Geração Automática de Metadados

José David Fernández Curado

Instituto de Matemática e Estatística - Universidade de São Paulo

November 17, 2010

"Projeto Euclid"

CLE-eprints

A revista "CLE-eprints"[a], deseja incluir sua revista ao *"Projeto Euclid"*. Para isso precisa gerar metadados referentes às suas edições

[a] http://www.cle.unicamp.br/e-prints/

Quais metadados? Exemplos:

- Para cada artigo: autor, título, número de páginas, palavras-chave.
- Para cada edição: numero da edição, editores.
- Para a revista: informações de contato.

Expectativas

Objetivo do trabalho

Para ajudar o CLE a gerar os dados exigidos pelo *"Projeto Euclid"*, propomos escrever uma ferramenta que consiga produzir automaticamente os metadados e escreve-los no formato adequado.

O que são Metadados?

Definição

- Dados sobre dados
- Dados referentes a um documento

Representação

São comumente representados em arquivos XML, associado a um schema, ou dtd, especificando a sua estrutura.

Exemplos

- musicas: nome do compositor, álbum, ano, gênero musical.
- livros: nome do autor, editora, ano do lançamento, gênero literário.
- artigos: nome do autor, abstract, palavras-chave.

Utilidade de Metadados

Para a internet

- Busca semântica na Web

Para o *"Projeto Euclid"*

- Busca de artigos: A busca do *"Projeto Euclid"*é feita pelas palavras-chave.
- Exibição de informação: Ao apresentar um artigo, informações como autor e abstract são extraidas do arquivo de metadados.

Estratégias

Mais usadas

1. Raciocínio baseado em regras.
2. Raciocínio baseado em casos.
3. Processamento de Linguagem Natural.
4. Expressões Regulares

Utilizada neste projeto

Neste projeto decidimos por uma estratégia mais simples baseada em um processamento do arquivo TEXcom extrações através de expressões regulares em textos já processados.

Raciocínio baseado em regras

Definição

Um sistema de raciocinio baseado em regras utiliza um conjunto de regras de inferencia para chegar a conclusões sobre um conjunto de premissas.

Na prática

São sistemas que permitem definição de uma base de conhecimento do tipo "If A then B else C", onde A é uma condição e B e C são ações que o sistema pode executar.

Aplicações

- Sistemas Especialistas: Usa regras para fazer deduções ou escolhas. Exemplo: sistema que ajuda um médico a fazer um diagnóstico.

Exemplo de uso

Extração de Caracterísica

Em um dos sistemas de extração de metadados estudados [HGM+03], o algoritmo utiliza regras extraídas de bancos de dados e de propriedades ortográficas para definir caracterisicas de uma palavra.
Essa classificação será usada no algoritmo para classificar linhas em separado e em contexto, para, por fim, conseguir extrair os metadados que estejam na linha.

Algoritmo

1. Extração de características das palavras (baseado em regra)
2. Classificação Independete das linhas
3. Classificação Contextual das linhas
4. Quebra das linhas e extração dos metadados

Vantagens e Desvantagens

Vantagens

- Implementação “direta”, sem necessidade de treinamento.

Desvantagens

- Regras precisam ser definidas à priori por um especialista.
- Limitado ao problema definido pelas regras

Raciocínio baseado em casos

Definição

Raciocínio baseado em casos é uma tecnica de Inteligência artificial que utiliza soluções conhecidas de um problema para chegar a solução de outros problemas semelhantes.

Aplicações

- Sistemas de perguntas e respostas [HSYY02]

Raciocínio baseado em casos

Na prática

São sistemas que utilizam uma base de soluções conhecidas e, ao receber um novo problema, utilizam as soluções conhecidas para montar um solução para o novo caso. Basicamente:

1. Retrieve: busca casos similares.
2. Reuse: reusa as informações sobre a solução dos casos resuperados.
3. Revise: revisa a solução proposta.
4. Retain: guarda as partes da experiência que poderão ser úteis no futuro.

Exemplo de uso

Extração de Metadados em Documentos Semelhantes [Kha10]

No sistema proposto em [Kha10], o algoritmo utiliza regras para agrupar documentos semelhantes e, então, aplica raciocinio baseado em casos para fazer a extração dos metadados.
Neste caso, a extração dos metadados está no algoritmo de Solução de Problemas, definindo a solução como um documento com marcações onde estão os metadados.

Algoritmo

1. Agrupamento de documentos semelhantes (baseado em regra)
2. Análise e marcação do documento (baseado em caso)
3. Extração dos metadados
4. Verificação e correção dos metadados extraídos

Vantagens e Desvantagens

Vantagens

- O sistema é capaz de aperfeiçoar suas soluções com o passar do tempo

Desvantagens

- O sistema precisa receber casos relativamente semelhantes

PLN

Definição

Um sistema de Processamento de Linguagem Natural (*PNL*) é um sistema que processa documentos com o objetivo de extrair informação

Aplicações [dO]

- Em textos: Buscas em bases de dados
- Em dialogo: Interfaces de Linguagem Natural

PLN

Na prática [dO]

O sistema guarda diversas informações das frases para interpretar o seu significado. Para isso realiza diversas análise no texto, como:

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

Exemplo de uso

Extração de Preferencias de Usuários [PYB+]

No sistema proposto em [PYB+], o algoritmo utiliza *PLN* para extrair preferências de usuários através de emails trocados com seus analistas financeiros.

Algoritmo

1. Identificação de orações
2. Análise sintática
3. Análise morfológica
4. Identificação de conceitos de várias palavras
5. Categorização de conceitos
6. Metadados implicitos
7. Extração de metadados

Vantagens e Desvantagens

Vantagens

- É uma forte ferramenta para a obtenção de metadados complexos

Desvantagens

- O sistema é limitado para uma linguagem especifica

Sobre "Regexp"

Ferramenta

Expressões Regulares certamente são uma forte ferramenta para a extração de dados que possuem uma estrutura incomum.

Exemplos

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

Algoritmo

Descrição

Propomos um algoritmo que recebe os arquivos TEXreferentes aos artigos e gera o arquivo de metadados correspondente.

Passos

1. Parsing dos arquivos TEX
2. Análise da Estrutura Abstrata do TEX
3. Geração do arquivo de metadados

Parsing

Gramática do TEX

O primeiro passo do algoritmo é realizar um parsing do arquivo TEX
Para executar essa etapa, desenvolvemos um parser incompleto de TEX, que é capaz de extrair corretamente, pelo menos, a estrutura básica do arquivo.
Consideramos que o preâmbulo e o começo do documento contém a maior parte dos metadados desejados.

Análise e Busca

Extração

A extração de metadados é feita com base na estrutura do arquivo TEXe com expressões regulares.
Consideramos que o autor irá usar os comandos TEXpara especificar título, autor e outras informações.

Outros Metadados

Alguns metadados não estão presentes nos arquivos em si. Estes serão fornecidos de outra forma, provavelmente como uma configuração do programa.

Geração do arquivo

Gerando XML

Várias ferramentas foram estudadas para a geração do XML.
A escolha inicial é gerar o XML de uma forma "hard coded", baseado na versão atual do DTD "euclid_issue.dtd" e extender, em uma versão futura da ferramenta, para geração de XML dado um DTD qualquer.

Possibilidades futuras para a ferramenta

Algoritmos

Seria interessante implementar outros algoritmos de extração de metadados para complementar os dados extraídos.

- Existem ótimos algoritmos de extração de palavras-chave que podem beneficiar os metadados referentes aos artigos
- Os metadados extraídos podem estar incompletos, existem possibilidades de extender o programa para utilizar mais estratégias ou estratégias melhores

Referência


Fabio Abreu Dias de Oliveira, *Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa*.


Hui Han, C.L. Giles, E. Manavoglu, Hongyuan Zha, Zhenyue Zhang, and E.A. Fox, *Automatic document metadata extraction using support vector machines*.


Peng Han, Rui-Min Shen, Fan Yang, and Qiang Yang, *The application of case based reasoning on q&a system*.


Krisda Khankasikam, *A hybrid case-based and rule-based for metadata extraction on heterogeneous thai documents*.


Woojin Paik, Sibel Yilmazel, Eric Brown, Maryjane Poulin, Stephane Dubon, and Christophe Amice, *Applying natural language processing based metadata extraction to automatically acquire user preferences*.