Geração Automática de Metadados

José David Fernández Curado
Orientadora: Renata Wassermann
Instituto de Matemática e Estatistica - Universidade de São Paulo

IME-USP

Introdução

Ao enviar uma edição de uma publicação eletrônica, para ser disponibilizada no site do *"Project Euclid"*, os editores precisam produzir um arquivo contendo diversos dados sobre o conteúdo que está sendo enviado, entre eles:

- Para cada artigo: nome do autor, titulo do artigo, abstract e palavras-chave
- Para cada volume (ou edição): identificadores, titulo do volume, editores

O objetivo principal do trabalho é desenvolver uma ferramenta que possibilite geração automática desses dados, a partir dos arquivos da publicação, no formato desejado pelo *"Project Euclid"*.

Metadados

A definição mais comum de metadados é "dados sobre dados". São comumente encontrados associados a:

- músicas: artista, álbum, gênero musical, ano, letra
- livros: autor, editora, ano, sinopse
- artigos: autor, título, abstract, palavras-chaves

A forma mais usada para representar metadados é atraves de *"XML"*.

Técnicas de Extração de Metadados

Dentre as técnicas mais utilizadas para implementar a extração automática de Metadados estão:

- Raciocínio baseado em regras
- Raciocínio baseado em casos
- Processamento de Linguagem Natural
- Expressões Regulares

Raciocínio baseado em regras

Os geradores que utilizam "Raciocínio baseado em regras" usam regras definidas à priori (possivelmente por um especialista) para chegar a conclusões sobre o documento analisado.
Um dos sistemas estudados [2] classifica tanto linhas como palavras através de regras extraídas de vários domínios:

- Classificação independente de linhas: classifica cada linha independentemente
- Classificação contextual de linhas: melhora a classificação de cada linha usando informações de linhas proximas

Após a etapa de classificação de linhas, o sistema possue uma estratégia para extrair os dados a partir das frases e da classificação

Raciocínio baseado em casos

Os geradores que utilizam "Raciocínio baseado em casos" usam um conjunto de soluções conhecidas do problema para chegar na solução do problema atual.
Outro sistema estudado [3] utiliza um método baseado em regras para agrupar documentos semelhantes, e aplica, então, "Raciocínio baseado em casos" para resolver os problemas semelhantes.
Sistemas deste tipo são capazes de aprender novas resoluções e melhorar sua precisão ao longo do tempo, oferecendo algumas vantagens em relação a sistemas baseados em regras.

Processamento de Linguagem Natural

O objetivo de um sistema de "Processamento de Linguagem Natural" é processar documentos e extrair deles informação.
Existem várias etapas no processamento de linguagem natural [1]:

- Análise Morfológica: Identificação de palavras isoladas
- Análise Sintática: Aplicação dos conhecimentos da Gramática da linguagem
- Análise Semântica: Identificar o sentido das palavras
- Pragmática: Análise do significado em contexto mais geral (sentido da parte quando relacionada com o todo)

O uso de *PNL* pode ser uma opção até obvia ao se estudar Extração de Metadados, mas tem um problema inerente à um sistema para reconhecer uma linguagem, que é a restrição de só trabalhar com a linguagem para a qual o sistema foi programado.

Expressões regulares

Por fim, uma ferramenta muito usada para extração de metadados é "Expressões regulares".
Apesar de ser uma ferramente bem restrita, pode funcionar muito bem para extração de dados que possuem um formato incomum, como emails e telefones. Por exemplo:

- encontrando telefones brasileiros: "$([0-9]\{4\}[-]?[0-9]\{4\})$"
- encontrando emails: "$([a-zA-Z0-9\_.]*@[a-zA-Z0-9\_.]*)$"

Método proposto

Muitos dos métodos estudados supõem que terão de analisar os dados em sua forma mais desestruturada, isto é, são sistemas que tem por objetivo conseguir extrair metadados de arquivos em formatos como pdf ou PostScript, ou ainda de emails entre pessoas.
Neste projeto, supomos que a ferramenta irá processar arquivos TEX, portanto podemos esperar que alguns dados estejam definidos na estrutura do arquivo. Uma regra muito simples para encontrar o autor de um arquivo TEXé simplesmente buscar pelo "author".

Método proposto

```
1 \documentclass{article}
2 \author{Anderson}
3 \title{Matrix - A Computer generated world}
4 \begin{document}
5 \begin{abstract}
6 The Matrix is a Computer generated World to
7 trap humans and turn them into...
8 \end{abstract}
9 \end{document}
```

No exemplo, depois de realizar uma análise do arquivo TEX, extrair os metadados título, autor e abstract, se resume a buscar numa árvore sintática, os nós correspondentes ao title, ao author e ao abstract. Podemos gerar, então, o arquivo com os metadados:

```
1 <?xml-stylesheet type="text/dtd" href="euclid_issue.dtd"?>
2 <euclid_issue>
3   <header>...</header>
4   <issue>
5     <issue_data>...</issue_data>
6     <record type="frontmatter" lang="EN">
7       <title>Matrix - A Computer generated world</title>
8       <author>
9         <name>
10          <given>Anderson</given>
11        </name>
12      </author>
13      <abstract>
14        <p>
15        The Matrix is a Computer generated World to trap
            humans and turn them into...
16        </p>
17      </abstract>
18    </record>
19  </issue>
20 </euclid_issue>
```

Bibliografia

- de Oliveira, F. A. D.
  Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa.
- Han, H., Giles, C., Manavoglu, E., Zha, H., Zhang, Z., and Fox, E.
  Automatic document metadata extraction using support vector machines.
- Khankasikam, K.
  A hybrid case-based and rule-based for metadata extraction on heterogeneous thai documents.