Um Sistema de Predição de Relacionamentos em Redes Sociais

Alternative title: A Link Prediction System in Social Networks

Luciano Antonio Digiampietri
University of Sao Paulo
Av. Arlindo Bettio, 1000
Sao Paulo, SP, Brazil
digiampietri@usp.br

William Takahiro Maruyama
University of Sao Paulo
Av. Arlindo Bettio, 1000
Sao Paulo, SP, Brazil
takahiro.mw@gmail.com

Caio Rafael do Nascimento Santiago
University of Sao Paulo
Av. Arlindo Bettio, 1000
Sao Paulo, SP, Brazil
caio.rns@gmail.com

Jamison José da Silva Lima
University of Sao Paulo
Av. Arlindo Bettio, 1000
Sao Paulo, SP, Brazil
jamison.lima@gmail.com

RESUMO

Predizer novos relacionamentos dentro de um grupo social é uma tarefa complexa, porém extremamente útil para potencializar ou maximizar colaborações por meio da indicação de quais seriam as parcerias mais promissoras. Nas redes sociais acadêmicas, a predição de relacionamentos é tipicamente utilizada para tentar identificar potenciais parceiros no desenvolvimento de um projeto e/ou coautores para a publicação de um artigo. Este artigo apresenta um sistema que combina técnicas de inteligência artificial com o estado da arte das métricas de predição de relacionamentos em redes sociais. O sistema resultante foi testado usando dados reais de pesquisadores em Ciência da Computação e atingiu uma precisão superior a 99,5% na predição de novas coautorias.

Palavras-Chave

predição de relacionamentos, redes sociais, redes acadêmicas

ABSTRACT

The prediction of new relationships in a social network is a complex and extremely useful task to enhance or maximize collaborations by indicating what the most promising partnerships are. 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 a system that combines artificial intelligence techniques with the state-of-the-art metrics for link prediction. The resulting system was tested using real data from Computer Science researchers and achieved a precision above 99.5% in the coauthorship prediction.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Information retrieval

General Terms

Information Retrieval

Keywords

link prediction, social networks, academic networks

1. INTRODUÇÃO

A predição de novos relacionamentos dentro de uma rede social é uma tarefa que ganhou bastante destaque nos últimos anos [25, 21, 30, 20, 27, 6], pois serve para desde encontrar amigos que ainda não estavam ligados em uma rede social online [35, 34, 31, 15, 36, 32], até para potencializar a realização de trabalhos em empresas ou na academia [23, 14, 5].

Alguns dos fatores que tornam esta tarefa complexa são: a identificação de quais atributos individuais (relacionados, por exemplo, ao perfil ou currículo das pessoas) serão utilizados; especificação ou seleção de métricas estruturais de redes sociais a serem usadas; utilização de estratégias para combinar estes atributos e métricas de forma a possibilitar a predição; e uso correto do fator temporal para maximizar a precisão e a sensibilidade[1] da solução produzida.

Nas redes sociais acadêmicas, a predição de relacionamentos tem sido utilizada principalmente para a predição de coautorias [13, 7, 16, 18, 26, 28], atividade que indica se um

[1]A *sensibilidade* ou *revocação* corresponde a quantidade dos elementos corretamente classificados de forma positiva (verdadeiro-positivos) dividida pela quantidade total de elementos pertencentes a classe positiva.

par de pesquisadores poderá/deverá colaborar na produção de um artigo. Este tipo de predição pode, assim, otimizar a produção destes pesquisadores por meio da indicação de pesquisadores cujas parcerias são mais promissoras.

O problema de predição de relacionamentos, por sua vez, pode ser dividido em predição de relacionamentos novos/ inéditos (isto é, prever quais pares de pessoas que nunca se relacionaram numa rede social irão começar a se relacionar) ou no problema geral de predição de relacionamentos (predizer que pares de pessoas irão se relacionar independentemente delas já terem ou não se relacionado). Tipicamente, a expressão "predição de relacionamentos" (ou de *links*) se refere ao primeiro problema (predição de relacionamentos novos/inéditos).

Este artigo utiliza um sistema gerenciador de workflows científicos (*Scientific Workflow Management System (SWMS)*) para a construção de um sistema geral e modular de predição de relacionamentos em redes sociais. A entrada básica para o sistema é um arquivo contendo a lista de relacionamentos criados ao longo do tempo entre um conjunto de pessoas. Adicionalmente, foram desenvolvidos módulos específicos para a extração de informações de currículos Lattes de forma a se extrair informações da rede social acadêmica de um conjunto de pessoas e, com base nessas informações, realizar a predição de coautorias.

O restante deste artigo está organizado da seguinte maneira. A Seção 2 apresenta os trabalhos correlatos. A Seção 3 contém a descrição da metodologia utilizada. Na Seção 4 é apresentado um estudo de caso sobre predição de coautorias em redes sociais acadêmicas. Por fim, a Seção 5 contém as conclusões e os trabalhos futuros.

2. TRABALHOS CORRELATOS

Nos últimos anos diversos trabalhos foram publicados sobre a predição de relacionamentos. De um modo geral, estes trabalhos podem ser divididos em três grupos: aqueles que utilizam apenas características da rede social (ou mais especificamente, do grafo que representa a rede social)[13, 3, 14, 16]; aqueles que propõe a utilização de atributos (primitivos ou derivados) específicos do domínio no qual a predição irá ocorrer [35]; e sistemas híbridos que combinam estes dois aspectos [24, 7, 34, 18].

Ao se analisar os trabalhos que propõe a utilização de um ou mais atributos (ou mesmo a criação de atributos derivados) é possível observar uma gama muito grande e diversificada de atributos. Mesmo ao se restringir o domínio para predição de coautorias o número de correlatos ainda é elevado [13, 7, 16, 18, 26, 28, 5, 4].

A partir de uma revisão sistemática realizada sobre este tema, foram selecionados os atributos e as métricas que obtiveram os melhores desempenhos para serem aplicadas no problema de predição de novos relacionamentos. Estas métricas serão apresentadas na próxima seção.

Além disso, foi encontrado um trabalho específico sobre predição de coautorias usando dados de currículos Lattes [7]. Nesse trabalho, todos os atributos utilizados são oriundos de informações dos currículos da Plataforma Lattes. Esse trabalho analisou a predição de coautorias dos docentes dos programas brasileiros de pós-graduação em Ciência da Computação e utilizou os seguintes atributos (para cada par de docentes): artigos publicados em conjunto (em periódicos e em conferências); relação de orientação entre os docentes; existência de orientadores em comum; existência de orientandos em comum; se os dois pertencem ao mesmo programa de pós-graduação; e quantos vizinhos em comum eles possuem na rede de coautoria. Além disto, estes atributos foram divididos em períodos de tempo: os dados de 1971 a 2000 foram considerados passados; de 2001 a 2005, dados atuais e o objetivo dos autores era predizer as coautorias que ocorreram de 2006 a 2010. O problema de predição de relacionamentos foi abordado como um problema de classificação no qual, dado o vetor de atributos correspondente às informações de um par de docentes, diferentes algoritmos de inteligência artificial foram usados para classificar estes docentes em "serão coautores" ou "não serão coautores". Esta estratégia permitiu aos autores uma acurácia de 99,709% na predição, valor bastante expressivo, porém é importante observar, como destacado pelos autores, que dado um par arbitrário de docentes da amostra utilizada, há 99,6% de chance deles não serem coautores (um sistema que classificasse todos os pares como "não serão coautores" acertaria em 99,6% dos casos). O sistema não utilizou nenhuma estratégia para balanceamento de dados e, provavelmente por isto, a principal limitação da abordagem utilizada nesse trabalho foi sua incapacidade de predizer novas coautorias, isto é, o sistema se mostrou muito bom em identificar quais pares de docentes iriam publicar novamente juntos, mas ao tratar especificamente do problema de identificar quais serão os novos coautores a acurácia ficou abaixo daquela que seria obtida por um classificador que sempre classificasse os pares como "não serão coautores".

No presente artigo será apresentado um sistema de recomendação de relacionamentos em redes sociais utilizando uma abordagem robusta, por meio da combinação de diferentes atributos estruturais de rede (oriundos dos trabalhos correlatos analisados) e, para a realização do estudo de caso em redes sociais acadêmicas, também serão utilizados diversos atributos específicos desse domínio. Os diferentes atributos serão combinados utilizando-se técnicas de inteligência artificial a fim de se predizer tanto relacionamentos inéditos como a reincidência de relacionamentos. Tanto os atributos a serem utilizados como o algoritmo de inteligência artificial que irá combiná-los podem ser escolhidos pelo usuário. Adicionalmente, há um parâmetro específico para indicar se o usuário pretende prever apenas novos relacionamentos ou também prever a reincidência de relacionamentos (por exemplo, se um par de docentes que já colaborou na produção de um artigo irá colaborar novamente).

Neste trabalho a predição de relacionamentos também será abordada como um problema de classificação e para cada par de pessoas será atribuída uma das seguintes classes "possuirão relacionamento" ou "não possuirão relacionamento".

O sistema desenvolvido neste trabalho opera sobre um SWMS, isto permite que novos módulos sejam facilmente incorporados ao sistema além de possibilitar a execução paralela ou distribuída das diferentes atividades/módulos do workflow. Há diversos SWMSs disponíveis na literatura [1, 33, 2, 17]. Optou por utilizar uma das extensões do SWMS chamado WOODSS [29] pelos seguintes motivos: (i) o sistema foi desenvolvido no Brasil e possui seu código fonte disponível na linguagem de programação Java; (ii) o sistema permite que as atividades básicas (módulos) dos workflows sejam de diferentes tipos (por exemplo, serviços Web ou métodos na linguagem Java) [9]; (iii) o sistema já foi estendido para diferentes aplicações [11, 8]; e (iv) a versão utilizada do

SWMS possui um módulo de execução distribuída que pode utilizar computação voluntária (quando disponível), o que possibilita a execução eficiente dos workflows [12].

3. METODOLOGIA

O trabalho apresentado neste artigo foi organizado em quatro atividades: revisão da literatura correlata (sumarizada na Seção 2); especificação do sistema; desenvolvimento dos módulos; e realização de um estudo de caso.

3.1 Especificação do sistema

Conforme apresentado, o sistema de predição foi construído sobre um SWMS. Desta forma, o sistema se beneficiou de diversas características desse SWMS (tais como, execução paralela e distribuída, interface gráfica para construção, edição dos workflows e preenchimento de parâmetros das atividades). Desta forma, o sistema de predição precisará, minimamente, conter os módulos de processamento (atividades de workflow) necessários para a realização da predição. A Figura 3.1 apresenta a organização do sistema em seus módulos. Detalhes sobre a versão atual do SWMS bem como exemplos de workflows são apresentados na dissertação de mestrado de Caio Nascimento Santiago [12].

Figura 1: Módulos do Sistema Implementado

O sistema é composto de seis módulos, que serão definidos a seguir e terão algumas características detalhadas na próxima subseção. O módulo *Extrator de Atributos Estruturais* recebe como entrada um conjunto de triplas que indicam os relacionamentos existentes ao longo dos anos, no formato $< id1, id2, ano >$ (identificador da pessoa 1, da pessoa 2 e ano em que ocorreu o relacionamento entre estas pessoas) e extrai um conjunto de métricas/atributos estruturais de rede que podem ser utilizadas para a predição de relacionamentos em diferentes domínios (por exemplo, redes sociais de amizade, de trabalho, etc).

O módulo *Extrator de Atributos Específicos* foi concebido para processar dados de currículos Lattes. Ele recebe como parâmetro uma lista de identificadores de currículos e extrai diferentes características específicas a serem utilizadas na predição. Além disso, quando este módulo é utilizado ele produz as triplas utilizadas como entrada pelo módulo *Extrator de Atributos Estruturais* (que neste caso corresponderão às coautorias realizadas entre os pesquisadores possuidores dos currículos que estão sendo analisados). A maioria dos módulos recebe ainda três intervalos de tempo como parâmetros de entrada, que são chamados de *passado*, *presente* e *futuro*. O *futuro* indica o período de tempo que inicialmente será utilizado para a validação do treinamento do sistema de predição. Corresponde a um intervalo de tempo cujos dados de entrada são conhecidos, mas serão utilizados pelo *Preditor* durante o treinamento e validação dos resultados. O *presente* constitui o conjunto de dados mais recente exceto os dados do *futuro*. Os dados do *passado* são opcionais e visam a agregar informação que pode ser útil para a predição mas que não é muito recente. Esta separação entre *passado* e *presente* ou a utilização de ponderações diferentes entre dados mais ou menos recentes tende a melhorar a acurácia na predição de relacionamentos [34, 3].

O módulo *Concatenador de Atributos* é responsável por concatenar os diferentes atributos de entrada de forma a produzir um conjunto de dados padronizado a ser utilizado pelos próximos módulos. Este módulo recebe um conjunto de tuplas no formato $< id1, id2, atributo, valor >$ correspondendo ao valor de um dado atributo para o par de pessoas identificadas por $id1$ e $id2$ e gera uma matriz na qual cada linha corresponde a cada par de pessoas da amostra utilizada e cada coluna corresponde a um atributo.

O módulo *Filtro Horizontal de Dados* recebe a matriz de pares de pessoas (linhas) por atributos (colunas) e elimina as linhas que não tenham nenhum valor diferente de nulo (ou de acordo com outras condições detalhadas na próxima subseção). As linhas filtradas/excluídas tipicamente são rotuladas como "não possuirão relacionamento" e são excluídas por não agregarem informação ao *Preditor* que irá tentar predizer os relacionamentos para os demais pares de pessoas (isto é, as linhas restantes da matriz). Há uma função adicional deste módulo: preparar a matriz de dados para o problema de predição de (novos) relacionamentos ou para o problema que envolve também a predição da reincidência de relacionamentos. Um dos parâmetros passado pelo usuário determinará o problema que se deseja abordar e a diferença na execução deste módulo é a seguinte: para o caso de predição de (novos) relacionamentos o módulo também deverá excluir as linhas que correspondem a pares de pessoas que já se relacionam no *presente*.

O módulo *Balanceador da Amostra* recebe a matriz resultante da execução do *Filtro Horizontal de Dados* e produz como saída uma matriz com dados balanceados segundo o atributo classe (isto é, "possuirão relacionamento" ou "não possuirão relacionamento"). Este módulo é particularmente importante para o problema de predição de relacionamentos, pois em redes sociais a existência ou não de relacionamentos tipicamente produz um conjunto de dados bastante desbalanceado. Isto é, dado um par arbitrário de pessoas é muito mais provável que estas pessoas não irão se relacionar do que o contrário. Esse tipo de desbalanceamento, se não tratado, pode afetar significativamente o desempenho do *Preditor* (que neste trabalho é um classificador), afetando principalmente a sensibilidade ou revocação da solução [22].

O módulo *Preditor* recebe a matriz resultante da execução do *Balanceador da Amostra* e produz como saída a classificação dos pares de pessoas indicando se possuirão ou não um relacionamento. Este módulo utiliza os dados do *passado* e do *presente* para treinamento e teste (verificando a precisão em relação ao *futuro*) e, em seguida, realiza a predição dos dados mais recentes (do *futuro*) utilizando os dados do *presente* como dados anteriores (isto é, como do *passado*).

Características importantes da implementação atual dos módulos serão descritas na próxima subseção.

3.2 Desenvolvimento dos módulos

Todos os módulos desenvolvidos neste trabalho foram implementados na linguagem de programação Java, o que permitiu sua automática incorporação como atividades de workflow no SWMS utilizado. A seguir são detalhadas algumas características de cada módulo e em especial as técnicas ou algoritmos utilizados.

A versão atual do módulo *Extrator de Atributos Estru-*

Tabela 1: Atributos Estruturais Utilizados

Atributo	Descrição
classe	Atributo que assume o valor "possuirão relacionamento" ou "não possuirão relacionamento" com base nos dados do intervalo de tempo chamado neste trabalho de *futuro*.
CN	Common Neighbors - número de vizinhos em comum no grafo correspondente à rede social.
SAL	Salton Index - índice que mede a coocorrência de dois elementos dividida pela raiz quadrada da multiplicação da ocorrência de cada elemento. Em redes sociais pode ser usado para medir relação entre o número de vizinhos que duas pessoas têm em comum divido pela raiz quadrada da multiplicação do número de vizinhos de cada um.
JAC	Jaccard's coefficient - índice que mede a similaridade entre dois conjuntos dividindo o número de elementos da intersecção dos dois conjuntos pelo número de elementos da união (por exemplo, número de vizinhos em comum dividido pela união dos vizinhos de duas pessoas).
AA	Adamic-Adar - índice que atribui peso na relação de duas pessoas favorecendo as relações entre pessoas que possuem poucos relacionamentos (o peso do relacionamento é calculado pela somatória de 1 divido pelo logaritmo do número de relacionamentos [grau] dos vizinhos em comum destas duas pessoas).
RA	Resource Allocation - índice que atribui peso na relação de duas pessoas favorecendo as relações entre pessoas que possuem poucos relacionamentos (o peso do relacionamento é calculado pela somatória de 1 divido pelo número de relacionamentos [grau] dos vizinhos em comum destas duas pessoas).
SOR	Sørensen Index - índice calculado como sendo duas vezes a intersecção entre dois conjuntos dividida pela soma dos elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pelo número de vizinhos da primeira pessoa mais o número de vizinhos da segunda).
HPI	Hub Promoted Index - índice calculado pela divisão do número de elementos da intersecção de dois conjuntos dividido pelo número mínimo de elementos entre estes dois conjuntos (por exemplo, número de vizinhos em comum de duas pessoas dividido pelo número mínimo de vizinhos destas pessoas).
HDI	Hub Depressed Index - índice calculado pela divisão do número de elementos da intersecção de dois conjuntos dividido pelo número máximo de elementos entre estes dois conjuntos (por exemplo, número de vizinhos em comum de duas pessoas dividido pelo número máximo de vizinhos destas pessoas).
LHM	Leicht-Holme-Newman Index - índice calculado pelo número de elementos da intersecção de dois conjuntos dividido pelo produto do número de elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pela multiplicação do número de vizinhos de duas pessoas).
PA	Preferential Attachment - índice dado pelo produto entre o número de elementos de dois conjuntos (por exemplo, produto do número de vizinhos de duas pessoas).
KATZ0,05 KATZ0,005 KATZ0,0005	Katz é um índice calculado de maneira iterativa para estimar a influência de um par de nós em uma rede considerando-se os caminhos existentes entre os nós. Para este cálculo existe a necessidade da definição de uma constante Beta. Os valores utilizados foram: 0,05 ; 0,005 ; e 0,0005.
SP	Shortest Path - caminho mínimo entre dois nós da rede.

turais possui 14 atributos (ou características) estruturais calculados (além do atributo *classe*). Estes atributos são oriundos de métricas estabelecidas por trabalhos correlatos que obtiveram bons desempenhos na predição de relacionamentos [25, 21, 30, 20, 27, 6]. Todos estes atributos são calculados utilizando-se apenas o grafo que representa a rede de relacionamentos. A Tabela 1 apresenta o nome e a descrição destes atributos. Observa-se que foram utilizados três atributos derivados do índice Katz. A maioria desses atributos faz sentido apenas para pares de pessoas pertencentes ao mesmo componente conexo da rede, para os demais pares alguns atributos possuem valores padrões ou simplesmente nenhum valor é produzido para o respectivo atributo.

O módulo *Extrator de Atributos Específicos* recebe como entrada a lista de identificadores de currículos Lattes, copia da internet a versão HTML dos respectivos currículos e produz 19 atributos, conforme pode ser visto na Tabela 2. Este módulo só é utilizado para a predição de coautorias que utilizem dados de currículos Lattes. Para o problema geral de predição de relacionamentos utiliza-se apenas o módulo *Extrator de Atributos Estruturais*. Os dois primeiros atributos são relacionados à classe. O primeiro, *coautorias a serem preditas* contém o número de coautorias que ocorrerão no período *futuro* (ou seja, aquelas que deseja-se predizer). O segundo contém a *classe* propriamente dita (no caso, uma indicação se o respectivo par de pessoas colaborará ou não em uma coautoria).

Todos os atributos específicos, exceto os seis últimos, foram extraídos do trabalho de Digiampietri *et al.*, 2013 [7], sendo que, para se calcular o atributo, *programas em comum* este módulo precisa receber uma tabela que relacione cada pesquisador ao programa de pós-graduação (ou departamento) em que atua, pois esta informação nem sempre é explícita nos currículos Lattes. Dos atributos específicos incluídos no presente trabalho, quatro se referem à produção individual de cada uma das pessoas envolvidas no par de pessoas em análise. Já os dois últimos são *distância* (distância euclidiana entre as coordenadas do endereço profissional de cada par de pessoas) e *subáreas em comum* (correspondendo a contagem de subáreas em comum declaradas pelas duas pessoas no campo "Áreas de atuação" de seus currículos Lattes). Destaca-se ainda o atributo *vizinhos em comum*, apesar de ser semelhante ao atributo estrutural *CN* a diferença é que o primeiro contém todos os "vizinhos" na rede social acadêmica formada pelos diferentes tipos de relacionamento (coautoria, coparticipação em projetos, orientação, etc), já o segundo contém a contagem de vizinhos apenas para a rede (ou grafo) de um relacionamento específico (por exemplo, o de coautoria).

O módulo *Concatenador de Atributos* apenas realiza a

Tabela 2: Atributos Específicos Utilizados

Atributo	Descrição
coautorias a serem preditas	Quantidade de artigos completos publicados em coautoria pelo par de pesquisadores em análise em conferências ou em periódicos no período *futuro*.
classe	Atributo que assume o valor "possuirão relacionamento" caso o atributo "coautorias a serem preditas" seja maior que zero e, caso contrário, "não possuirão relacionamento".
periódicos anterior	Quantidade de artigos publicados em periódicos em coautoria pelo par de pesquisadores no período *passado*.
conferências anterior	Quantidade de artigos completos publicados em conferências em coautoria pelo par de pesquisadores no período *passado*.
periódicos atual	Quantidade de artigos publicados em periódicos em coautoria pelo par de pesquisadores no período *presente*.
conferências atual	Quantidade de artigos completos publicados em conferências em coautoria pelo par de pesquisadores no período *presente*.
orientação anterior	Atributo que recebe o valor 1 (um) caso um dos pesquisadores tenha sido orientador do outro no período *passado*, ou 0 (zero) caso contrário.
orientação atual	Atributo que recebe o valor 1 (um) caso um dos pesquisadores tenha sido orientador do outro no período *presente*, ou 0 (zero) caso contrário.
orientação em andamento	Atributo que recebe o valor 1 (um) caso um dos pesquisadores seja orientador, em uma orientação em andamento, no período *presente*, ou 0 (zero) caso contrário.
orientadores em comum	Quantidade de orientadores e coorientadores que foram orientadores dos dois pesquisadores em análise.
orientandos em comum	Quantidade de orientandos e coorientandos que foram orientados pelos dois pesquisadores em análise.
vizinhos em comum	Quantidade de vizinhos em comum entre os dois pesquisadores na rede social acadêmica (incluindo os diferentes relacionamentos: coautoria, orientação, etc).
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.
artigos periódico1	Quantidade de artigos publicados em periódicos no período *presente* pela pessoa 1.
artigos anais1	Quantidade de artigos completos publicados em anais de conferências no período *presente* pela pessoa 1.
artigos periódico2	Quantidade de artigos publicados em periódicos no período *presente* pela pessoa 2.
artigos anais2	Quantidade de artigos completos publicados em anais de conferências no período *presente* pela pessoa 2.
distância	Distância geográfica entre os endereços profissionais do dois pesquisadores.
subáreas em comum	Número de subáreas de atuação que os dois pesquisadores possuem em comum.

concatenação dos atributos recebidos como entrada, executando ou não a normalização dos valores de cada atributo (de acordo com um parâmetro de entrada) e produz uma matriz na qual cada linha corresponde a um par de pessoas e cada coluna corresponde a um atributo.

O módulo *Filtro Horizontal de Dados* possui três formas de filtragem de dados atualmente implementadas. Na primeira são descartadas as linhas (correspondendo a um par de pessoas) nas quais todos os valores de atributos são nulos. Na segunda são eliminadas todas as linhas cujos pares de pessoas não estão no mesmo componente conexo. Na terceira forma o usuário passa como parâmetro uma expressão booleana que será aplicada sobre cada uma das linhas e indicará se a linha deverá ou não ser descartada.

A versão atual do módulo *Balanceador da Amostra* permite apenas uma estratégia de balanceamento, o *Random Oversampling*. Esta estratégia adiciona aleatoriamente elementos da classe minoritária até que o número de elementos dessa classe se iguale ao número de elementos da classe majoritária.

Conforme apresentado, o problema de predição é, neste trabalho, visto como um problema de classificação em inteligência artificial. O módulo *Preditor*, além de receber o conjunto de dados a ser utilizado no treinamento e teste também receberá um parâmetro indicando qual o planejador a ser utilizado, o caminho (dentro do sistema de arquivos) onde esse planejador se encontra e a estratégia de validação a ser utilizada. Ao invés de se implementar um ou mais classificadores, optou-se por utilizar os classificadores disponíveis no ambiente Weka [19]. Com base nos parâmetros de entrada, o *Preditor* utiliza de reflexão (*reflection*[2]) para invocar o classificador e obter o resultado da classificação/predição.

4. ESTUDO DE CASO

O estudo de caso realizado neste trabalho envolveu a predição de coautorias. A amostra selecionada corresponde aos 657 docentes permanentes dos programas de pós-graduação em Ciência da Computação com doutorado e/ou mestrado acadêmico que atuaram nos triênios 2004-2006 e 2007-2009. As três principais motivações para o uso deste conjunto de dados são: (i) disponibilidade dos dados; (ii) validação: estes dados já foram utilizados para a predição de coautorias

[2]http://docs.oracle.com/javase/tutorial/reflect/

Tabela 3: Resultados da Predição de Novas Coautorias

	Taxa de verdadeiro-positivos	**Taxa de falso-positivos**	**Precisão**	**Revocação**	**F-Measure**	**Área ROC**
Não serão coautores	0,992	0	1	0,992	0,996	1
Serão coautores	1	0,008	0,992	1	0,996	1
Média ponderada	0,996	0,004	0,996	0,996	0,996	1

em [7], então será possível comparar os resultados da solução proposta com os resultados obtidos pelos autores; (iii) ser um amostra de interesse: pelo fato da amostra conter dados de docentes permanentes dos programas de pós-graduação de uma única área do conhecimento há diversos tipos de relacionamentos pertinentes para a predição de coautorias: relações de orientação; de trabalho (num mesmo programa de pós-graduação); e de coautorias [10].

Os parâmetros utilizados para a realização do estudo de caso são: intervalo *passado* envolveu os anos de 1971 a 2000; o *presente* de 2001 a 2005 e o *futuro* de 2006 a 2010. Na filtragem horizontal dos dados foram excluídos os pares de elementos para os quais mais de metade dos atributos possuem valores nulos. Adicionalmente foi passado o parâmetro que indica que o problema a ser tratado é o de predição de relacionamentos novos/inéditos. O classificador utilizado foi o *Rotation Forest* que experimentalmente havia apresentado os melhores resultados para diferentes conjuntos de dados e a estratégia de validação selecionada foi *10-fold-cross-validation*.

Ao se combinar os 657 pesquisadores da amostra dois a dois obtêm-se 215.496 pares. Nesta amostra, o número de pares que efetivamente serão coautores (no *futuro*) é muito menor (apenas 804 pares). Dos 215.496 pares apenas 10.996 foram mantidos após a execução do filtro horizontal de dados. Os demais 204.500 pares (excluídos pelo filtro) são automaticamente classificados como "não possuirão relacionamento" (que neste estudo de caso significa que não serão coautores no *futuro*). Obviamente esta classificação poderia implicar em alguns falso-negativos, porém nenhum destes 204.500 pares foram coautores entre 2006 e 2010. Dos 10.996 pares de docentes restantes, 10.094 pares não haviam colaborado em nenhuma publicação antes de 2005 (ou seja, são candidatos a predição de coautorias novas/inéditas).

A Tabela 3 apresenta os resultados da execução do classificador *Rotation Forest* utilizando a estratégia *10-fold-cross-validation* sobre o conjunto de dados balanceado. As métricas utilizadas para a apresentação dos resultados são: *Taxa de verdadeiro-positivos*, isto é, a quantidade de elementos corretamente classificados como pertencentes à classe dividida pela quantidade de elementos que são da classe; *Taxa de falso-positivos*, isto é, a quantidade de elementos erroneamente classificados como pertencentes à classe dividida pela quantidade de elementos pertencentes à outra classe; *Precisão*, isto é, o número de elementos verdadeiro-positivos dividido pela soma dos verdadeiro-positivos com os falso-positivos; *Revocação*, isto é, a quantidade de elementos classificados como pertencente à classe dividido pela quantidade total de elementos da classe; *F-Measure*, que corresponde a duas vezes o valor da precisão vezes a revocação dividido pela soma da precisão com a revocação; e *Área ROC*.

Conforme pode ser observado na Tabela 3 os resultados foram bastante satisfatórios atingindo uma precisão média acima de 99% para o conjunto balanceado e, ao se considerar também os elementos excluídos pelo filtro e que são classificados como "não serão coautores" a precisão sobe para 99,96% (considerando todo o conjunto de dados). De fato, ao se comparar o resultado da predição para o período *futuro* com o que efetivamente ocorreu (isto é, se os pares de pessoas realmente colaboraram em uma publicação nesse período) a predição errou apenas em 81 dos 215.496 pares de pessoas e o erro foi dizer que estes 81 pares seriam coautores quando na verdade não foram. Pode-se atribuir os resultados obtidos a três características principais. A primeira é que foram utilizados diferentes atributos estruturais encontrados na literatura com bons desempenhos para a predição de relacionamentos. Adicionalmente, foram utilizados atributos específicos do domínio (alguns desenvolvidos especificamente neste trabalho e outros encontrados na literatura). Por fim, o conjunto de dados utilizados nos testes é composto por elementos relativamente homogêneos (todas as pessoas são professores orientadores permanentes por ao menos dois triênios em programas de pós-graduação em Ciência da Computação) o que deve ter simplificado o processo de classificação.

Para uma comparação mais detalhada dos resultados, são apresentados na Tabela 4 os resultados médios de alguns atributos individuais, do conjunto de atributos específicos utilizados em trabalho correlato [7] e, na última linha, a média dos resultados da solução proposta. Todos estes resultados calculados para o conjunto balanceado de dados, utilizando o classificador *Rotation Forest* e a estratégia de validação *10-fold-cross-validation*. Observa-se nesta tabela que os resultados de atributos individuais estão bem aquém dos resultados utilizando-se todos os atributos (última linha) ou ao menos um conjunto de atributos específicos (penúltima linha). Este fato confirma a premissa deste trabalho de que um sistema de predição que combine diferentes atributos/características tende a ser mais preciso e o robusto.

A Tabela 5 apresenta os resultados, utilizando-se a mesma amostra, para o problema que considera tanto a predição de novas coautorias como a predição da reincidência de coautorias. Os resultados utilizaram dados balanceados, o classificador *Rotation Forest* e a estratégia de validação *10-fold-cross-validation*. A precisão, se calculada para todo conjunto de entrada (incluindo os elementos que foram excluídos pelo filtro) é de 99,90%.

5. CONCLUSÕES

Este trabalho apresentou um sistema para a predição de relacionamentos em redes socais baseado em estratégias de classificação em inteligência artificial para combinar atributos que podem ser estruturais da rede, mas também específicos do domínio.

Adicionalmente foi realizado um estudo de caso real utilizando-se dados da Plataforma Lattes de professores de pro-

Tabela 4: Comparações de Resultados

	Taxa de verdadeiro-positivos	**Taxa de falso-positivos**	**Precisão**	**Revocação**	**F-Measure**	**Área ROC**
Subáreas em comum	0,547	0,453	0,547	0,547	0,547	0,542
Vizinhos em comum	0,524	0,476	0,56	0,524	0,438	0,522
PA	0,598	0,402	0,599	0,598	0,597	0,611
Distância	0,596	0,404	0,636	0,596	0,565	0,645
CN	0,61	0,39	0,639	0,61	0,588	0,606
Katz 0,05	0,647	0,353	0,685	0,647	0,628	0,707
Específicos[7]	0,993	0,007	0,993	0,993	0,993	1
Solução Proposta	**0,996**	**0,004**	**0,996**	**0,996**	**0,996**	**1**

Tabela 5: Resultados da Predição de Coautorias (Novas e Reincidentes)

	Taxa de verdadeiro-positivos	**Taxa de falso-positivos**	**Precisão**	**Revocação**	**F-Measure**	**Área ROC**
Não serão coautores	0,98	0	1	0,98	0,99	0,999
Serão coautores	1	0,02	0,98	1	0,99	0,999
Média ponderada	0,99	0,01	0,99	0,99	0,99	0,999

gramas de pós-graduação em ciência da computação, no qual foi obtida uma precisão para o problema de predição de (novas/inéditas) coautorias acima de 99,5% para os dados filtrados e de 99,96% considerando-se o conjunto total de dados.

Também foi possível verificar que a abordagem proposta pode ser utilizada para o problema geral de predição de coautorias (e não apenas de coautorias inéditas), atingindo a precisão de 99% para o conjunto filtrado de dados e 99,9% para o conjunto total de dados.

Como trabalhos futuros pretende-se desenvolver uma interface gráfica específica para o sistema de predição que atualmente utiliza a interface genérica de um sistema de gerenciamento de workflows científicos. Adicionalmente, pretende-se testar e incorporar outras características para a predição de relacionamentos.

Agradecimentos

O trabalho desenvolvido neste artigo foi parcialmente financiado pela CAPES, CNPq e FAPESP.

6. REFERÊNCIAS

[1] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. Kepler: An extensible system for design and execution of scientific workflows. In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*, pages 423–424, Washington, DC, USA, 2004.

[2] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Managing the evolution of dataflows with VisTrails. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*, page 71, 2006.

[3] W. Cukierski, B. Hamner, and B. Yang. Graph-based features for supervised link prediction. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1237–1244, July 2011.

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

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

[6] Y. Dhote, N. Mishra, and S. Sharma. Survey and analysis of temporal link prediction in online social networks. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1178–1183, Aug 2013.

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

[8] L. Digiampietri, B. Teodoro, C. Santiago, G. Oliveira, and J. Araújo. Um sistema de informação extensível para o reconhecimento automático de libras. In *SBSI 2012 - Trilhas Técnicas (Technical Tracks)*, may 2012.

[9] L. A. Digiampietri, J. C. Araujo, E. H. Ostroski, C. R. N. Santiago, and J. J. Perez-Alcazar. Combinando workflows e semântica para facilitar o reuso de software. *Revista de Informática Teórica e Aplicada: RITA*, 20:73–89, 2013.

[10] L. A. Digiampietri, J. P. Mena-Chalco, P. O. S. Vaz de Melo, A. P. R. Malheiro, D. N. O. Meira, L. F. Franco, and L. B. Oliveira. Brax-ray: An x-ray of the brazilian computer science graduate programs. *PLoS ONE*, 9(4):e94541, 04 2014.

[11] L. A. Digiampietri, V. M. Pereira, C. I. Costa, G. J. dos Santos Junior, F. M. Stefanini, and C. R. Santiago. An extensible framework for genomic and metagenomic analysis. In S. Campos, editor, *Advances in Bioinformatics and Computational Biology*, volume 8826 of *Lecture Notes in Computer Science*, pages 1–8. Springer International Publishing, 2014.

[12] C. R. do Nascimento Santiago. Desenvolvimento de

um ambiente de computação voluntária baseado em computação ponto-a-ponto. Master's thesis, Universidade de Sao Paulo, 2015.

[13] Y. Dong, Q. Ke, J. Rao, and B. Wu. Predicting missing links via local feature of common neighbors. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 2, pages 1038–1042, July 2011.

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

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

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

[17] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.

[18] J. Guo and H. Guo. Multi-features link prediction based on matrix. In *Computer Design and Applications (ICCDA), 2010 International Conference on*, volume 1, pages V1–357–V1–361, June 2010.

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

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

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

[22] H. He, Y. Bai, E. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322–1328, June 2008.

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

[24] J. Kunegis, J. Preusse, and F. Schwagereit. What is the added value of negative links in online social networks? In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 727–736, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

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

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

[27] L. Lu and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150 – 1170, 2011.

[28] M. Makrehchi. 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, 2011. ACM.

[29] C. Medeiros, J. Perez-Alcazar, L. Digiampietri, G. Pastorello, A. Santanche, R. Torres, E. Madeira, and E. Bacarin. WOODSS and the Web: Annotating and Reusing Scientific Workflows. *ACM SIGMOD Record*, 34(3):18–23, 2005.

[30] T. Murata and S. Moriyasu. Link prediction based on structural properties of online social networks. *New Generation Computing*, 26(3):245–257, 2008.

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

[32] D. Quercia and L. Capra. 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, 2009. ACM.

[33] I. Taylor, M. Shields, I. Wang, and A. Harrison. Visual grid workflow in triana. *Journal of Grid Computing*, 3(3-4):153–169, 2005.

[34] Y. Tian, Q. He, Q. Zhao, X. Liu, and W.-c. Lee. Boosting social network connectivity with link revival. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 589–598, New York, NY, USA, 2010. ACM.

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

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