



Iam Vita Jabour

O Impacto de Atributos Estruturais na Identificação de Tabelas e Listas em Documentos HTML

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática da PUC-Rio

Orientador : Prof. Eduardo Sany Laber
Co-Orientador: Prof. Raúl Pierre Rentería

Rio de Janeiro
Novembro de 2010



Iam Vita Jabour

O Impacto de Atributos Estruturais na Identificação de Tabelas e Listas em Documentos HTML

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Eduardo Sany Laber

Orientador

Departamento de Informática — PUC–Rio

Prof. Raúl Pierre Rentería

Co–Orientador

Departamento de Informática — PUC–Rio

Prof. Ruy Luiz Milidiú

Departamento de Informática – PUC–Rio

Prof. Rogério Ferreira Rodrigues

Departamento de Informática – PUC–Rio

Prof. Alexandre Plastino de Carvalho

Departamento de Ciência da Computação – UFF

Prof. José Eugenio Leal

Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 25 de Novembro de 2010

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Iam Vita Jabour

Graduou-se em Bacharelado em Informática na PUC-Rio.

Ficha Catalográfica

Jabour, Iam

O Impacto de Atributos Estruturais na Identificação de Tabelas e Listas em Documentos HTML / Iam Vita Jabour; orientador: Eduardo Sany Laber; co-orientador: Raúl Pierre Rentería. — Rio de Janeiro : PUC-Rio, Departamento de Informática, 2010.

v., 64 f: il. ; 29,7 cm

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Tese. 2. extração de informação. 3. extração de tabelas. 4. extração de listas. 5. segmentação de documentos HTML. 6. isomorfismo em árvore. I. Laber, Eduardo. II. Rentería, Raúl. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

Agradecimentos

Aos meus pais, por todo o apoio e carinho.

A minhas irmãs, por serem pessoas especiais em minha vida.

Aos orientadores Prof. Eduardo Laber e Prof. Raúl Rentería, pelo apoio e incentivo, proporcionando uma orientação exemplar e estando ambos sempre disponíveis e presentes durante o desenvolvimento desta dissertação.

Aos professores da banca examinadora, pelas críticas e sugestões que contribuíram para o aprimoramento deste trabalho.

Ao CNPq e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

A todos os meus colegas da pós-graduação, que participaram contribuindo com sugestões e incentivos.

Resumo

Jabour, Iam; Laber, Eduardo; Rentería, Raúl. **O Impacto de Atributos Estruturais na Identificação de Tabelas e Listas em Documentos HTML**. Rio de Janeiro, 2010. 64p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A segmentação de documentos HTML tem sido essencial para as tarefas de extração de informações, como mostram vários estudos na área. Nesta dissertação investigamos a relação entre o documento HTML e sua representação visual, mostrando como esta ligação ajuda na abordagem estrutural para a identificação de segmentos. Também investigamos como utilizar algoritmos de distância de edição em árvores para encontrar padrões na árvore DOM, tornando possível resolver duas tarefas de identificação de segmentos. A primeira tarefa é a identificação de tabelas genuínas, aonde foi obtido 90,40% de F_1 utilizando o corpus fornecido por (Wang e Hu, 2002). Mostramos através de um estudo experimental que este resultado é competitivo com os melhores resultados da área. A segunda tarefa que consideramos é a identificação de listas de produtos em sites de comércio eletrônico, nessa obtivemos 94,95% de F_1 utilizando um corpus com 1114 documentos HTML, criado a partir de 8 sites. Concluimos que os algoritmos de similaridade estrutural ajudam na resolução de ambas as tarefas e acreditamos que possam ajudar na identificação de outros tipos de segmentos.

Palavras-chave

extração de informação; extração de tabelas; extração de listas; segmentação de documentos HTML; isomorfismo em árvore;

Abstract

Jabour, Iam; Laber, Eduardo(Advisor); Rentería, Raúl. **The Impact of Structural Attributes to Identify Tables and Lists in HTML Documents**. Rio de Janeiro, 2010. 64p. MSc Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The segmentation of HTML documents has been essential to information extraction tasks, as showed by several works in this area. This paper studies the link between an HTML document and its visual representation to show how it helps segments identification using a structural approach. For this, we investigate how tree edit distance algorithms can find structural similarities in a DOM tree, using two tasks to execute our experiments. The first one is the identification of genuine tables where we obtained a 90.40% F_1 score using the corpus provided by (Wang e Hu, 2002). We show through an experimental study that this result is competitive with the best results in the area. The second task studied is the identification of product listings in e-commerce sites. Here we get a 94.95% F_1 score using a corpus with 1114 HTML documents from 8 distinct sites. We conclude that algorithms to calculate trees similarity provide competitive results for both tasks, making them also good candidates to identify other types of segments.

Keywords

information extraction; table extraction; list extraction; webpage segmentation; isomorphism;

Sumário

1	Introdução	10
1.1	Definição do problema	11
1.2	Contribuições	13
1.3	Organização da dissertação	15
2	Conceitos básicos	16
2.1	A linguagem HTML	16
2.2	Formas de visualização de um documento HTML	19
2.3	Document Object Model (DOM)	22
3	Abordagem	24
3.1	Algoritmos de similaridade em árvore	24
3.2	O ambiente de experimentação	33
4	Identificação de tabelas	37
4.1	Trabalhos existentes	40
4.2	Métricas de avaliação	42
4.3	Corpus de exploração	43
4.4	Abordagem proposta	43
5	Extraindo listas de produtos em sites de comércio eletrônico	51
5.1	Trabalhos existentes	52
5.2	Corpus de exploração	54
5.3	Métrica	55
5.4	Abordagem proposta	55
6	Conclusões	60
	Referências Bibliográficas	62

Lista de figuras

1.1	Exemplo de segmentos retirado do site do UOL.	11
1.2	Exemplo de identificação de um produto do site bestbuy.com	12
2.1	Fluxo principal do WebKit	19
2.2	Comparação entre o documento HTML e sua árvore DOM	22
2.3	Comparação entre a visualização do documento em um navegador e sua árvore DOM	23
3.1	Exemplo de dois conjuntos generalizadores de tamanho 5	25
3.2	Exemplo de <i>data regions</i> (Regions), retirado de (Liu et. al., 2003)	26
3.3	Ilustração de dois nós generalizadores de tamanho 5	28
3.4	Ilustração de duas árvores A e B retirado de (Yang, 1991)	29
3.5	Ilustração da tabela final do procedimento (C) e do cálculo da primeiro nível de mapeamento (D) retirado de (Yang, 1991)	30
3.6	Procedimento Simple Tree Matching	31
3.7	Procedimento Casamento Simples	31
3.8	Exemplo de estrutura que utiliza um conjunto generalizador de tamanho 2 para apresentar os itens	33
3.9	Diagrama de classes da ferramenta	34
3.10	Diagrama de sequência da aplicação Benchmark	36
4.1	Tabela de difícil compreensão retirada de (Tengli et. al., 2004)	37
4.2	Tabela não genuína com a árvore HTML correspondente à direita	39
4.3	Tabela genuína com a árvore HTML correspondente a direita	40
5.1	Exemplo de uma lista de produtos do site bestbuy.com	52
5.2	Exemplo de uma lista de produtos do site americanas.com	53

Lista de tabelas

4.1	Classificações possíveis de uma tabela para o cálculo das métricas	42
4.2	Resultados sobre o conjunto de treino sem utilizar técnicas específicas para tabelas	44
4.3	Resultados no conjunto de treino utilizando a função razão de linhas (RL) com as técnicas de semelhança de estrutura	46
4.4	Resultado das técnicas sobre o conjunto de teste	47
4.5	Resultados de aprendizado de máquina com validação cruzada sobre o corpus completo	48
4.6	Comparação dos resultados de identificação de tabelas genuínas com os trabalhos relacionados	49
4.7	Tempo de processamento em segundos dos 1393 documentos com os algoritmos propostos	49
5.1	Classificações possíveis de uma tabela para o cálculo das métricas	55
5.2	Testes iniciais com 24 documentos do corpus de treino	56
5.3	Escolha do melhor método com todos os documentos do corpus de treino	57
5.4	Resultados no corpus de teste	57
5.5	Regras específicas sobre o corpus de treino	58
5.6	Resultado F_1 das regras específicas no corpus de teste	59