X WORCAP
20 e 21 de outubro
Instituto Nacional de Pesquisas Espaciais - INPE
São José dos Campos - SP

Especificação e Implementação de uma Linguagem de Domínio Específico para Análise de Redes Sociais Acadêmicas

Alexandre D. Alves[1], Horacio H. Yanasse[1], Nei H. Soma[2]

[1]Programa de Doutorado em Computação Aplicada – CAP
Instituto Nacional de Pesquisas Espaciais – INPE

[2]Instituto Tecnológico de Aeronáutica (ITA)
Praça Marechal Eduardo Gomes, 50 – Vila das Acácias
12.228–010 – São José dos Campos – SP – Brasil

alexdonizeti@gmail.com, horacio@lac.inpe.br, soma@ita.br

Abstract. *The explosive growth and popularity of the Web has resulted in many sources of information on the Internet. Increasingly tools that automatically extract only the data of interest to a user are needed, facilitating the access and the manipulation of these information. This paper proposes the specification and implementation of a domain-specific language that allows the extraction of information for identification, visualization and analysis of academic social network. A high-level language allows any user to implement his/her own applications with a high-level abstraction and expression power.*

Resumo. *O crescimento explosivo e a popularidade da Web têm resultado em uma grande quantidade de fontes de informação na Internet. Cada vez mais se fazem necessárias ferramentas capazes de extrair automaticamente apenas os dados de interesse de um usuário, facilitando o acesso e a manipulação dessas informações. Este artigo propõe a especificação e implementação de uma linguagem de domínio específico que permita a extração de informações para a identificação, visualização e análise de redes sociais acadêmicas. Uma linguagem de alto nível possibilita a qualquer usuário implementar as suas próprias aplicações com alto nível de abstração e poder de expressão.*

Palavras-chave: *Redes Sociais Acadêmicas, Linguagem de Domínio Específico, Extração de Informação.*

1. Introdução

O crescimento explosivo e a popularidade da Web têm resultado em uma grande quantidade de fontes de informação na Internet. A Web é hoje uma grande fonte de informação, fazendo com que o processo de extração de informação de conteúdos Web seja considerado um problema importante. Cada vez mais fazem-se necessárias ferramentas capazes

de extrair automaticamente os dados de interesse de um usuário, facilitando o acesso e a manipulação dessas informações. Isto traz grandes desafios na elaboração de metodologias eficazes para pesquisa, acesso e integração de informação [Vadrevu et al. 2007]

O volume de informações disponíveis na Web hoje em dia é muito grande. Percebe-se que há a necessidade de mecanismos que permitam extrair essas informações e transformá-las em conhecimento. Entretanto, alguns problemas dificultam a extração de informações de páginas Web.

Um problema sério é a ambigüidade. Há uma variedade de maneiras de se referir a uma mesma pessoa. Por exemplo, `J. Silva`, `João Silva` e `João da Silva` podem todos se referir à mesma pessoa. Esse problema também reduz a possibilidade de identificação de redes sociais acadêmicas e pode introduzir informações erradas.

Um dos problemas mais difíceis de se resolver é decidir se dois objetos localizados em fontes diferentes se referem à mesma entidade. Como no caso de bases científicas, em que cada uma adota um padrão de citação, dificultando ainda mais a extração de informações sobre um determinado autor.

O objetivo principal é a especificação e implementação de uma DSL que permita a extração de informações para a identificação, visualização e análise de redes sociais acadêmicas. A importância de se desenvolver uma linguagem de alto nível é que ela possibilita a outros usuários implementarem as suas próprias aplicações com alto nível de abstração e poder de expressão. O foco inicial da DSL será a Plataforma Lattes.

Uma Linguagem de Domínio Específico (*Domain-Specific Language* - DSL) tem como objetivo resolver um problema em particular, tornando-a mais acessível ao público comparada às linguagens de programação tradicionais [Taha 2008]. O processo de aprendizagem poderia ser bem mais rápido e intuitivo, uma vez que a linguagem poderia ser mais próxima da língua dos usuários não exigindo especialistas em programação.

A Plataforma Lattes é hoje, sem dúvida, a principal fonte de informações sobre os pesquisadores brasileiros e tem um elevado potencial para extração de informação. Por exemplo, é possível identificar as redes sociais acadêmicas existentes entre os pesquisadores, embora isso não seja uma tarefa simples e imediata. Entretanto, não existem mecanismos que permitam que isso seja feito de maneira simples e rápida, e sem o auxílio de desenvolvedores experientes.

Os problemas já citados anteriormente para extração de informações também ocorrem na Plataforma Lattes. Além desses problemas, há outros que são bem específicos da Plataforma Lattes. Um dos problemas em extrair dados de um currículo Lattes disponível na Web é a falta de padronização dos dados registrados. Muitos currículos são parcialmente preenchidos e muitos pesquisadores não atualizam seus currículos periodicamente.

Os trabalhos encontrados na literatura que usam a Plataforma Lattes como fonte de extração de informação apresentam limitações, tais como:

- A aquisição dos currículos Lattes normalmente é feita manualmente;
- Quando os currículos são extraídos em formato XML (*Extensible Markup Language*) há restrições, pois somente instituições licenciadas têm acesso e é permitido extrair somente os currículos dessa instituição;

- Quando uma análise é feita é muito complicado repetir o mesmo processo, principalmente pelo fato da análise não ser sido feita automaticamente. Nesse caso, evidentemente, é um processo demorado.

A Extração de Informação na Plataforma Lattes poderia ajudar o governo a descobrir informações de determinadas áreas e assim, por exemplo, investir de acordo com a região onde estão ou mantêm vínculo os principais pesquisadores daquela determinada área. Também poderia ajudar a responder perguntas como: o ambiente em que estou pode influenciar na minha carreira profissional e acadêmica? Se trabalho com algum pesquisador que está trabalhando ativamente, a minha chance de sucesso na vida acadêmica aumenta?

O número atual de currículos Lattes cadastrados já passa de um milhão e está constantemente aumentando. A diversidade de dados e o número de possíveis relacionamentos entre pesquisadores tornam a tarefa de encontrar os relacionamentos existentes complexa.

Recentemente, a própria Plataforma Lattes incorporou no sistema Currículo Lattes um novo recurso que permite descobrir a rede de colaboração de um pesquisador. Com base nas suas publicações, é possível visualizar graficamente a rede de co-autores de um pesquisador desde que os mesmos também tenham currículo Lattes. A rede é composta por outros pesquisadores que trabalharam em conjunto com o pesquisador em questão em co-autoria de artigos científicos.

Muitos dos pesquisadores que têm usado dados da Plataforma Lattes para a identificação de redes sociais acadêmicas. Entretanto, a maioria é de outra área e mesmo para quem é da área da Computação, não é uma tarefa simples extrair informações do currículo Lattes ou mesmo de qualquer outra base científica.

Dessa forma, há a necessidade de encontrar mecanismos que permitem realizar essas tarefas com um maior nível de abstração, por um maior número de usuários e também de forma mais eficiente.

Este artigo está organizado da seguinte maneira: na Seção 2 são apresentados os trabalhos relacionados. Na Seção 3 é apresentada a API *LattesMiner*, destacando os seus principais componentes. Na Seção 4 é apresentada uma comparação entre essas ferramentas. Na Seção 5 é apresentada a descrição da proposta. Finalmente, na Seção 6 são apresentados os resultados esperados.

2. Trabalhos Relacionados

Nos últimos anos, muitos trabalhos têm sido realizados usando dados disponíveis na base de dados da Plataforma Lattes [de Oliveira et al. 2004, Cardoso and Machado 2008, Pacheco et al. 2007]. Alguns trabalhos na literatura têm analisado o perfil de pesquisadores bolsistas de produtividade em pesquisa em áreas como Saúde Coletiva [Barata and Goldbaum 2003] e Odontologia [Scarpelli et al. 2008, Cavalcante et al. 2008]. Essas análises foram feitas usando informações contidas no currículo Lattes desses pesquisadores. Um problema comum apresentado em alguns trabalhos é que os currículos e as informações extraídas foram obtidas manualmente. [Cavalcante et al. 2008] descrevem que levaram quase 3 anos para analisar 132 currículos. Dessa forma, é muito complicado realizar uma outra análise.

Pela pesquisa efetuada na literatura, existem apenas duas ferramentas que per-

mitem a extração de dados a partir de currículos Lattes. São elas: Lattes Extrator e ScriptLattes.

Lattes Extrator foi desenvolvida pelo próprio CNPq e é uma das ferramentas que compõem a Plataforma Lattes. É acessível via Web e seu acesso é restrito. Atualmente, somente instituições licenciadas podem extrair informações diretamente do banco de dados de currículos Lattes do CNPq e somente informações de seus pesquisadores, docentes, estudantes e colaboradores [CNPq 2009]. As informações extraídas são disponibilizadas em arquivos no formato XML definido pela comunidade LMPL (Linguagem de Marcação da Plataforma Lattes) e as instituições podem desenvolver rotinas para a importação dessas informações para as suas próprias bases. As extrações são feitas em lote e podem ser configuradas de acordo com o interesse e as permissões de cada usuário.

ScripLattes é um *script* desenvolvido em Perl (*Practical Extraction and Report Language*) para extração e compilação de produções bibliográficas, produções técnicas, produções artísticas e orientações de um grupo de pesquisadores cadastrados no sistema Currículo Lattes [Mena-Chalco and Junior 2009]. A primeira versão, lançada em 2005, foi desenvolvida para auxiliar a secretaria do Programa de Pós-Graduação do IME-USP na elaboração de relatórios sobre a produção bibliográfica dos docentes do Departamento de Ciência da Computação. Esses relatórios foram baseados nas informações cadastradas nos currículos Lattes desses docentes. Atualmente, os relatórios podem ser gerados em português, inglês e espanhol.

O *script* baixa os currículos Lattes, compila as listas de publicações e orientações e gera páginas Web contendo essas informações, um grafo de colaborações entre os pesquisadores e um mapa de pesquisa. Para executar o *script* é necessário criar um arquivo no formato texto contendo o número dos pesquisadores. Esse número contém 16 digítos e é usado como um identificador (ID) para cada currículo Lattes.

O grafo de colaborações é obtido a partir de relações entre um grupo de pesquisadores. O grafo é gerado considerando publicações com títulos iguais ou similares e o número de relações encontradas entre os pesquisadores é exibido nas arestas. O grafo é estástico, ou seja, não permite qualquer tipo de interação; permite apenas clicar nos nomes dos pesquisadores. Esta ação abre uma página contendo o currículo Lattes do pesquisador em que o nome foi clicado. Apesar de existir uma versão interativa do grafo de colaborações, a interação é muito limitada, sendo permitido somente funções simples de *zoom* e arrastar todo o grafo usando o mouse ou o teclado.

O mapa de pesquisa é gerado baseando-se nos CEPs (Código de Endereçamento Postal) cadastrados nos currículos dos pesquisadores, calculando a latitude e longitude de cada endereço. O mapa é exibido usando a API do *Google Maps*.

A licença da ferramenta ScriptLattes é GPL (*General Public License*) e é executada apenas no sistema operacional Linux. É necessário ter um compilador Perl configurado e bibliotecas gráficas instaladas.

3. A API *LattesMiner*

LattesMiner é uma API orientada a objetos para extração de informações de currículos Lattes e identificação de redes sociais acadêmicas. A API é composta por um conjunto de classes escritas em Java que permite que outros desenvolvedores implementem as suas

próprias aplicações [Alves et al. 2009].

A API está em desenvolvimento e faz parte de um projeto maior denominado "Sistema Unificado de Currículos e Programas: Identificação de Redes Acadêmicas - SUCUPIRA" (processo Capes 23038-029609/2008-02). O objetivo principal desse projeto é desenvolver ferramentas computacionais acessíveis pela Web para auxiliar na obtenção de indicadores de desempenho de docentes, pesquisadores e programas de pós-graduação. As informações serão apresentadas na forma de grafos contendo os diversos relacionamentos entre pesquisadores e os programas de pós-graduação.

A API *LattesMiner* é composta por seis componentes. O componente **Descoberta de Dados** é opcional, ou seja, é necessário somente se o número (ID) dos pesquisadores não estiver disponível. O componente **Aquisição de Dados** também é opcional, uma vez que o currículo Lattes pode ser baixado diretamente do sítio do CNPq, sendo necessário apenas que o currículo seja armazenado como arquivo HTML.

Uma visão geral da arquitetura de componentes da API *LattesMiner* é ilustrada na Figura 1. Os componentes **Descoberta de Dados** e **Aquisição de Dados** acessam o sistema Currículo Lattes através do código ou número (ID) do pesquisador. Para o componente **Descoberta de Dados** é retornado apenas a parte inicial do currículo Lattes para verificar se o nome contido no currículo é igual ao nome procurado. Para o componente **Aquisição de Dados** é retornado uma cópia do currículo Lattes que é armazenado como arquivo HTML.

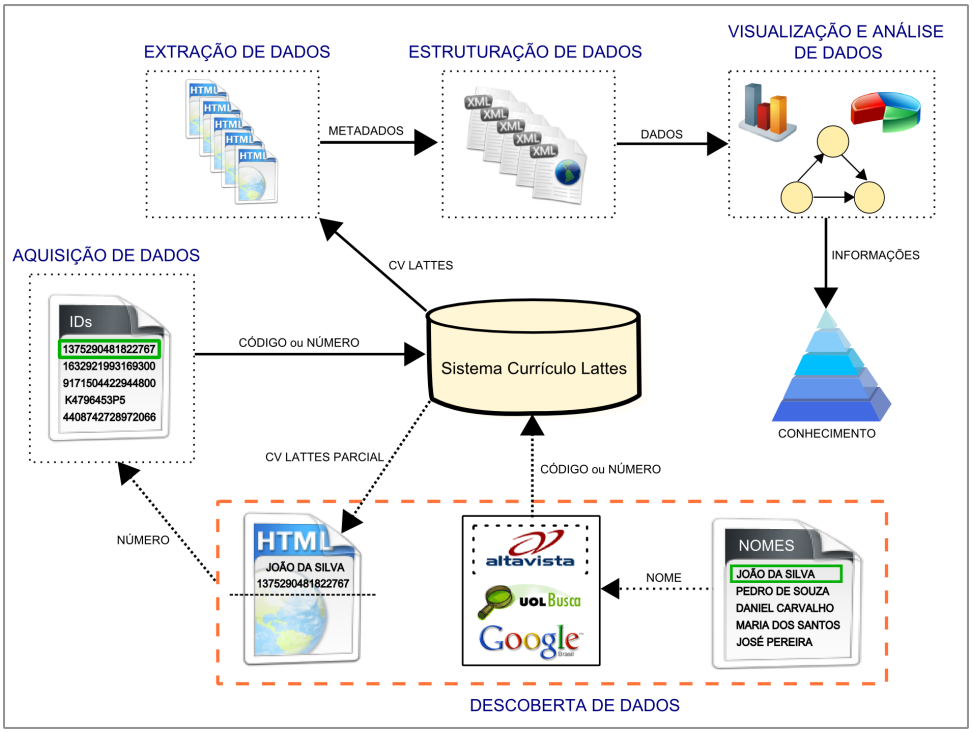

Figura 1. Arquitetura de componentes da API *LattesMiner*.

O principal componente da API *LattesMiner* é **Extração de Dados**. Este componente é responsável pela extração de dados dos arquivos HTML. Para a extração de dados de um currículo Lattes a partir do arquivo HTML foi usada a técnica de extração

de informação baseada em expressões regulares [Vadrevu et al. 2007, Xiao et al. 2004]. Foi observado que trechos de código no arquivo HTML do currículo Lattes têm uma estrutura de repetição, ou seja, têm a mesma formatação HTML [Nanno et al. 2003]. Por essa razão é que a técnica de extração de informação baseada em expressões regulares foi usada.

O componente **Visualização de Dados** é responsável pela identificação e visualização de redes sociais acadêmicas. A identificação dessas redes sociais é feita verificando os relacionamentos entre os pesquisadores. E como essa identificação considera apenas as informações acadêmicas dos pesquisadores, essas redes foram chamadas de redes sociais acadêmicas.

A visualização das redes sociais é feita na forma de grafos e para isso foi usado o framework JUNG (*Java Universal Network/Graph*). JUNG é uma biblioteca de código aberto escrita em Java que fornece uma linguagem comum e extensível para a modelagem, análise e visualização de dados que podem ser representados como um grafo ou uma rede [JUNG 2009].

O componente **Análise de Dados** é responsável pela análise dos dados extraídos e também pela análise dos relacionamentos identificados. No momento, a API *LattesMiner* permite apenas análise simples das relações identificadas, como a identificação de cliques e da clique máxima.

4. Comparação entre as ferramentas

Nas seções anteriores foram apresentadas três ferramentas para a extração de informações de currículos Lattes. Lattes Extrator tem a vantagem de extrair os currículos diretamente do banco de dados da Plataforma Lattes e são extraídos como arquivos XML. As ferramentas ScriptLattes e LattesMiner extraem os currículos a partir do sistema Currículo Lattes disponível na Web. Nessas ferramentas, os currículos são extraídos como páginas Web, o que dificulta a extração de informações, pois uma página Web é um documento semi-estruturado.

Lattes Extrator permite configurar quais informações serão extraídas e pode extrair qualquer informação disponível, ao contrário das ferramentas ScriptLattes e LattesMiner que extraem apenas algumas informações. LattesMiner também é configurável, ou seja, é possível definir as informações a serem extraídas.

ScriptLattes permite a identificação de alguns relacionamentos entre um grupo de pesquisadores, mas não é um biblioteca que pode ser programada pelo usuário. Outro problema é que as páginas são geradas em HTML e JSP (*JavaServer Pages*), o que obriga o usuário a ter um servidor Web instalado e devidamente configurado para executar páginas dinâmicas em Java.

A principal vantagem da API LattesMiner é o fato de ser uma biblioteca que permite aos desenvolvedores programarem suas próprias aplicações, ao contrário das ferramentas Lattes Extrator e ScripLattes. Essas ferramentas realizam busca apenas pelo número (ID) do pesquisador, enquanto a API LattesMiner também permite a busca pelo nome do pesquisador. Dessa forma, é possível, por exemplo, buscar por qualquer nome citado em um currículo Lattes, aumentando assim o número de relacionamentos identificáveis.

Outra vantagem da API é que as informações extraídas podem ser importadas por outras ferramentas, uma vez que essas informações podem ser armazenadas em XML. Também é possível exportar os grafos de relacionamentos para formatos populares, como Pajek [de Nooy et al. 2005] e GraphML [GraphML 2009]. Dessa forma, os grafos podem também ser analisados por outras ferramentas e algoritmos diferentes.

Outra vantagem é o fato da API ser implementada em Java. Como Java é uma linguagem independente de plataforma, a API também pode ser usada em qualquer plataforma e sistema operacional. Além disso, a API permite um alto nível de abstração por ser orientada a objetos. A API também faz uso de *threads*, o que permite um melhor desempenho, principalmente, em máquinas com mais de um processador.

A Tabela 1 apresenta uma comparação entre as ferramentas, destacando as suas principais características.

Tabela 1. Quadro comparativo entre as ferramentas Lattes Extrator, ScriptLattes e LattesMiner.

Tópicos considerados	Lattes Extrator	ScriptLattes	LattesMiner
Linguagem de desenvolvimento	JSP	Perl e JSP	Java
Local de desenvolvimento	CNPq	IME-USP	INPE/ITA
Formato dos currículos extraídos	XML	HTML	HTML
Restrição de Sistema Operacional	-	Linux	-
Busca pelo nome do pesquisador	não	não	sim
Biblioteca programável	não	não	sim
Visualização de Redes Sociais	não	sim	sim
Análise de Redes Sociais	não	não	sim
Relatórios e Gráficos	não	sim	não
Exportação de Redes Sociais	não	não	sim

5. Descrição da proposta

Os principais objetivos são:

- Especificar e implementar uma DSL que permita a extração de informações para a identificação, visualização e análise de redes sociais acadêmicas.
- Especificar e implementar uma ferramenta para extração de conhecimento de redes sociais acadêmicas. A ferramenta deve permitir descobrir e visualizar os relacionamentos existentes entre grupos de usuários, sendo que as informações serão apresentadas em forma de grafos, gráficos, tabelas e mapas.

O foco inicial da DSL será a Plataforma Lattes e para a definição da DSL, a API LattesMiner será usada como base.

A ferramenta para extração de conhecimento que será proposta com base na DSL deverá ser flexível o suficiente para permitir que o próprio usuário possa configurar e obter as visualizações, em tempo real, das informações que lhe são mais relevantes. A ferramenta também deverá ser acessível via Web e ser de acesso público. Outra característica importante da ferramenta é permitir a colaboração entre os usuários, pois cada usuário tem diferentes experiências e conhecimentos que podem ser compartilhados com os demais.

6. Resultados esperados

Os principais resultados esperados são:

- Uma DSL para a identificação, visualização e análise de redes sociais acadêmicas.
- Uma ferramenta para extração de conhecimento de redes sociais acadêmicas.

De maneira geral, espera-se que o projeto proposto possa auxiliar a extração de informações relevantes sobre pesquisadores. Dessa forma, será possível analisar a colaboração entre pesquisadores, instituições e até mesmo países. Também permitirá verificar o surgimento ou desaparecimento de áreas de pesquisa, permitindo ao governo e as agências de fomento saberem onde melhor investir.

Outra contribuição do projeto é auxiliar na identificação de pontos vulneráveis na rede. Isso é importante pois se um determinado pesquisador ocupa uma posição chave na rede social acadêmica de uma instituição e esse pesquisador se aposenta, por exemplo, a instituição pode vir a ter sérios problemas. Além disso, esse pesquisador também pode ser o elo de ligação com outras instituições e isso pode ser perdido a qualquer momento, principalmente, se esse pesquisador for o único elo de ligação.

Referências

Alves, A. D., Yanasse, H. H., and Soma, N. H. (2009). Extração de informação na plataforma lattes para identificação de redes sociais acadêmicas. In *IX Workshop do Curso de Computação Aplicada - WORCAP*, São José dos Campos, SP.

Barata, R. B. and Goldbaum, M. (2003). Perfil dos pesquisadores com bolsa de produtividade em pesquisa do cnpq da área de saúde coletiva. *Cadernos de Saúde Pública*, 19(6):1863–1876.

Cardoso, O. N. P. and Machado, R. T. M. (2008). Gestão do conhecimento usando data mining: estudo de caso na universidade federal de lavras. *Revista de Administração Pública*, 42(3):495–528.

Cavalcante, R. A., Barbosa, D. R., Bonan, P. R. F., de Oliveira Pires, M. B., and Martelli-Júnior., H. (2008). Perfil dos pesquisadores da área de odontologia no conselho nacional de desenvolvimento científico e tecnológico (cnpq). *Revista Brasileira de Epidemiologia*, 11(1):106–113.

CNPq (2009). Cnpq - plataforma lattes - lattes extrator.

de Nooy, W., Mrvar, A., and Batagelj, V. (2005). *Exploratory Social Network Analysis With Pajek*. Cambridge University Press.

de Oliveira, E., de Souza Bermejo, P. H., and Kern, V. M. (2004). Geralattes: extração de informação gerencial de currículos de pesquisadores usando xml. In *WorkCompSul 2004 - I Workshop de Computação da Região Sul*, Florianópolis, SC.

GraphML (2009). The graphml file format.

JUNG (2009). Java universal network/graph framework.

Mena-Chalco, J. P. and Junior, R. M. C. (2009). Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39.

Nanno, T., Saito, S., and Okumura, M. (2003). Structuring web pages based on repetition of elements. In *Second International Workshop on Web Document Analysis*, Japão.

Pacheco, R. C. S., Forcellini, F. A., Kern, V. M., Gonçalves, A. L., and Igarashi, W. (2007). Uma análise da pesquisa em engenharia e ciências mecânicas no brasil a partir dos dados da plataforma lattes. *Associação Brasileira de Engenharia e Ciências Mecânicas (ABCM)*, 12(1):18–24.

Scarpelli, A. C., Sardenberg, F., Goursand, D., Paiva, S. M., and Pordeus, I. A. (2008). Academic trajectories of dental researchers receiving cnpq's productivity grants. *Brazilian Dental Journal*, 19(3):252–256.

Taha, W. (2008). Plenary talk iii domain-specific languages. In *Computer Engineering & Systems, 2008. ICCES 2008. International Conference*, pages xxiii–xxviii.

Vadrevu, S., Gelgi, F., and Davulcu, H. (2007). Information extraction from web pages using presentation regularities and domain knowledge. *World Wide Web*, 10(2):157–179.

Xiao, L., Wissmann, D., Brown, M., and Jablonski, S. (2004). Information extraction from the web: System and techniques. *Applied Intelligence*, 21(2):195–224.