Predição de Novas Coautorias na Rede Social Acadêmica dos Programas Brasileiros de Pós-Graduação em Ciência da Computação

Luciano A. Digiampietri[1], **William T. Maruyama**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
Av. Arlindo Bettio, 1000 – CEP 03828-000 – São Paulo – SP – Brasil

Abstract. *This paper extends a previous work, presenting an approach to predict new co-authorships combining artificial intelligence techniques with the state-of-the-art metrics for predicting relationships in social networks.*

Resumo. *Este artigo estende um trabalho prévio, apresentando uma técnica de predição de novas coautorias combinando técnicas de inteligência artificial com o estado da arte das métricas de predição de relacionamentos em redes sociais.*

1. Introdução

Nos últimos anos a predição de relacionamentos em redes sociais ganhou bastante destaque, especialmente nas redes sociais online [Vasuki et al. 2010, Tian et al. 2010, Perez et al. 2012, Fire et al. 2011, Zhong et al. 2013, Quercia and Capra 2009, Hsieh et al. 2013, Dong et al. 2012, de Sa and Prudencio 2011, Lu and Zhou 2011, Hasan and Zaki 2011].

Predizer relacionamentos é uma tarefa complexa pois é necessário identificar os atributos que serão utilizados na predição (que podem ser características de cada indivíduo ou medidas extraídas da rede a que fazem parte) e definir a estratégia para combinar os atributos. Adicionalmente, analisar cada par de indivíduos dentro de uma rede para avaliar se existirá ou não um relacionamento entre eles pode ser um tarefa computacionalmente inviável, sendo necessário definir para quais pares de indivíduos serão realizados todos os cálculos utilizados para a predição.

A predição de relacionamentos tem sido aplicada a redes sociais de diferentes naturezas, servindo, por exemplo, para a sugestão de amizades ou de potenciais colaboradores. Em redes sociais acadêmicas, isto é, redes formadas por pesquisadores, professores e alunos, a predição de relacionamentos é tipicamente utilizada para prever quais pessoas serão coautoras de publicações científicas [Dong et al. 2011, Digiampietri et al. 2013, Gao et al. 2012, Guo and Guo 2010, Lin et al. 2012, Makrehchi 2011]. Este tipo de predição pode ser útil para otimizar a produção destas pessoas através da indicação de pesquisadores cujas parcerias são mais promissoras.

A predição de coautorias pode ser estudada de duas formas. A primeira é o problema geral de identificar entre todos os pares de pessoas/pesquisadores quais serão coautores no futuro. A segunda, é um problema mais específico que visa a predizer coautorias novas (inéditas), isto é, predizer apenas as relações de coautoria entre duas pessoas que nunca haviam colaborado na publicação de um artigo científico.

Este artigo está focado na predição de coautorias inéditas em redes sociais acadêmicas e apresenta uma primeira abordagem para este problema, estendendo um trabalho prévio [Digiampietri et al. 2013] cujo enfoque era no problema geral de predição de coautorias.

2. Trabalhos Correlatos

Nos últimos anos diversos trabalhos foram publicados sobre a predição de relacionamentos, incluindo revisões que apresentam um panorama sobre a área [Lu and Zhou 2011, Hasan and Zaki 2011]. 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)[Dong et al. 2011, Cukierski et al. 2011, Dong et al. 2012, Gao et al. 2012]; aqueles que propõe a utilização de atributos (primitivos ou derivados) específicos do domínio no qual a predição irá ocorrer [Vasuki et al. 2010, Makrehchi 2011, de Sa and Prudencio 2011, da Silva Soares and Bastos Cavalcante Prudencio 2012]; e sistemas híbridos que combinam estes dois aspectos [Kunegis et al. 2013, Digiampietri et al. 2013, Tian et al. 2010, Guo and Guo 2010].

Em um trabalho prévio [Digiampietri et al. 2013], a influência de diferentes atributos extraídos da plataforma Lattes para a predição de relacionamentos foi analisada, bem como o uso da combinação destes atributos. Esse trabalho utilizou como estudo de caso a rede acadêmica de coautorias formada pelos dados dos currículos Lattes dos professores permanentes dos programas de pós-graduação em Ciência da Computação. Nesse trabalho, o problema geral de predição de coautorias (e não o problema específico de predição de coautorias inéditas) foi tratado como um problema de classificação em inteligência artificial no qual cada par de docentes era classificado como "serão coautores" ou "não serão coautores". Esta estratégia atingiu uma alta acurácia na predição, porém a abordagem utilizada foi incapaz de predizer novas coautorias.

O presente trabalho visa a estender o trabalho prévio de forma a possibilitar a predição de coautorias novas/inéditas. Para isto, novos atributos que vem sendo utilizados no estado-da-arte da predição de relacionamentos em redes sociais serão combinados por meio do uso de técnicas de inteligência artificial.

3. Metodologia

A metodologia deste trabalho foi composta de seis atividades: revisão da literatura correlata; seleção da amostra; cálculo dos atributos adicionais; filtragem horizontal dos dados; execução do experimento; e análise dos resultados.

Por meio de uma **revisão sistemática da literatura** correlata foi possível identificar os atributos relacionados à análise de redes sociais que apresentam os melhores resultados na predição de relacionamentos. Neste trabalho foi utilizada a **mesma amostra do trabalho prévio**: 657 docentes permanentes dos programas de pós-graduação em Ciência da Computação no Brasil nos triênios 2004-2006 e 2007-2009.

A principal extensão deste trabalho corresponde ao cálculo e combinação de **16 atributos adicionais**, extraídos dos trabalhos correlatos. A Tabela 1 apresenta o nome e a descrição destes atributos. Os 11 atributos originais podem ser

vistos em [Digiampietri et al. 2013]. Os dados da plataforma Lattes foram copiados e tabulados segundo a metodologia apresentada em [Digiampietri et al. 2012a, Digiampietri et al. 2012b].

Tabela 1. Descrição dos novos atributos utilizados

	Nome do atributo	Descrição
1	Distância	Distância geográfica entre os endereços profissionais do dois pequisadores.
2	CN	Common Neighbors (vizinhos em comum, considernado-se apenas as coautorias de 2001 a 2005).
3	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 quadradada da multiplicação do número de vizinhos de cada um.
4	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 das união (por exemplo, número de vizinhos em comum dividido pela união dos vizinhos de duas pessoas).
5	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).
6	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).
7	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).
8	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).
9	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).
10	LHM	Leicht-Holme-Newman Index - índice calculado pelo número de elementos da intersecção de dois conjuntos dividido pela multiplicação do número de elementos de cada conjunto (por exemplo, número de vizinhos em comum dividido pela multiplicação do número de vizinhos de duas pessoas).
11	PA	Preferential Attachment - índice dado pela multiplicação entre o número de elementos de dois conjuntos (por exemplo, multiplicação do número de vizinhos de duas pessoas).
12	KATZ 0.05	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. Neste artigo três valores de Beta foram considerados: 0,05 ; 0,005 ; e 0,0005.
13	KATZ 0.005	
14	KATZ 0.0005	
15	SP	Shortest Path - caminho mínimo entre dois nós da rede.
16	Subáreas em comum	Número de subáreas de atuação que os dois pesquisadores possuem em comum.

A combinação dos 657 docentes dois a dois resulta em 215.496 pares sendo que a grande maioria destes pares não será coautores (apenas 804 serão). Assim, para tornar o processamento mais eficiente foi realizada uma **filtragem horizontal dos dados** sendo que os pares excluídos foram automaticamente classificados como "não serão coautores". O critério utilizado para a filtragem foi: "excluir os pares de docentes cujos 11 atributos (originais) têm valor igual a zero ou que não possuam ao menos uma aresta na rede de coautorias entre 2001 e 2005". Esta filtragem excluiu 204.500 dos 215.496 pares (ou seja, restaram apenas 10.996 pares para os processamentos subsequentes). Vale observar que os 204.500 pares filtrados de fato não se tornaram coautores no período analisado (de 2006 a 2010).

Na **execução do experimento**, os 27 atributos calculados serão utilizados como entrada para o algoritmo de classificação *Rotation Forest*. Será utilizada a implementação deste algoritmo disponível no arcabouço *Weka* [Hall et al. 2009]. Este algoritmo foi escolhido por ter apresentado os melhores resultados nos testes preliminares e por ter

sido utilizado com bastante sucesso em trabalhos correlatos [Digiampietri et al. 2013, Digiampietri et al. 2012c]. Os **resultados serão analisados** conforme as seguintes métricas: acurácia[1] e sensibilidade[2]. Para a execução dos testes será utilizada a estratégia *10-fold cross-validation*.

4. Experimento e Análise dos Resultados

Conforme apresentado, o problema de predição de novas coautorias foi abordado como um problema de classificação. Após a execução da filtragem horizontal dos dados, restaram 10.996 pares de docentes dos quais 804 serão coautores.

O classificador *Rotation Forest* atingiu uma acurácia de 95,956%. Apesar desta taxa ser elevada, é importante ressaltar que um classificador que classificasse todos os pares como "não serão coautores" teria acurácia de 95,621%. A melhora não muito grande obtida pela solução proposta é justificada pelo grande desbalanceamento do conjunto de dados, mas melhoras, mesmo que pequenas, são significativas devido a este desbalanceamento e à complexidade do problema.

A solução proposta se destacou pela sensibilidade, que foi de 98,958%, isto é, dentre todos os pares que foram classificados como "serão coautores" a solução acertou esta classificação em 98,958% dos casos. A Tabela 2 apresenta a matriz confusão produzida pelo algoritmo *Rotation Forest*. A maior limitação da abordagem é seu baixo valor preditivo positivo[3], que foi de apenas 21,493%. Isto significa que a solução proposta foi capaz de identificar apenas pouco mais de um quinto dos casos de novas coautorias.

Tabela 2. Matriz confusão da predição de novas coautorias

	Classificados como:	
Cassificação Correta:	Serão Coautores	Não serão coautores
Serão coautores	95	347
Não serão coautores	1	9651

5. Conclusões

Este trabalho apresentou nossos primeiros experimentos para a predição de novas coautorias em redes sociais acadêmicas baseada na classificação em inteligência artificial para combinar características individuais de cada pesquisador com características extraídas da rede de colaboração.

Os resultados iniciais foram satisfatórios, obtendo-se acurácia de 95,956% e sensibilidade de 98,958% para o conjunto filtrado de dados e apresentando como limitação um baixo valor preditivo positivo (21,5%).

Como trabalhos futuros pretende-se melhorar o valor preditivo positivo, bem como desenvolver uma análise detalhada da contribuição de cada um dos atributos para a solução geral.

[1] A **acurácia** corresponde a quantidade de acertos (verdadeiro-positivos mais falso-positivos) dividida pela quantidade total de dados.

[2] A **sensibilidade** corresponde a quantidade de predições de coautoria corretas (verdadeiro-positivo) dividida pela quantidade total de predições positivas (verdadeiro-positivos mais falso-positivos).

[3] O **valor preditivo positivo** corresponde a quantidade dos elementos corretamente classificados de forma positiva (verdadeiro-positivos) dividida pela quantidade total de elementos pertencentes a classe de positivos (verdadeiro-positivos mais falso-negativos).

Agradecimentos

Agradecemos a CAPES, CNPq e FAPESP pelo financiamento.

Referências

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

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

de Sa, H. and Prudencio, R. (2011). Supervised link prediction in weighted networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 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-BraSNAM 2013*.

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

Dong, Y., Ke, Q., Rao, J., and Wu, B. (2011). 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.

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 International Conference 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 international conference on and 2011 ieee third international 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 International Conference Companion on World Wide Web*, WWW '12 Companion, pages 507–508, New York, NY, USA. ACM.

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

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.

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.

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 International Conference on World Wide Web*, WWW '13, pages 571–582, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Kunegis, J., Preusse, J., and Schwagereit, F. (2013). 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. International World Wide Web Conferences Steering Committee.

Lin, Z., Yun, X., and Zhu, Y. (2012). 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. IEEE Computer Society.

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

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.

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 International Conference 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.

Tian, Y., He, Q., Zhao, Q., Liu, X., and Lee, W.-c. (2010). 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. 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.

Zhong, E., Fan, W., Zhu, Y., and Yang, Q. (2013). 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. ACM.