Modelagem e Caracterização de Redes Científicas: Um Estudo Sobre a Plataforma Lattes

Thiago M. R. Dias[1], Gray F. Moita[1], Patrícia M. Dias[1], Tales H. Moreira[1], Leandro R. Santos[1]

[1]Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

{thiagomagela,patriciamdias,talesinf,leandroroc}@gmail.com,
gray@dppg.cefetmg.br

Abstract. *The analysis of social networks has been the focus of several studies in different areas of knowledge. Mainly scientific collaboration networks, where data on productivity and intensity of collaboration reveal important information for understanding the evolution of research. This article presents the process of data extraction from Plataforma Lattes and modeling the extracted data in a huge network of scientific collaboration. From this network are produced several results that provide an overview of the productivity and collaboration of all users of the platform.*

Resumo. *A análise de redes sociais tem sido foco de vários estudos nas diversas áreas do conhecimento. Principalmente as redes de colaboração científicas, onde dados sobre produtividade e intensidade de colaboração revelam informações importantes para o entendimento da evolução das pesquisas. Este artigo apresenta o processo de extração de dados da Plataforma Lattes e a modelagem dos dados extraídos em uma imensa rede de colaboração científica. A partir desta rede são produzidos diversos resultados que apresentam uma visão geral da produtividade e colaboração de todos os usuários da plataforma.*

1. Introdução

As redes sociais tem sido objetos de estudos a muitos anos. Em (Barabási, 2003) são apresentados diversos estudos que motivaram e agregaram valor a teoria dos grafos e que já se utilizava do conceito de redes sociais.

Rede social é um conjunto de objetos, onde cada um deles está conectado a outro objeto. Uma rede social pode ser representada por um grafo no qual os nós estão relacionados ou não por arestas. As redes sociais refletem uma estrutura social que pode ser representada por indivíduos ou organizações e suas relações. Em geral, as relações representam um ou mais tipos de interdependência (como idéia e religião) ou relacionamentos mais específicos (como troca de conhecimento, informação e amizade) [Stroele et al. 2012].

Devido à possibilidade de extração de informações tanto das características particulares de cada um dos indivíduos que compõem as redes como também das características topológicas que estas possuem, a análise de redessociais tem sido objetos de estudos de diversas áreas do conhecimento.

Com a análise de redes sociais busca-se entender os relacionamentos e o fluxo de informações entre pessoas, grupos e organizações. A unidade na análise de redes sociais não é o indivíduo, mas sim a coleção de indivíduos e os relacionamentos entre eles [Revoredo et al. 2012].

Newman (2001), apresenta uma avaliação extensa sobre características sociais das redes de co-autoria em redes científicas de computação, biologia, física e medicina no período de 1995 a 1999. Já em (Newman 2004), o autor faz análise de co-autoria para identificar propriedades estatísticas, procurando por padrões comuns.

É proposto neste trabalho, a construção de uma rede de colaboração científica, a partir de dados extraídos de um repositório de currículos de pesquisadores e profissionaisde diversas áreas de pesquisa. Para caracterizar a relação de colaboração entre pares de pesquisadores, são analisados trabalhos que 2 (dois) ou mais pesquisadores realizaram em conjunto. Estes trabalhos podem ser de natureza diversa como artigos publicados em conjunto e orientações de trabalhos ou projetos.

2. Modelagem e Caracterização

Para a construção da rede de colaboração científica deste trabalho, foi utilizado a Plataforma Lattes como principal fonte de informação.

A Plataforma Lattes, criada em 1997 pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) com participação do Grupo Stela do PPGEP/UFSC, tornou-se grande conhecida da comunidade científica e tecnológica no Brasil, como também, em países latino-americanos e europeus. A Plataforma Lattes foi concebida para integrar os sistemas de informação das agências federais, racionalizando o processo de gestão de Ciência e Tecnologia (C&T), tanto do ponto de vista do usuário quanto das agências de fomento e das instituições de ensino e pesquisa [CNPQ, 2013].

Diversos trabalhos para extração de dados científicos tem explorado a Plataforma Lattes como principal fonte de dados [Alves e Yanasse 2011(a); Alves e Yanasse2011(b); Alves et al. 2011; Fernandes et al. 2011; Farias et al. 2012; Mena-Chalco et al. 2012; Dias et al. 2013].

O objetivo deste trabalho é utilizar tecnologias envolvidas no processo de extração de dados da *web* para realizar a extração de todos os currículos da Plataforma Lattes e posterior transformação dos currículos em formato XML.

Com os currículos em formato XML diversas métricas e análises estatísticas poderão ser empregadas com o intuito de se obter informações relevantes sobre a rede científica composta de todos os usuários com currículos cadastrados.

Todo este processo pode ser dividido em 3 partes conforme Figura 1 que apresenta a arquitetura geral do sistema.

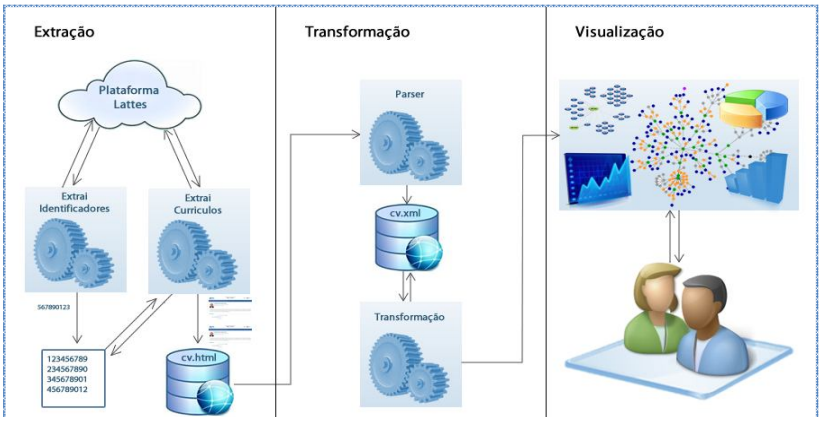

Figura 1 – Arquitetura Geral

2.1. Extração

O processo de aquisição inicia-se com uma requisição do extrator que faz a extração da lista dos identificadores dos usuários que possuem currículos cadastrados na plataforma. Esta lista é armazenada em arquivo para posterior utilização do extrator de currículos.

É importante ressaltar que a requisição retorna uma listagem com identificadores dos currículos por ordem de atualização sendo os primeiros registros aqueles currículos com maior tempo de atualização e os últimos registros os currículos recentemente atualizados e cadastrados.

De posse do arquivo com a listagem atualizada contendo todos os identificadores dos currículos, o extrator de currículos permite selecionar uma quantidade de currículos a serem extraídos, extrair os currículos a partir de uma determinada data de atualização, ou todo o repositório.

Todos os currículos são armazenados em disco um a um até o fim da listagem enviada ao extrator. Para o armazenamento, alguns elementos não necessários para a análise dos currículos como foto e imagens como da bandeira do país de origem dos integrantes são ignoradas. A extração dos currículos é eficiente e permite por exemplo, que o processo seja interrompido e retomado de qualquer parte da lista de identificadores.

2.2. Transformação

A etapa de transformação é caracterizada por uma série de algoritmos que transformam os dados extraídos dos currículos Lattes em formato estruturado que visa facilitar o uso destes em outras etapas da arquitetura.

Ainda no módulo de transformação, algoritmos são responsáveis por gerar consultas e análises dos currículos. É possível visualizar todos os currículos como um grande grafo ou gerar redes específicas a partir de termos que são informados pelos usuários e localizados nos currículos.

Para a identificação de colaboraçõespodem ser utilizados vários elementos para relacionar dois indivíduos como, publicação de trabalhos em conjunto (coautorias), orientações, participação de bancas de avaliação, organização de eventos e participação em projetos.

Além da identificação das colaborações a etapa de transformação é responsável por fazer cálculos de algumas métricas de análise de redes sociais e gerar a rede como um grafo não orientado de toda a rede pesquisada. Arquivos padronizados que podem ser importados por ferramentas de visualização de grafos também são gerados. Todos os resultados são enviados para a etapa de visualização que é responsável pela interação entre toda a arquitetura e o usuário final.

2.3. Visualização

Esta etapa é caracterizada por apresentar aos usuário os resultados das análises e possibilitar a interação entre usuários e sistema.

Com o auxílio de uma interface gráfica o usuário pode escolher qual a rede ele quer visualizar e calcular métricas, pode selecionar para ser gerada toda a rede de colaboração de todos os currículos Lattes ou por exemplo somente daqueles pesquisadores que possuem como Grande Área de atuação, Ciência da Computação.

Após a geração da rede, é possível aplicar algumas métricas de análises de redes sociais informando o que se deseja e um algoritmo de visualização é responsável por atualizar os dados e gerar novas visualizações.

3. Análise da Rede

A Plataforma Lattes é uma fonte extremamente rica de informações. O grande volume de dados presente nos currículos podem ser melhores trabalhados e fornecer informações valiosas até então desconhecidas. Em fevereiro de 2013 a Plataforma Lattes alcançou a marca de 3.000.000 de currículos cadastrados. Todos os currículos foram extraídos como prova de conceito para a arquitetura aqui apresentada.

Como resultado do algoritmo de identificação de colaborações científicas descrito anteriormente, é possível observar a rede com todos os currículos na Figura 2.

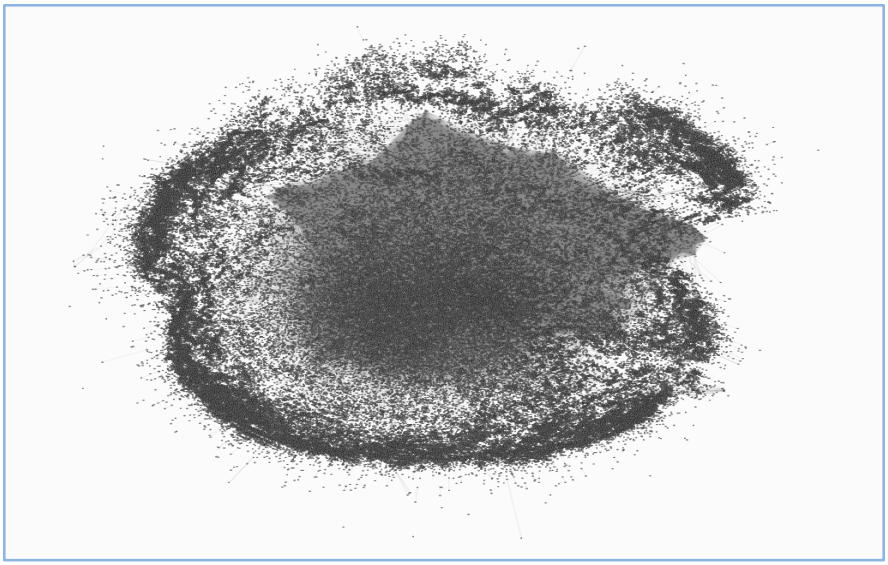

Figura 2 – Rede com 3.001.980 currículos coletados em fevereiro de 2013

É possível ainda a geração de redes específicas tendo como base termos dentro dos currículos. Exemplos de redes específicas podem ser visualizadas na Figura 3.

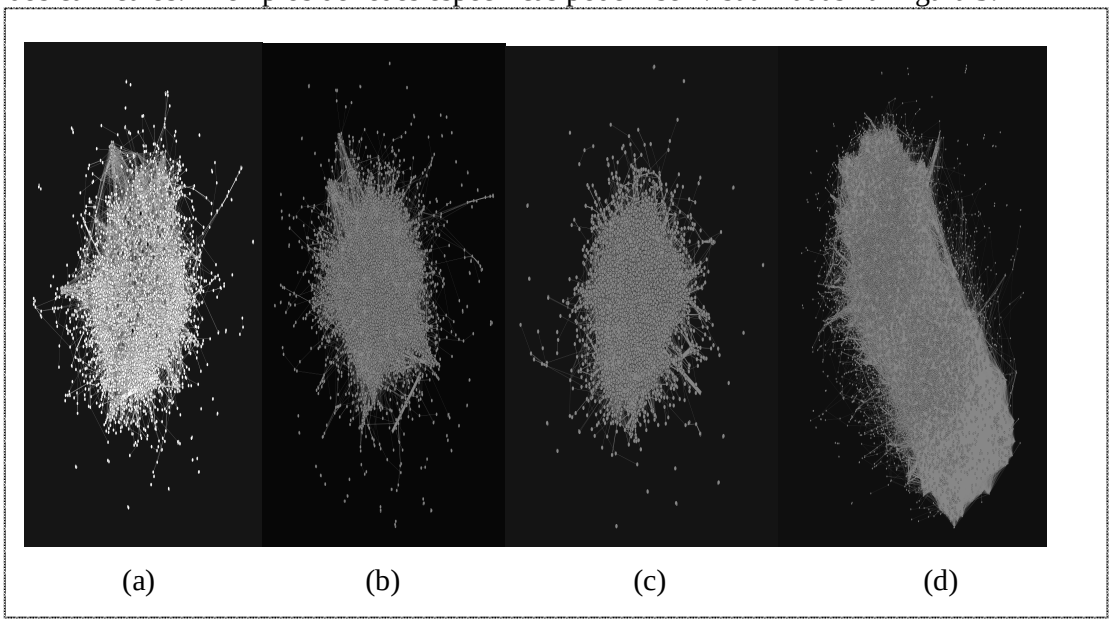

Figura 3 – Redes: a) FAPEMIG, b) UFMG, c) CSBC, d) Ciência da Computação

A rede de currículos que possui o termo FAPEMIG é composta por um total de 20.307 currículos cadastrados na plataforma e um total de 89.336 currículos possuem o termo UFMG em qualquer um de seus cursos de formação.

Análises como participação em congressos ou publicações de artigos em periódicos podem ser pesquisados e obter a rede deste subconjunto de autores. Exemplo disto é a rede de participantes do Congresso da Sociedade Brasileira de Computação que possui 6.354 integrantes que declararam ter participação de alguma edição do congresso e uma rede de 142.571pessoas que informam ter como Área de Atuação o termo Ciência da Computação.

4. Conclusões

Este trabalho apresenta uma arquitetura para extração, transformação e visualização de todos os currículos cadastrados na Plataforma Lattes. A extração acontece com grande eficiência e possibilita extrair os currículos na ordem em que foram atualizados ou cadastrados, sendo os últimos currículos os mais recentemente atualizados.

Após a extração os currículos são padronizados em formato XML para posterior consulta e análises. Uma técnica eficiente para identificar colaboração entre os currículos é implementada e as redes de colaboração são geradas com maior precisão. Pode se gerar redes específicas por termos, ou toda a rede da plataforma.

Referências

Alves, A.D.; Yanasse, H. H. (2011) Perfil dos Bolsistas PQ das Áreas de Engenharia de Produção e de Transportes do CNPq: Enfoque na Subárea de Pesquisa Operacional.

In: XLIII Simpósio Brasileiro de Pesquisa Operacional, 2011, Ubatuba. Anais do XLIII SBPO. v. 1. p. 144-155

Alves, A.D. ; Yanasse, H. H. (2011) Sucupira: um Sistema de Extração de Informações da Plataforma Lattes para Identificação de Redes Sociais Acadêmicas. In: CISTI'2011 (6ª Conferência Ibérica de Sistemas e Tecnologias de Informação). Anais do CISTI`2011.

Alves, A. D.; Yanasse, H. H.; Soma, N. Y. (2011) LattesMiner: a multilingual DSL for information extraction from lattes platform. In Proceedings of the compilation of the co-located workshops on DSM'11, TMC'11, AGERE!'11, AOOPES'11, NEAT'11, &VMIL'11(SPLASH '11 Workshops). ACM, New York, NY, USA, 85-92

Barabasi, A.L. Linked: How Everything Is Connected to Everything Else and What It Means. Plume. 2003. USA 1ª edição.

CNPQ. Sobre a Plataforma Lattes. 2013. URL http://www.http://lattes.cnpq.br/. Acessado: 08/04/2013.

Dias, T. M. R.; Moita, G. F.; Dias, P. M. Analysis Collaboration Networks of Scientific Publications. In: X International Conference of Information System and Technology Management (CONTECSI). São Paulo, SP, Brasil.

Farias, L. R.; Vargas, A. P.; Borges, E. N. (2012).Um sistema para análise de redes de pesquisa baseado na Plataforma Lattes. Escola Regional de Banco de Dados, Curitiba - PR.

Fernandes, G.O.; Sampaio, J. O. ; Souza, J. M. (2011). XMLattes - A Tool for Importing and Exporting Curricula Data. In: WORLDCOMP'11 - The 2011 World Congress in Computer Science, Computer Engineering, and Applied Computing, Las Vegas - Nevada.

Mena-Chalco, J. ; Digiampietri, Luciano A. ; Cesar-Jr, R. M. (2012). Caracterizando as redes de coautoria de currículos Lattes. In: Brazilian Workshop on Social Network Analysisand Mining (BraSNAM), Curitiba, PR, Brasil.

Newman, M. E. (2001) The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America, v. 98, n. 2, p. 404-9.

Newman, M. E. (2004) Co-authorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America, v. 101, p. 5200-5205.

Revoredo, k.,Araújo R., Silveira B. and Muramatsu T. (2012). Minerando publicações científicas para análise da colaboração em comunidades de pesquisa.In: Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), Curitiba- PR.

Stroele, V., Zimbrão, G. Souza, J. M. (2012).Análise de Redes Sociais Científicas: Modelagem Multi-relacional. In: Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012, Curitiba- PR.