LattesMiner: uma linguagem de domínio específico para extração automática de informações da Plataforma Lattes

Alexandre Donizeti Alves[1], Horacio Hideki Yanasse[1], Nei Yoshihiro Soma[2]

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

[2]Instituto Tecnológico de Aeronáutica (ITA)
12.228–010 – São José dos Campos – SP – Brasil

alexandre.alves@inpe.br, horacio@lac.inpe.br, soma@ita.br

Abstract. *The Lattes CV system, a curricular information system maintained by CNPq, is the main component of the Lattes Platform. Currently, the system stores around 2.000.000 curricula of professionals from diverse areas of knowledge. This system is undoubtedly a major source of information on Brazilian researchers and it has a very high information extraction potential. This paper describes "LattesMiner", an internal multilingual domain-specific language for automatic information extraction and identification of academic social networks from Lattes curricula.*

Resumo. *O sistema Currículo Lattes, que é um sistema de informação curricular mantido pelo CNPq, é o principal componente da Plataforma Lattes. Atualmente, o sistema armazena mais de dois milhões de currículos de profissionais das diversas áreas do conhecimento. Esse sistema é hoje, sem dúvida, a principal fonte de informações sobre os pesquisadores brasileiros e tem um elevado potencial para extração de informações. Este artigo descreve "LattesMiner", uma linguagem de domínio específico interna e multilíngue para extração automática de informações e identificação de redes sociais acadêmicas a partir de currículos Lattes.*

Palavras-chave: *Linguagem de Domínio Específico, Extração de Informação, Plataforma Lattes*

1. Introdução

A Plataforma Lattes (PL) é um sistema de informação desenvolvido e mantido pelo CNPq para gerenciar informações relacionadas a pesquisadores e instituições no país. Recentemente, a PL foi citada como exemplo de banco de dados completo e altamente qualificado em um artigo publicado na Nature [Lane 2010]. O principal componente da plataforma é o sistema Currículo Lattes, que é um sistema de informação curricular. Atualmente, o sistema armazena mais de dois milhões de currículos de pesquisadores, docentes, estudantes e profissionais das diversas áreas do conhecimento que atuam em ciência, tecnologia e inovação, principalmente no Brasil. Percebe-se, portanto que esse sistema tem um elevado potencial para extração de informação.

Este artigo descreve "LattesMiner", uma linguagem de domínio específico (LDE) para extração automática de informações de currículos Lattes. As informações extraídas podem ser analisadas e usadas, por exemplo, para identificar redes sociais acadêmicas, competências regionais, perfil de grupos de diferentes áreas de pesquisa e conhecimento etc.

2. Trabalhos Relacionados

De nosso conhecimento, há duas ferramentas atualmente que permitem extrair informações do sistema Currículo Lattes de forma automática: Lattes Extrator e scriptLattes.

Lattes Extrator é uma ferramenta acessível via Web[1] que foi desenvolvida pelo próprio CNPq e é uma das ferramentas que compõe a PL. O acesso é restrito a instituições licenciadas que podem extrair somente informações de seus pesquisadores, docentes, estudantes e colaboradores. As informações são extraídas diretamente do banco de dados do sistema Currículo Lattes e disponibilizadas em arquivos no formato XML. Dessa forma, as instituições precisam 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.

scriptLattes é um *script* desenvolvido em Python para extração e compilação automática de produções bibliográficas, produções técnicas, produções artísticas, orientações, projetos de pesquisa, prêmios e títulos, grafos de colaborações e mapa de geolocalização de um grupo de pesquisadores cadastrados no sistema Currículo Lattes [Mena-Chalco and Junior 2009].

Na revisão de literatura realizada não foi encontrada nenhuma biblioteca ou LDE que extraia informações de currículos Lattes e/ou identifique relacionamentos entre eles, principalmente quando se levam em consideração grupos com muitos pesquisadores.

3. Linguagem LattesMiner

LattesMiner é uma LDE interna para extração automática de informações de currículos Lattes. É composta por um conjunto de classes escritas em Java que permite que outros desenvolvedores implementem suas próprias aplicações com alto nível de abstração e poder de expressão [Alves et al. 2011a].

A linguagem LattesMiner faz parte de um projeto maior denominado "Sistema Unificado de Currículos e Programas: Identificação de Redes Acadêmicas - SUCUPIRA"[2]. O projeto SUCUPIRA, cujo acrônimo traz a lembrança do sobrenome do falecido Professor Emérito da Universidade Federal do Rio de Janeiro, Newton Lins Buarque Sucupira, relator do importante Parecer 977/65 sobre a Pós-Graduação, visa ser uma ferramenta computacional automatizada e de domínio público que possa eventualmente auxiliar na obtenção de indicadores de desempenho de docentes, pesquisadores e programas de pós-graduação.

3.1. Projeto

A primeira tarefa no projeto de uma LDE é definir os termos do problema [van Deursen et al. 2000]. Vale a pena mencionar que, embora o currículo Lattes esteja disponível em Português, também é possível disponibilizá-lo em Inglês. Além disso, o sistema Currículo Lattes já está sendo usado em outros idiomas e países, como Argentina, Chile, Colômbia, Cuba, Equador, México, Panamá, Paraguai, Peru, Portugal e Venezuela.

[1] http://lattesextrator.cnpq.br/lattesextrator/

[2] processo Capes 23038-029609/2008-02

Assim, quando a LDE LattesMiner foi projetada, estes fatos foram levados em consideração. Atualmente, a linguagem LattesMiner permitir usar os principais termos do currículo Lattes em Português e Inglês, sendo facilmente estendido para outros idiomas. A definição dos termos do problema é muito importante, pois os mesmos são considerados e usados para projetar a LDE, que deve descrever concisamente aplicações de um domínio particular (nesse caso, do currículo Lattes), permitindo uma solução no idioma e no nível de abstração do domínio do problema.

3.2. Implementação

LattesMiner é composta por um conjunto de classes escritas em Java e sua classe principal fornece a maioria das funcionalidades da LDE. Foi implementada usando uma interface fluente, que fornece uma representação compacta e fácil de ler do domínio do problema. Interfaces fluentes são implementadas usando o método de encadeamento (*method chaining*). É importante notar que o método de encadeamento por si só não é suficiente para criar uma LDE. Por exemplo, a classe `StringBuilder` da linguagem Java tem um método `append()` que sempre retorna uma instância da própria classe. Porém, ela não resolve o problema de um domínio especifíco, portanto não é uma LDE.

Também é interessante notar que usando o método de encadeamento, qualquer método da linguagem LattesMiner pode ser usado em qualquer ordem e várias vezes [Ruiz and Bay 2008]. Além do método de encadeamento, a linguagem LattesMiner também faz uso de métodos estáticos que permitem criar códigos mais compactos e ainda assim legíveis.

Uma questão que mereceu uma atenção especial foi a extração de informações em currículos Lattes. Inicialmente, foi constatado que o currículo Lattes baixado como arquivo HTML não era balanceado e portanto, não foi possível utilizar um *parser*. Porém, 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 essas razões a técnica de extração de informação baseada em expressões regulares foi usada.

4. Exemplo de Uso

Nesta Seção será apresentado um estudo de caso com os pesquisadores que receberam o Prêmio Anisio Teixeira em 2011, mostrando passo a passo como usar as principais funcionalidades da linguagem LattesMiner.

O primeiro passo é criar um arquivo texto contendo o nome dos pesquisadores. Neste caso, foi criado o arquivo “nomes.txt” contendo cada nome em uma linha separada.

Passo 1 [nomes.txt]

```
Nelson Maculan Filho
Luiz Bevilacqua
Fernando Galembeck
Alvaro Toubes Prata
João Fernando Gomes de Oliveira
```

O próximo passo é obter o número (ID) dos pesquisadores. A listagem a seguir apresenta o código de uma aplicação Java para descobrir o número (ID) dos pesquisadores usando a linguagem LattesMiner.

Passo 2 [Exemplo.java]

```
import java.util.*;
import lattes.util.Util;
import static lattes.miner.LattesMiner.*;

public class Exemplo
{
 public static void main(String[] args)
 {
  List<String> list = new ArrayList<String>();

  for (String nome : Util.getList("nomes.txt"))
   list.add(search(nome));

  Util.setList(list, "ids.txt");
 }
}
```

O método `search()` realiza a busca pelo nome do pesquisador no sistema Currículo Lattes. Se for encontrado, é retornado o número (ID) do pesquisador. Caso contrário, é retornado o nome do pesquisador. Nos casos em que mais de um currículo com o mesmo nome é encontrado, são retornados todos os números (ID) concatenados e separados por vírgula. Assim, é possível verificar no arquivo gerado se algum problema ocorreu. O resultado foi armazenado em um arquivo texto denominado "ids.txt". Neste caso, todos os números (ID) foram encontrados, conforme apresentado a seguir.

SAÍDA [ids.txt]

```
K4783153E3
K4787137U2
K4787937A7
K4781599Z8
K4787011P6
```

Em seguida, a lista de números (ID) gerada anteriormente é lida e o currículo Lattes correspondente é baixado. Para baixar o currículo do pesquisador identificado é usado o método `download()`. O método `save()` armazena o currículo baixado como arquivo HTML e o número (ID) do pesquisador é usado como nome de arquivo. O método `dir()` é opcional e permite definir um diretório no qual o currículo baixado é armazenado. Se o diretório não existir, ele é criado automaticamente. O trecho de código a seguir ilustra como isso pode ser feito.

Passo 3

```
dir("cvs");
for (String id : Util.getList("ids.txt"))
 download(id).save();
```

Após executar esses passos, é possível extrair as informações dos currículos Lattes. A lista de números (ID) dos pesquisadores é lida novamente, conforme ilustrado no trecho de código a seguir.

Passo 4

```
props("mysql");
for (String id : Util.getList("ids.txt"))
{
 load(id).biodata().address();
 areas().formations().languages();
 publications(JOURNAL).publications(CONFERENCE);
 advisories().boards().save();
}
```

O método `load()` é usado para carregar o arquivo HTML do pesquisador em questão na memória como uma string. Assim, é possível usar qualquer um dos métodos disponíveis na linguagem LattesMiner para extração de informações. Por exemplo, o método `boards()` permite extrair todas as participações em bancas de um determinado pesquisador, tanto em nível de mestrado quanto doutorado. É importante destacar que a ordem dos métodos é indiferente, pois cada um deles retorna uma instância da própria classe principal (**`LattesMiner`**), permitindo o encadeamento de métodos.

O método `save()` usado nesse trecho de código tem uma funcionalidade diferente do mesmo método usado no Passo 3. Este método armazena as informações extraídas, de acordo com os métodos de extração usados, em um banco de dados definido em um arquivo de propriedades (por exemplo, `mysql.properties`). O arquivo de propriedades é definido através do método `props()`. Outra possibilidade seria armazenar as informações extraídas em arquivos XML. Nesse caso, o método `xml()` seria usado ao invés do método `save()`.

Depois que as informações extraídas estão armazenadas em um banco de dados, outras consultas podem ser feitas e informações diferentes podem ser obtidas. Apesar de ser possível obter essas informações diretamente dos currículos Lattes armazenados como páginas HTML, isso não é viável principalmente quando o grupo que está sendo analisado contém muitos pesquisadores.

Diversas outras informações podem ser obtidas com a Linguagem LattesMiner. O exemplo apresentado aqui pode ser generalizado para qualquer grupo de pesquisadores cadastrados na PL.

5. Conclusões e trabalhos futuros

Atualmente o currículo Lattes está disponível em HTML, o que impõe um esforço adicional para a extração de informações. A linguagem LattesMiner é independente do formato dos dados contidos no currículo Lattes e permite que os usuários programem suas próprias aplicações com um alto nível de abstração e poder de expressão. Se o formato de dados for eventualmente modificado, a interface da linguagem permanecerá a mesma. Além disso, é possível realizar a busca também pelo nome do pesquisador, o que não ocorre com o Lattes Extrator e o scriptLattes.

A linguagem LattesMiner foi usada para o desenvolvimento do sistema SUCUPIRA [Alves et al. 2011c] e também para analisar o perfil dos bolsistas de produtividade em pesquisa de outra subárea do conhecimento [Alves et al. 2011b]. A principal contribuição deste trabalho é permitir analisar grupos de pesquisadores cadastrados na PL, de forma rápida e automática e em diferentes períodos de tempo.

O próximo passo que já está sendo implementado na linguagem LattesMiner é a análise de redes sociais acadêmicas. A versão beta da linguagem LattesMiner estará disponível em breve para testes. Para usar a linguagem será necessário apenas importar a biblioteca (`LattesMiner.jar`) e a biblioteca para banco de dados (por exemplo, `mysql-connector-java-5.1.21-bin.jar`), caso o usuário queira armazenar os dados em um banco de dados.

Referências

Alves, A. D., Yanasse, H. H., and Soma, N. Y. (2011a). Lattesminer: a multilingual dsl for information extraction from lattes platform. In *11th SPLASH Workshop on Domain-Specific Modeling (DSM'11)*, pages 85–92.

Alves, A. D., Yanasse, H. H., and Soma, N. Y. (2011b). 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*, Ubatuba, SP.

Alves, A. D., Yanasse, H. H., and Soma, N. Y. (2011c). Sucupira: a system for information extraction of the lattes platform to identify academic social networks. In *6th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 371–376, Chaves, Portugal.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288):488–489.

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.

Ruiz, A. and Bay, J. (2008). An approach to internal domain-specific languages in java. `http://www.infoq.com/articles/internal-dsls-java`.

van Deursen, A., Paul, K., and Joost, V. (2000). Domain-specific languages: an annotated bibliography. *ACM SIGPLAN Notices*, 35(6):26–36.