Desenvolvimento de Ferramentas para a Análise de Redes Sociais Acadêmicas

Jamison José da Silva Lima, Luciano Antonio Digiampietri

Universidade de São Paulo/Escola de Artes, Ciências e Humanidades

jamison.lima@usp.br

Resumo

No Brasil existe uma base ímpar de dados sobre pesquisadores e suas pesquisas: a Plataforma Lattes. Nessa plataforma, diversas informações referentes a publicações, produções técnicas e projetos de pesquisa são encontradas e podem ser utilizados para a identificação de diversos tipos de relações entre pesquisadores, possibilitando assim a criação de redes sociais de pesquisadores. O objetivo deste projeto foi desenvolver ferramentas que realizem a identificação de especialistas e a clusterização de redes de pesquisadores formadas a partir de dados da Plataforma Lattes, de forma a melhorar a análise de produtividade destas redes.

Palavras Chaves: Análise de Redes Sociais, currículos Lattes, Grupos de Pesquisa

Abstract

In Brazil there is a unique database related to researchers and their researches: the Lattes Platform. In this platform, different information about publications, technical productions, and research projects are found and can be used to identify different types of relationships among researchers, enabling the creation of academic social networks. In this work, tools were developed for the expert finding and the clustering on researcher's networks based on Lattes Curriculum, improving the productivity analysis on these networks.

Key words: Social Network Analysis, Lattes Curriculum, Research Groups

Introdução e Objetivos

Atualmente, uma grande quantidade de informações referentes à produção científica está disponível na Web, como publicações científicas, informações sobre projetos de pesquisa e mesmo currículos de pesquisadores.

Ao se tratar de dados referentes à pesquisa, o Brasil apresenta uma característica peculiar: a existência de um cadastro nacional de currículos de pesquisadores, por meio da Plataforma Lattes, contendo informações sobre publicações, orientações, projetos de pesquisa, entre outras. Esse grande volume de informações ainda não foi amplamente usado e, tipicamente, é consultado para avaliar (ou verificar dados de) pesquisadores individualmente ou orientadores de programas de pós-graduação.

Currículos de pesquisadores podem ser cruzados de forma a estabelecer relações entre os mesmos. O conjunto de pesquisadores e suas relações podem ser vistos como uma Rede Social (*Social Network*). Rede Social é uma estrutura social composta por indivíduos ou organizações chamados nós que são conectados por um ou mais tipos de relação de interdependência como, por exemplo, amizade, crença ou trabalho. Há diversos trabalhos desenvolvidos na área da análise de redes sociais. Muitos destes utilizam a teoria das redes e teoria dos grafos para avaliar e caracterizar Redes Sociais (Berkowitz, 1982, Breiger 2004, Brandes e Erlebach 2005).

Este projeto está contextualizado dentro do Grupo de Pesquisa de Análise de Redes Sociais e Cientometria sediado na Escola de Artes, Ciências e Humanidades da Universidade de São Paulo.

Objetivos

Este projeto visou a estender trabalhos prévios (Alcázar et al., 2011; Digiampietri et al., 2011; Digiampietri, 2012a; Digiampietri et al., 2014; Mena-Chalco et al., 2014) sobre a extração e organização de informações de Currículos Lattes por meio da implementação de um conjunto de métricas e índices a serem calculados tanto diretamente sobre os dados dos Currículos Lattes O objetivo específico deste trabalho foi o desenvolvimento de ferramentas para a identificação automática de grupos e de especialistas em redes sociais acadêmicas criadas a partir dos dados dos currículos Lattes. Com base nesta ferramenta foi realizado um estudo de caso utilizando-se dados reais da área de Ciência Computação.

Materiais e Métodos

Há muitos estudos desenvolvidos na área de análise de redes sociais. Neste projeto foi dado enfoque a dois tipos de trabalhos relacionados à área de ciências exatas: os trabalhos relacionados à caracterização das redes por meio do cálculo de métricas utilizando, por exemplo, teoria dos grafos e teoria de redes (Berkowitz, 1982; Breiger, 2004; Brandes e Erlebach, 2005) e os trabalhos que analisam redes sociais formadas a partir de dados da Plataforma Lattes.

A metodologia utilizada para o desenvolvimento de cada uma das ferramentas foi iniciada com a realização de uma revisão bibliográfica sobre os algoritmos e ferramentas relacionados. A partir do estudo realizado, uma abordagem dentre as encontradas foi escolhida e implementada na linguagem Java. Com base nas ferramentas desenvolvidas, algumas análises foram realizadas.

Primeiramente foi implementado um algoritmo de *clusterização* (ou agrupamento). A clusterização consiste na identificação de grupos em uma rede de modo a maximizar a similaridade entre os elementos de um grupo e minimizar a similaridade entre elementos de diferentes grupos. Para o contexto deste trabalho, o grau de proximidade que será encontrado pelo algoritmo refere-se a sua rede de coautorias, onde autores que possuam coautorias entre si tendem a pertencer ao mesmo grupo. O algoritmo implementado foi baseado na proposta de Schaeffer (2005), utilizando um método de busca local baseado no *simulated annealing* (Kirkpatrick et al, 1983). Os principais motivos para a escolha dessa abordagem foram a escalabilidade do algoritmo, que utiliza uma abordagem local, facilitando a clusterização de grandes redes, e também a ausência de parâmetros, o que facilita a utilização do algoritmo em diferentes redes. Os principais conceitos utilizados para essa implementação são os graus interno e externo de um *cluster* (grupo). O grau interno representa a quantidade de arestas tais que seus dois nós pertençam ao mesmo cluster. Já o grau externo indica o número de arestas onde cada nó pertence a um cluster diferente.

Para a realização da identificação de especialistas, foi realizada a extração de alguns dados dos currículos dos pesquisadores. Foram extraídos dados referentes às suas áreas de pesquisa/atuação, que estão subdivididas em grandes áreas, áreas, subáreas e especialidades. Além disso, foram extraídas informações sobre todas as publicações de artigos completos (em revistas ou anais de eventos). Estas informações foram cruzadas de forma a se formar as redes de coautoria.

A implementação do algoritmo de identificação de especialistas foi baseada na abordagem de Zhang et al. (2007), a qual realiza um cálculo de um valor inicial (*baseline*) para cada autor, baseando-se na quantidade de artigos publicados pelo autor e se o autor publica artigos na subárea que está sendo investigada. Esse valor inicial é utilizado como parâmetro para o cálculo do *AuthorRank* (Liu et al, 2005). O *AuthorRank* é um algoritmo iterativo baseado no *PageRank* (Page e Brin, 1999), idealizado por Liu et al. (2005), que leva em consideração o peso das relações entre os autores (em nosso contexto, nos baseamos na proporção da quantidade de coautorias entre cada par de autores u e v), portanto, quanto maior o número de relações de alto valor do autor com outros autores, maior será o valor do *AuthorRank* final desse autor. Basicamente, o valor do *AuthorRank* de um autor é dado de maneira iterativa como a somatória de parte do *AuthorRank* de seus colaboradores, como demonstrado na fórmula a seguir:

$$AR(u) = c * \sum_{\forall v \in Bu} \frac{AR(v)}{Nv} * W_{u,v} \quad (1)$$

onde Bu é o conjunto de vizinhos de u, Nv é o número de vizinhos de v, Wu,v é o peso da relação entre u e v e c é um fator de normalização. O algoritmo é executado até que convirja, baseando-se na variação total do *AuthorRank* dos nós entre duas iterações.

Resultados

Com as ferramentas implementadas, realizamos um estudo de caso envolvendo os orientadores permanentes dos programas de pós-graduação da área de Ciência da Computação, ativos no ano de 2010.

A Figura 1 apresenta a rede social acadêmica dos orientadores da área de Ciência da Computação. Nesta figura uma cor pode representar mais de um cluster, entretanto, clusters adjacentes possuem cores diferentes. O tamanho de cada nó é relacionado a seu grau de especialidade da subárea “Metodologia e Técnicas da Computação”. É possível notar que existem dois *clusters* principais (o roxo na parte superior e o marrom próximo ao centro), contendo os autores com os maiores valores de *AuthorRank*.

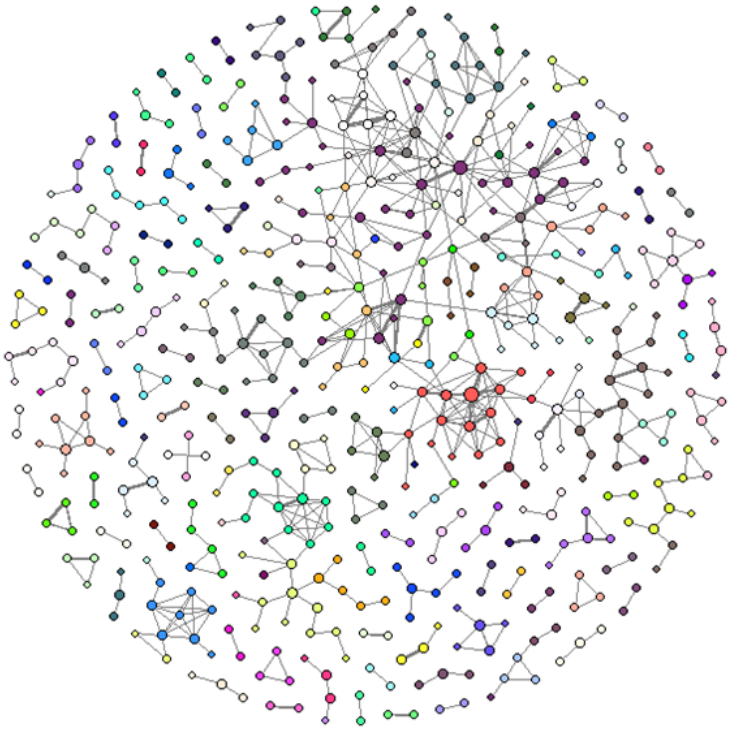

Figura 1: Representação da rede social dos orientadores permanentes da área de Ciência da Computação

Conclusões

Este projeto foi contextualizado dentro de um grupo de pesquisa trabalhando na análise de redes sociais acadêmicas formadas a partir de dados de currículos da Plataforma Lattes. Este grupo está estudando, desenvolvendo e disponibilizando um conjunto de ferramentas para o cálculo de diferentes métricas para a análise de redes sociais.

As ferramentas que desenvolvidas neste projeto objetivaram facilitar a identificação e análise de grupos e especialistas.

Trabalhos futuros poderão adicionar novos parâmetros de seleção para o refinamento na escolha de especialistas. É possível também a utilização das ferramentas atuais na análise de outras bases de dados e análises focadas nos trabalhos científicos, isto é, ao invés de procurar por especialistas poderia se procurar pelos principais artigos de uma determinada área.

Referências Bibliográficas

ALCÁZAR, J. J. P.; ESCALANTE, M.; KURECK, E.; DIGIAMPIETRI, L. A.; TUESTA, E. F. Avaliação de Redes de Inovação usando uma ferramenta baseada em redes sociais - caso Brasileiro de Nanotecnologia. In: XIV Congreso Latino-Iberoamericano de Gestión Tecnológica (ALTEC 2011), 2011, Lima, Perú. Anais do XIV Congreso Latino-Iberoamericano de Gestión Tecnológica, 2011.

BERKOWITZ, S.D. An Introduction to Structural Analysis: The Network Approach to Social Research. Butterworths, Toronto. 1982.

BREIGER, R.L. The Analysis of Social Networks. Handbook of Data Analysis. London, Sage Publications. pp 505–526, 2004.

BRANDES, U; ERLEBACH, T. Network analysis: methodological foundations. Springer. 2005.

DIGIAMPIETRI, L. A., DA SILVA, E. E. A Framework for Social Network of Researchers Analysis. Iberoamerican Journal of Applied Computing, v. 1, n. 1, p. 1-24, 2011.

DIGIAMPIETRI, L. A.; MENA-CHALCO, J.; ALCAZAR, J. J. P.; TUESTA, E. F.; DELGADO, K. V.; MUGNAINI, R.; SILVA, G. S. Minerando e Caracterizando Dados de Currículos Lattes. In proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012, Curitiba, PR, Brasil.

DIGIAMPIETRI, L. A.; MENA-CHALCO, J.; SILVA, G. S.; OLIVEIRA, L. B.; MALHEIROS, A. P.; MEIRA, D. Dinâmica das Relações de Coautoria nos Programas de Pós-Graduação em Computação no Brasil. In proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012, Curitiba, PR, Brasil.

DIGIAMPIETRI, L. A. ; MENA-CHALCO, J. P. ; MELO, P. O. V. ; MALHEIROS, A. P. ; MEIRA, D. N. O. ; FRANCO, L. F. ; OLIVEIRA, L. B. . BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. Plos One, v. 9, p. e94541, 2014.

MENA-CHALCO, J. ; DIGIAMPIETRI, LUCIANO A. ; CESAR-JR, R. M. ; LOPES, F. M. . Brazilian bibliometric coauthorship networks. Journal of the Association for Information Science and Technology, v.65, p. 1424-1445, 2014.

KIRKPATRICK, S; GELATT JR, C.D.; VECCHI, M. P. Optimization by simulated annealing. Science, v. 220, n. 4598, p. 671-680, 1983.

LIU, X., BOLLEN, J., NELSON, M., de SOMPEL, H. V. Co-authorship networks in the digital library research community. Information Processing and Management, 41(6):1462–1480, 2005.

SCHAEFFER, S. E. Stochastic local clustering for massive graphs. In: Advances in knowledge discovery and data mining. Springer Berlin Heidelberg, 2005. p. 354-360.