

Extração Semântica de Dados Semi-estruturados Através de Exemplos e Ferramentas Visuais

Iraci C. da Silveira

Universidade Federal do Rio Grande do Sul - UFRGS

Instituto de Informática

Porto Alegre – RS, Brasil, CEP 91501.970

iraci@unimed-ners.com.br Caixa Postal 15064

Carlos A. Heuser

Universidade Federal do Rio Grande do Sul - UFRGS

Instituto de Informática

Porto Alegre – RS, Brasil, CEP 91501.970

heuser@inf.ufrgs.br Caixa Postal 15064

ABSTRACT

Nowadays there is a latent necessity of researching and manipulating information, including the available one on WWW(World Wide Web). These types of data are semistructured because they do not have an explicit and regular structure, what makes difficult to manipulate them. This article deals with a tool whose aim is to realize the extraction of semistructured data, focalizing more specifically on HTML (HyperText Markup Language). This tool utilizes the method of semantic extraction which links the text information up to the shaped objects in ontology. The user specifies, through a visual interface, an example of hierarchic structure of the document and its relationship with the concepts of ontology, which generates a delineated grammar of its implicit structure. From this grammar the tool realizes the extraction of further documents in an automatic way, restructuring the result in a data regular format, so called XML (eXtensible Markup Language).

Keywords: semistructured data, data extraction, semantic extraction, XML, ontology, grammar.

RESUMO

Atualmente, existe uma necessidade latente de pesquisar e manipular informações, inclusive as disponíveis na WWW(World Wide Web). Estes tipos de dados são semi-estruturados, pois não possuem uma estrutura explícita e regular, dificultando sua manipulação. Este artigo trata de uma ferramenta para realizar a extração de dados semi-estruturados, focando mais especificamente fontes HTML(HyperText Markup Language). Esta ferramenta emprega o método de extração semântica que associa as informações do texto aos objetos modelados em uma ontologia. O usuário especifica através de uma interface visual um exemplo da estrutura hierárquica do documento e de seu relacionamento com os conceitos da ontologia, gerando uma gramática descritiva da estrutura implícita do mesmo. A partir desta gramática, a ferramenta realiza a extração dos próximos documentos de forma automática, reestruturando o resultado em um formato regular de dados, neste caso, XML (eXtensible Markup Language).

Palavras-chave: dados semi-estruturados, extração de dados, extração semântica, XML, ontologia, gramática.

1 INTRODUÇÃO

Nos dias atuais existe uma necessidade latente: pesquisar, filtrar e manipular informações relevantes. Esta necessidade atinge os mais diversos níveis de organização, desde o indivíduo até grandes corporações. Neste contexto, a WWW (World Wide Web) aparece como o principal veículo para troca e pesquisa de informações.

Com este cenário fica comprovado que já não são mais suficientes às pesquisas operadas sobre bancos de dados, na maior parte das vezes relacional, onde a informação é estruturada e regular, possibilitando consultas rápidas. Uma nova necessidade se estabelece: a manipulação de dados semi-estruturados. Estes tipos de dados se caracterizam por possuírem estrutura irregular e implícita[22,1], e mesmo que possuam alguma estrutura, os dados são heterogêneos. O principal exemplo deste tipo de dado são as páginas HTML(HyperText Markup Language), mas podem ser citados muitos outros, como mensagens de correio eletrônico e arquivos texto.

O desafio atual é construir ferramentas que possibilitem a consulta e manipulação de dados semi-estruturados. Já existe um considerável número de utilitários que permitem realizar consultas por meio de palavras-chave. Porém, estas ferramentas apenas buscam porções do texto de forma fixa, não satisfazendo a necessidade completa de pesquisa, além