautores" e "não serão coautores") no filtro, de forma que os resultados do classificador possam ser aplicados em situações reais que se enquadrem neste filtro (porém, os elementos de fora deste filtro ou não seriam classificados ou precisariam de um classificador específico). Para este filtro, optou por selecionar apenas os pares que possuíam ao menos um dos 11 atributos selecionados com valor diferente de zero, ou seja, exclui-se todos os pares de pesquisadores que não tinham nenhuma das características em comum (nunca colaboraram em um artigo, não estão no mesmo programa de pós-graduação, não possuem vizinhos em comum e assim por diante). Após a execução deste filtro, restaram 11.800 dos 215.496 e, coincidentemente, este conjunto contém todos os 804 pares rotulados como "serão coautores" no período de 2006 a 2010. Em outras palavras, dado o conjunto de dados original todos os pares cujos 11 atributos analisados possuem valor zero no período anterior a 2006 não foram coautores entre 2006 e 2010. Desta forma, não será necessário o desenvolvimento de um classificador específico para o conjunto de dados excluído pelo filtro, pois a predição para este conjunto será sempre "não serão coautores". Assim, 94,52% dos dados já estariam classificados corretamente e, no restante deste artigo, trabalharemos com a classificação dos demais 5,48%.

Vale destacar que na utilização de uma filtragem inicial é comum que ocorra a eliminação de alguns registros da classe de interesse (por exemplo, eliminar 20 pares da classe "serão coautores"). Mesmo que isto tivesse acontecido neste trabalho, a perda de precisão seria baixa pois, já que o filtro eliminou mais de 200.000 pares, mesmo que 20 pertencessem a classe "serão coautores" a classificação dos mais de 200.000 como "não serão coautores" ainda teria uma taxa de acerto superior a 99,99%.

Conforme apresentado, dos 11.800 pares selecionados pelo filtro, 804 são coautores no período de 2006 a 2010, ou seja, cerca de 93,19% não serão. Assim, esta porcentagem será utilizada como limite inferior para os classificadores testados.

A Tabela 2 apresenta os resultados de todos os 71 classificadores disponíveis pelo arcabouço Weka para a classificação de dados utilizando classes não numéricas (neste caso, "serão coautores" e "não serão coautores"). Os classificadores estão ordenados pela taxa de acerto e além do nome de cada classificador também é apresentado o tipo. A esquerda são apresentados os 45 classificadores que obtiveram resultados melhores do

que o valor de referência enquanto que a direita estão os que obtiveram resultados piores.

Tabela 2. Classificadores utilizados e taxa de acerto

Tipo	Nome	Taxa de acerto
meta	Bagging	94,689%
meta	EnsembleSelection	94,6526%
meta	RotationForest	94,6162%
trees	FT	94,5526%
meta	Decorate	94,5344%
trees	LMT	94,5071%
trees	J48graft	94,4707%
trees	J48	94,4616%
meta	OrdinalClassClassifier	94,4616%
meta	nestedDichotomies	94,4616%
meta	nestedDichotomiesDataNearBalancedND	94,4616%
meta	nestedDichotomiesClassBalancedND	94,4616%
meta	END	94,4616%
rules	PART	94,3798%
trees	RandomForest	94,3434%
trees	RandomTree	94,2525%
meta	RandomCommittee	94,2434%
trees	REPTree	94,207%
meta	ClassificationViaRegression	94,1433%
meta	LogitBoost	93,916%
rules	Ridor	93,8887%
trees	ADTree	93,8068%
bayes	BayesianLogisticRegression	93,8068%
trees	LADTree	93,7977%
functions	MultilayerPerceptron	93,7977%
meta	Vote	93,7887%
meta	MultiClassClassifier	93,7796%
meta	FilteredClassifier	93,7796%
functions	Logistic	93,7796%
trees	SimpleCart	93,7705%
functions	SimpleLogistic	93,7705%
meta	RandomSubSpace	93,7523%
trees	NBTree	93,7341%
meta	AttributeSelectedClassifier	93,725%
rules	DecisionTable	93,7159%
rules	JRip	93,6704%
meta	RacedIncrementalLogitBoost	93,5431%
trees	BFTree	93,534%
lazy	KStar	93,4613%
bayes	DMNBtext	93,4613%
functions	VotedPerceptron	93,4249%
rules	DTNB	93,3703%
functions	SMO	93,3612%
lazy	LWL	93,2703%
meta	Dagging	93,1339%
bayes	NaiveBayesMultinomialUpdateable	93,1339%
bayes	NaiveBayesMultinomial	93,1339%
trees	DecisionStump	93,1157%
meta	MultiBoostAB	93,1157%
meta	AdaBoostM1	93,1157%
lazy	IB1	93,0975%
rules	ConjunctiveRule	93,0156%
misc	FLR	92,8792%
misc	HyperPipes	92,8701%
bayes	NaiveBayesUpdateable	92,8429%
bayes	NaiveBayes	92,8429%
bayes	BayesNet	92,7337%
rules	ZeroR	92,6883%
meta	Vote	92,6883%
meta	Vote	92,6883%
meta	StackingC	92,6883%
meta	Stacking	92,6883%
meta	MultiScheme	92,6883%
meta	Grading	92,6883%
meta	CVParameterSelection	92,6883%
functions	RBFNetwork	92,6883%
meta	ThresholdSelector	92,5064%
bayes	NaiveBayesSimple	91,0877%
bayes	ComplementNaiveBayes	78,9924%
meta	ClassificationViaClustering	68,9069%
misc	VFI	60,5493%

Dentre os dez classificadores com maiores taxas de acertos estão apenas meta classificadores e classificadores que usam árvores. Estes dez classificadores foram escolhidos para serem utilizados nos demais experimentos apresentados a seguir.

Duas estratégias foram utilizadas a fim de se selecionar conjuntos de atributos a serem utilizados nos testes. A primeira foi o uso de seletores de atributos para identificar quais subconjuntos de atributos são mais relevantes para representar o conjunto de dados. A segunda foi o cálculo da correlação entre o atributo classe e os demais atributos. A Tabela 3 contém o resultado da execução de todos os seletores de atributos disponíveis no Weka que retornaram subconjuntos não vazios. Alguns dos seletores ordenam os atributos de acordo com sua importância, para estes, o valor da ordem de seleção foi colocado na tabela (onde 1 representa o atributo mais importante/informativo, 2 o segundo mais importante/informativo e assim por diante). Outros seletores apenas indicam quais atributos

foram selecionados, para estes o valor "x" foi colocado na tabela. É possível observar que o atributo com maior importância é o "conferências atual". Além deste, outros atributos que se destacam são periódicos atual e vizinhos em comum. Por outro lado, o atributo com menor importância foi o "orientadores em comum". Vale ressaltar que o conjunto de dados é formado por docentes permanentes de programas de pós-graduação, desta forma, não é de se estranhar que o atributo "orientadores em comum" não seja muito informativo.

Tabela 3. Seleção de atributos

	periódicos anterior	conferências anterior	periódicos atual	conferências atual	orientação anterior	orientação atual	orientação em andamento	orientadores em comum	orientandos em comum	vizinhos em comum	programas em comum
FilteredAttributeEval	6	5	3	1	7	9	8	11	4	2	10
ChiSquaredAttributeEval	6	5	3	1	8	9	7	11	4	2	10
GainRatioAttributeEval	4	5	2	1	8	7	3	11	9	6	10
FilteredAttributeEval	6	5	3	1	7	9	8	11	4	2	10
ConsistencySubsetEval	x	x	x	x	x	x	x		x	x	x
CfsSubsetEval			x	x							
SpreadSubsample			x	x							

A Figura 1 contém a correlação entre a classe e os demais onze atributos. Para a identificação dos atributos mais relevantes será utilizada apenas a primeira linha da matriz de correlações, as demais linhas servem apenas para ilustrar algumas características do conjunto de dados. Para o cálculo das correlações, o valor do atributo classe "não serão coautores" foi substituído por 0 (zero) e o valor "serão coautores" foi substituído por 1 (um). Destaca-se que os maiores valores de correlação com o atributo classe ocorrem com os atributos "vizinhos em comum" (0,35), "conferências atual" (0,34) e "periódicos atual" (0,30). Também é possível observar que estes três atributos estão fortemente correlacionados entre si.

Com base nos seletores e na correlação dos dados nove conjuntos de dados foram criados para se comparar os resultados dos classificadores. Estes conjuntos foram criados tanto para mostrar a quantidade de informação relevante para a classificação presente em um ou mais atributos como para testar conjuntos de atributos que não contém informações sobre as coautorias atuais e anteriores, a fim de identificar o quão dependente desta informação a classificação é.

Tabela 4. Conjuntos de atributos utilizados

	c1	c2	c3	c4	c5	c6	c7	c8	c9
periódicos anterior	x					x	x		
conferências anterior	x					x	x		
periódicos atual	x				x	x	x		
conferências atual	x		x	x	x	x	x		
orientação anterior	x							x	
orientação atual	x							x	
orientação em andamento	x						x	x	
orientadores em comum	x							x	
orientandos em comum	x					x	x	x	x
vizinhos em comum	x	x		x	x	x	x	x	x
programas em comum	x							x	

Figura 1. Correlação entre todos os atributos para os 11.800 pares

A Tabela 5 contém os resultados dos classificadores selecionados aplicados a cada conjunto de dados. Vale lembrar que se um classificador classificasse todos os pares como "não serão coautores" ele teria uma taxa de acerto de 93,19% (valor a ser utilizado como referência). O conjunto de dados c1 contém todos os atributos e é o conjunto que possibilitou a melhor classificação. Os conjuntos c2 e c3 contêm apenas um atributo cada ("vizinhos em comum" e "conferências atual", respectivamente), conforme indicado pelos dados de correlação o conjunto c2 apresentou melhores resultados do que o c3, porém, os resultados ficaram muito próximos ao valor de referência. O conjunto c4 contém os atributos dos conjuntos c2 e c3 e os resultados da classificação já foram superiores ao valor de referência. O conjunto c5 contém os atributos "vizinhos em comum", "conferências atual" e "periódicos atual" e os resultados da classificação são cerca de 0,5% melhores do que os resultados utilizando c4. Os conjuntos c6 e c7 contém alguns atributos adicionais em relação a c5 (escolhidos de acordo com os seletores de atributos) e seus resultados são um pouco melhores do que os anteriores. O conjunto c8 contém todos os atributos exceto o número de coautorias atuais e anteriores. Os resultados utilizando-se este conjunto são melhores do que o valor de referência mas estão cerca de 0,8% abaixo do resultado utilizando-se todos os atributos. Por fim, o conjunto c9 contém os atributos "vizinhos em comum" e "orientandos atuais" que são os dois atributos com maior correlação com o

atributo classe (excluindo-se os atributos de coautorias atuais e anteriores). Os resultados dos classificadores para o conjunto c9 é cerca de 0,2% melhor do que o valor de referência.

Além dos conjuntos apresentados foram criados outros onze conjuntos de dados, cada um com dez dos onze atributos, com a finalidade de se verificar se é possível manter ou melhorar o desempenho dos classificadores retirando-se um atributo (ou seja, verificar se um dos atributos é redundante ou, até mesmo, se este atributo estaria atrapalhando a classificação). Como resultado foi verificado que o desempenho de cada um destes conjuntos foi pior do que o desempenho dos classificadores utilizando-se todos os atributos. Em outras palavras, todos os atributos estão ajudando na classificação.

Tabela 5. Taxa de acerto dos classificadores para cada conjunto de atributos

	c1	c2	c3	c4	c5	c6	c7	c8	c9
Bagging	94,69%	93,16%	92,63%	93,28%	93,79%	93,78%	93,86%	93,83%	93,45%
EnsembleSelection	94,65%	93,18%	92,72%	93,26%	93,76%	93,79%	93,74%	93,75%	93,41%
RotationForest	94,62%	93,21%	92,63%	93,19%	93,78%	93,82%	93,84%	93,63%	93,28%
FT	94,55%	93,21%	92,63%	93,28%	93,78%	93,89%	93,78%	93,61%	93,47%
Decorate	94,53%	93,19%	92,63%	93,34%	93,71%	93,70%	93,75%	93,55%	93,42%
LMT	94,51%	93,16%	92,63%	93,30%	93,76%	93,82%	93,74%	93,48%	93,44%
J48graft	94,47%	93,21%	92,63%	93,31%	93,70%	93,69%	93,73%	93,60%	93,42%
nestedDichotomies.ND	94,46%	93,21%	92,63%	93,31%	93,70%	93,68%	93,72%	93,62%	93,41%
OrdinalClassClassifier	94,46%	93,21%	92,63%	93,31%	93,70%	93,68%	93,72%	93,62%	93,41%
J48	94,46%	93,21%	92,63%	93,31%	93,70%	93,68%	93,72%	93,62%	93,41%

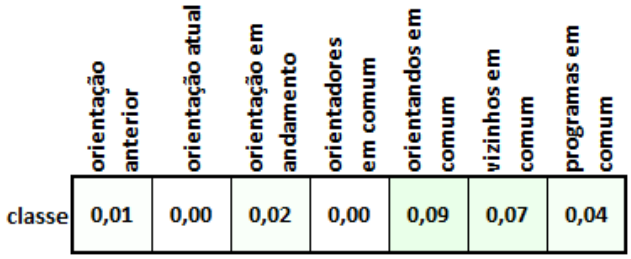
Conforme pôde ser observado na Figura 1 e nas Tabelas 3 e 5 a quantidade de coautorias atuais e anteriores tem grande influência no resultado da classificação. Isto motivou a investigação da classificação sem estes atributos e, mais especificamente, da identificação de coautorias entre pesquisadores que nunca haviam colaborado em nenhum artigo completo em anais ou em periódicos. Este conjunto possui as potenciais "novas coautorias".

Dos 11.800 pares de pesquisadores que possuem valores diferentes de zero em ao menos um dos onze atributos, 9.999 nunca tiveram uma coautoria atual (de 2001 a 2005) ou anterior ao ano 2001. Destes pares 347 foram coautores entre 2006 e 2010, ou seja, devem ser classificados como "serão coautores" (cerca de 3,47% de 9.999). Assim, cerca de 96,53% dos pares não serão coautores e esta porcentagem será utilizada como valor de referência aqui.

Os classificadores selecionados foram testados para os nove conjuntos de atributos para este novo conjunto de dados, porém nenhum deles obteve resultados melhores do que 96,53% indicando que nenhum classificador teve taxa de acerto superior a um classificador que classificasse todos os pares como "não serão coautores".

A Figura 2 contém a correlação entre o atributo classe e os demais considerando-se apenas os 9.999 pares de pesquisadores que não foram coautores antes de 2006. Os atributos de coautoria (atual e anterior) não aparecem nesta figura pois seus valores eram todos zero. É possível observar que não há nenhuma correlação relevante entre o atributo classe e os demais atributos. Isto explica, em partes, o fato de não ter sido possível obter um classificador melhor do que o valor de referência.

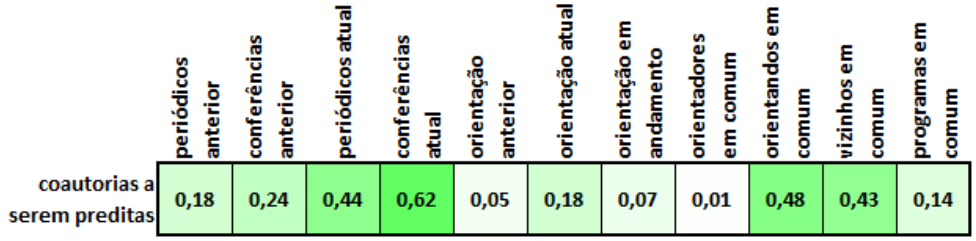
A última tarefa realizada neste estudo foi uma investigação das características dos 804 pares de pesquisadores cuja classificação correta é "serão coautores". Na média cada par colaborou em 3,57 artigos (com desvio padrão de 5,79 e mediana igual a 2).



	orientação anterior	orientação atual	orientação em andamento	orientadores em comum	orientandos em comum	vizinhos em comum	programas em comum
classe	0,01	0,00	0,02	0,00	0,09	0,07	0,04

Figura 2. Correlação dos atributos para o conjunto de dados "novas coautorias"

Foram identificadas as correlações entre a quantidade de artigos que estes pesquisadores publicaram conjuntamente no período de 2006 a 2010 e os demais atributos (Figura 3) e foram executados algoritmos para criar funções que calculem (estimem) a quantidade de "coautorias a serem preditas" tendo como parâmetros os demais atributos.



	periódicos anterior	conferências anterior	periódicos atual	conferências atual	orientação anterior	orientação atual	orientação em andamento	orientadores em comum	orientandos em comum	vizinhos em comum	programas em comum
coautorias a serem preditas	0,18	0,24	0,44	0,62	0,05	0,18	0,07	0,01	0,48	0,43	0,14

Figura 3. Correlação utilizando os 804 pares da classe "serão coautores"

Os algoritmos para a estimativa de valores costumam ser avaliados usando 5 medidas: coeficiente de correlação, erro absoluto médio, erro quadrático médio, erro absoluto relativo e erro quadrático relativo. Estas medidas avaliam a diferença entre o valor predito e o valor correto e os valores dos erros, em muitos casos, são maiores do que 1. A Tabela 6 apresenta o resultado de nove destes algoritmos (as implementações utilizadas foram as disponíveis no Weka). Destaca-se a correlação e o erro absoluto relativo obtidos pelo algoritmo *SMOreg*, e o erro quadrático relativo obtido por *Gaussian Processes*.

Tabela 6. Algoritmos de regressão e resultados

	Gaussian Processes	Isotonic Regression	Least MedSq	Linear Regression	Multilayer Perceptron	Pace Regression	RBF Network	Simple Linear Regression	SMOreg
Correlation coefficient	0,64	0,62	0,20	0,61	0,37	0,63	0,24	0,61	0,64
Mean absolute error	2,34	2,45	2,51	2,42	3,66	2,39	2,93	2,44	2,10
Root mean squared error	4,47	4,54	6,11	4,58	9,50	4,52	5,62	4,60	4,77
Relative absolute error	75,62%	79,03%	80,93%	77,98%	118,18%	77,20%	94,45%	78,67%	67,83%
Root relative squared error	77,00%	78,32%	105,30%	79,00%	163,79%	77,98%	96,93%	79,28%	82,15%

A maioria dos algoritmos utilizados criou uma função de estimativa do número de "coautorias a serem preditas" utilizando os onze atributos disponíveis, porém, o algoritmo *SimpleLinearRegression* criou uma função utilizando apenas o atributo "conferências atual": $coautorias_a_serem_preditas = 0,76 * conferencias_atual + 2,12$, ou seja, o número de coautorias entre os anos 2006 e 2010 pode ser estimado como sendo 0,76 vezes o número de coautorias em conferências no período de 2001 a 2005 mais 2,12.

4. Trabalhos Correlatos

Bartal et al [Bartal et al. 2009] combinaram técnicas de análise de redes sociais com mineração de texto para predizer coautorias em publicações acadêmicas. Eles combinaram treze atributos (um relacionado à mineração de texto e os demais relacionados à análise de redes sociais) e atingiram uma taxa de acerto máxima de 91% em seus testes utilizando dados da base bibliográfica DBLP[1].

Hoseini et al [Hoseini et al. 2012] utilizaram as técnicas de clusterização e co-clusterização para a predição de relacionamentos de coautoria na base DBLP. Nesse trabalho o menor erro quadrático médio (medida utilizada pelos autores para avaliar a predição) foi de 0,2384. Apenas para permitir a comparação desse estudo com os resultados presentes no trabalho atual, o erro quadrático médio do classificador *Bagging* sobre os 11.800 pares de currículos selecionados foi de 0,2152 e, conforme visto na Tabela 5, neste exemplo foi obtida uma taxa de acerto de cerca de 94,69%.

Sun et al [Sun et al. 2012] desenvolveram um algoritmo que além de tentar prever links em redes heterogêneas (isto é, redes onde existem arestas de diferentes tipos) também tenta identificar quando esses novos links ocorrerão. Seus resultados tiveram taxas de acerto inferiores a 80%. Em outro trabalho [Sun et al. 2011], Sun et al focaram na predição de links em redes bibliográficas usando diferentes métricas de vizinhança e de caminhos em grafos. Eles utilizaram a base DBLP e obtiveram uma taxa de acerto máxima em torno de 75%.

Narayanan et al [Narayanan et al. 2011] ganharam o *Kaggle Social Network Challenge*, um desafio de implementação de algoritmos para a predição de links e identificação de elementos em redes sociais, utilizando uma combinação de florestas de rotação com casamento de grafos ponderados. A competição utilizou dados do site de compartilhamento de fotos Flickr[2] e eles foram capazes de identificar 64,7% dos elementos.

5. Conclusões

Este artigo apresentou e aplicou uma metodologia para a predição de coautorias em redes sociais acadêmicas. Para isto foram utilizados dados da Plataforma Lattes, onde além das informações sobre publicações também estão disponíveis informações sobre orientações, áreas de atuação e vínculos empregatícios. Esta variedade de informações possibilita o desenvolvimento de uma grande gama de sistemas de predição.

Para a predição do relacionamento de coautoria foram testados diferentes classificadores e a taxa de acerto atingida pelo classificador de melhor desempenho (correspondendo à combinação dos 94,689% de acerto sobre o conjunto de 11.800 pares e ao acerto de 100% para os 203.696 pares eliminados pelo filtro) foi de 99,709% onde 311 pares foram corretamente classificados como "serão coautores" e 213.797 foram corretamente classificados como "não serão coautores", além disto, 91 pares foram incorretamente classificados como "serão coautores" e 493 pares foram incorretamente classificados como "não serão coautores". Apesar dos bons resultados, observou-se que não foi possível predizer novas coautorias quando foram utilizados apenas os dados dos pares de pesquisadores que nunca foram coautores no passado.

[1]http://www.informatik.uni-trier.de/ ley/db/

[2]http://www.flickr.com/

Para os pares de pesquisadores que são coautores foi possível calcular uma função para estimar a quantidade de coautorias que eles terão com uma taxa de erro médio absoluto de cerca de 2,1 publicações.

Como trabalhos futuros pretende-se identificar e analisar a influência de outros atributos na predição de coautorias em redes acadêmicas, bem como estudar outros filtros nos dados com o objetivo principal de predizer novas coautorias, isto é, coautorias entre pesquisadores que nunca haviam coautorado anteriormente.

Agradecimentos

O trabalho apresentado neste artigo foi parcialmente financiado pela FAPESP (Projeto Jovem Pesquisador processo 2009/10413-5), pelo CNPq (Bolsa de Iniciação Científica e Bolsa Produtividade em Pesquisa processo 304937/2010-0) e pelo Programa de Educação Tutorial (MEC/SESu).

Referências

Bartal, A., Sasson, E., and Ravid, G. (2009). Predicting links in social networks using text mining and sna. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, ASONAM '09, pages 131–136, Washington, DC, USA. IEEE Computer Society.

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*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Hoseini, E., Hashemi, S., and Hamzeh, A. (2012). Link prediction in social network using co-clustering based approach. In *26th International Conference on Advanced Information Networking and Applications Workshops*, pages 795–800.

Mugnaini, R., Digiampietri, L. A., de Oliveira, L. C., and Ferreira, S. M. S. P. (2012). Normalização de nomes de autores em fontes de informação institucionais: proposta de um método automático de verificação de erros. *Em Questão*, 18(3):263–279.

Narayanan, A., Shi, E., and Rubinstein, B. (2011). Link prediction by de-anonymization: How we won the kaggle social network challenge. In *International Joint Conference on Neural Networks*, pages 1825–1834.

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C., and Han, J. (2011). Co-author relationship prediction in heterogeneous bibliographic networks. In *IEEE/ACM International Conference 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 International Conference on Web Search and Data Mining*, WSDM '12, pages 663–672, New York, NY, USA. ACM.