1
Introdução

As tarefas de extração de informação ganharam destaque durante a última década devido ao grande crescimento dos mecanismos de busca. Com isso, puderam ser observados esforços no contexto de identificação de segmentos de informações, onde o objetivo é delimitar grupos de dados que apresentam alguma unidade lógica ou semântica de informação. Nessa direção podem ser encontrados diversos trabalhos como template (Chakrabarti et. al., 2007, Viera et. al., 2006, Chuang et. al., 2004), notícias (Laber et. al., 2009, Reis et. al., 2004), títulos de notícias (Xue et. al., 2007, Hu et. al., 2005), tabelas (Liu et. al., 2003, Tengli et. al., 2004, Krüpl et. al., 2006) e listas (Zhai et. al., 2005, Liu et. al., 2003).

Essas tarefas facilitam as técnicas de extração de informação, pois torna possível o tratamento específico de um conjunto de dados, diante "aparente" desordem do conteúdo existente na Web. A Figura 1.1 ilustra o cenário onde um documento HTML, que apresenta um artigo, contém diversos segmentos distintos. Vale ressaltar que, conhecendo a classe do segmento, é possível aplicar processamentos específicos ao segmento.

Uma abordagem, aparentemente simples, para resolver o problema de identificação de segmentos é a criação de marcações para cada segmento. Com isso, bastaria verificar a existência da marcação para identificar um segmento. Essa é uma abordagem conhecida e podem ser observados esforços nessa direção, se analisarmos a evolução da linguagem HTML, principalmente nos últimos anos com a HTML5 (Kestern, 2009).

No entanto, essa abordagem pode apresentar problemas como a má utilização de uma marcação. Um exemplo é a marcação de tabela, que tem sido amplamente utilizada para a diagramação do documento, mas seu objetivo é a criação de estruturas tabulares para a exposição de informação, ou, como definido pela World Wide Web Consortium (W3C), para a organização de dados (Chisholm et. al., 2000).

Na tentativa de solucionar problemas, como a má utilização de uma marcação, a W3C apresenta uma série de documentos, explicando o objetivo de cada marcação e qual sua correta utilização. Como exemplo desses documentos,

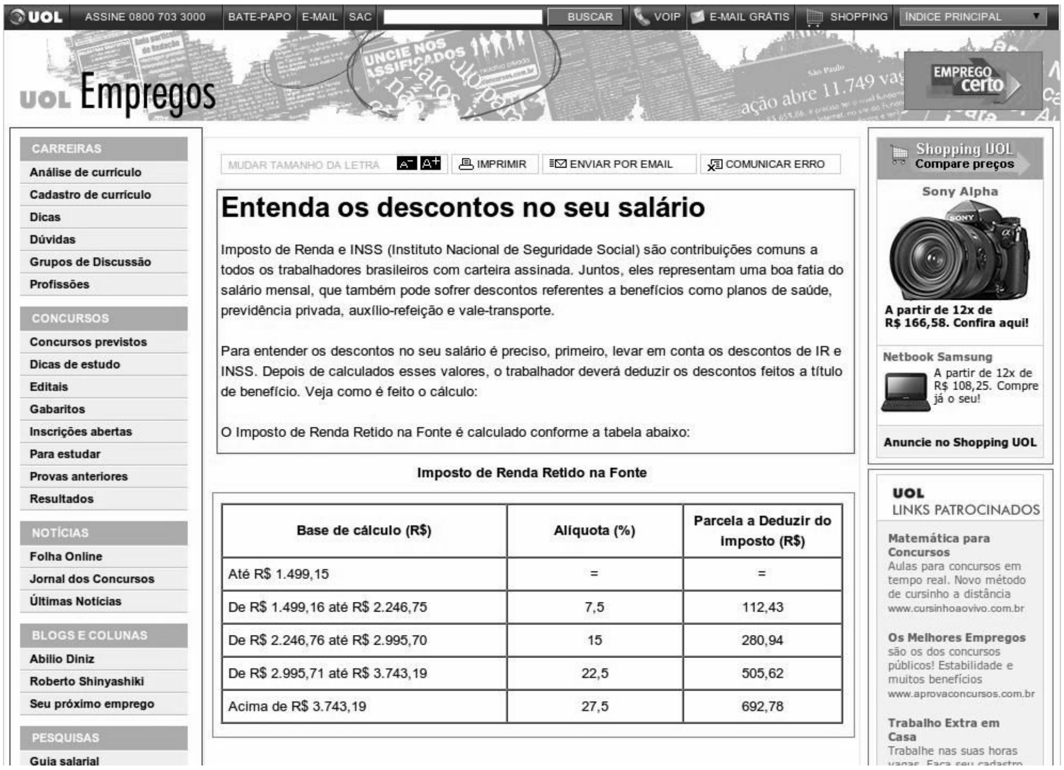

Figura 1.1: Exemplo de segmentos retirado do site do UOL.

destacamos o guia de introdução (Chisholm et. al., 2000), que condena a utilização da marcação de tabela para a diagramação do documento.

No entanto, mesmo com toda essa preocupação e cuidado da W3C, a conformidade com as normas é uma opção do autor durante a autoria de um documento HTML, pois os responsáveis pela exibição dos documentos são os navegadores.

Tendo em vista esse cenário, esta dissertação considera uma abordagem estrutural para a identificação de tabelas e listas. Para isso, é utilizada uma adaptação da abordagem proposta em (Zhai et. al., 2005). Essa abordagem consiste em aplicar algoritmos de isomorfismo em árvores para a identificação de *data records*, listagens que apresentam informações com alguma forma de repetição, como listas (menus, produtos, detre outras) e tabelas. Acreditamos que o entendimento da estrutura do documento HTML contribui para a tarefa de segmentação, proporcionando um avanço interessante para a realização dessa tarefa.

1.1 Definição do problema

Consideramos duas tarefas de identificação de segmentos. A primeira é a identificação de tabelas, que tem o objetivo de identificar a existência de uma tabela em um documento HTML, tornando possível aplicar processamentos específicos à tabela. Já a segunda tarefa é a identificação de listas de produtos em sites de comércio eletrônico, que tem como objetivo identificar as listas que expõem os produtos.

A identificação de segmentos pode ser entendida como a capacidade de encontrar uma ou mais *tags* do documento HTML que representa o segmento desejado. Por exemplo, na Figura 1.2, pode ser observada a estrutura de um documento HTML e sua exibição em um navegador. A capacidade de identificar a região A, como um item da lista de produtos, é equivalente a adicionar um rótulo 'item' à *tag* A'. Com isso, as tarefas de identificação se resumem à capacidade de rotular as *tags* de um documento HTML ou elementos do documento HTML, já que cada elemento representa uma região visual, como será discutido no Capítulo 2.

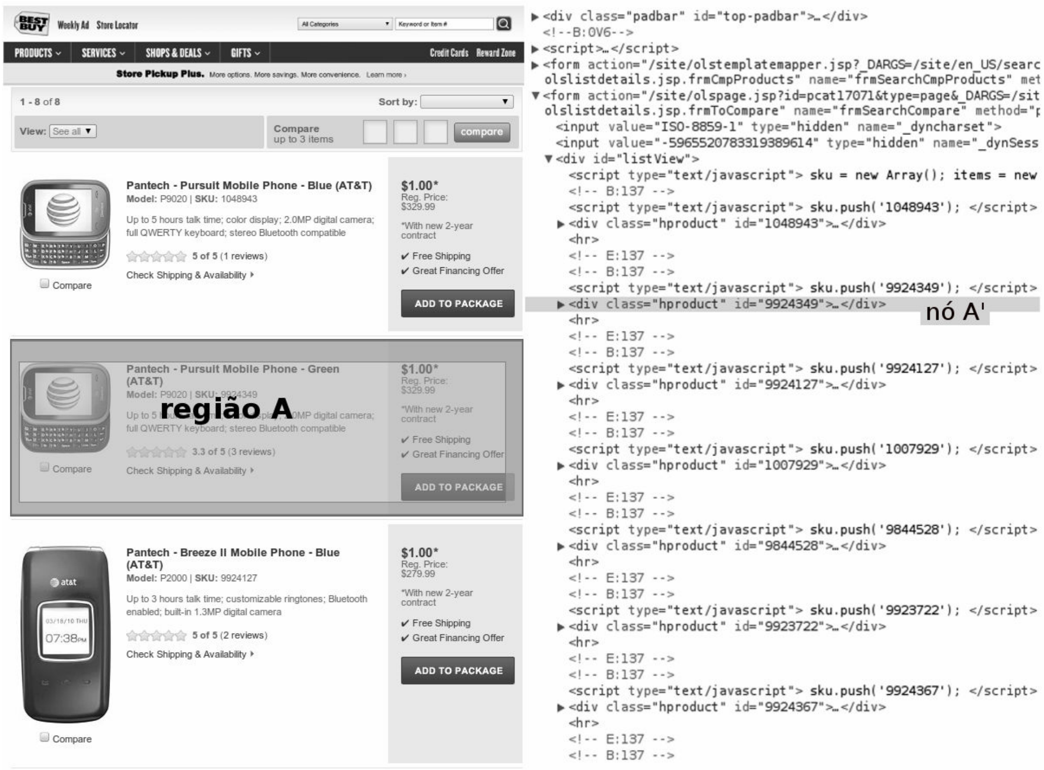

Figura 1.2: Exemplo de identificação de um produto do site bestbuy.com

A identificação de tabelas é uma tarefa conhecida, com diversos trabalhos relacionados (Liu et. al., 2003, Tengli et. al., 2004, Krüpl et. al., 2006), pois agrega informações importantes para as técnicas de recuperação de informação. Essa importância é destacada em (Pinto et. al., 2003), onde são reportados ganhos na tarefa de *question-answering*, quando as tabelas são processadas de forma específica.

Diferentemente da tarefa de identificação de tabelas, a identificação de listas de produtos em sites de comércio eletrônico não é conhecida, no sentido de que não foi possível encontrar trabalhos na literatura direcionados à identificação de listas dessa natureza, ou mesmo de algum tipo específico de lista. Durante esses esforços, foram encontrados trabalhos com o objetivo de identificar listas, de uma forma geral, como o trabalho de Zhai *et al* (Zhai et. al., 2005) que utiliza algoritmos de isomorfismo em árvores para a identificação de listas.

As listas de produtos são segmentos importantes no domínio dos sites de comércio eletrônico. Tais listas apresentam todos os itens existentes em um portal de vendas e resumem de forma coerente a coleção de produtos do portal. Com isso, diminuem o número de requisições necessárias para a obtenção da coleção de itens de um site de comércio eletrônico. Outro ponto favorável é a capacidade de obter algumas informações importantes nessas listas, sem a necessidade de visitar cada produto individualmente, como preço, disponibilidade, promoções e até mesmo a imagem que ilustra um produto. No entanto, a utilidade dessas informações não é abordada, pois o objetivo desta dissertação é apresentar e avaliar a qualidade da identificação dessas estruturas utilizando uma abordagem estrutural.

Uma dificuldade imposta pelas listas de produtos de sites de comércio eletrônico é que essas estruturas apresentam arranjos diferentes. Essas formas variam de site para site e dificultam tanto as técnicas que utilizam a estrutura de marcação do documento HTML, quanto as técnicas que não são direcionadas a um único domínio, como as apresentadas nesta dissertação. Achamos interessante essas dificuldades, pois testam a capacidade de generalização e adaptação das abordagens propostas.

1.2 Contribuições

Inicialmente, estudamos o funcionamento dos navegadores, principais responsáveis pela exibição de documentos HTML, para o entendimento do elo entre a estrutura do documento HTML e sua representação visual. Com isso, foi possível identificar alguns pontos importantes não reportados por trabalhos da área. Um dos pontos que mais chama a atenção é a necessidade de um longo processamento para a obtenção de informações como tamanho de fonte, cor e posicionamento de um elemento HTML na tela, como será discutido no Capítulo 2. Nesse capítulo, discutimos como a árvore DOM, que é uma estrutura básica para representação de documentos HTML, é importante para abordagens que necessitam processar um grande volume de documentos.

Em seguida, avaliamos como a árvore DOM pode contribuir para as tarefas de identificação de segmentos. Optamos por utilizar como base a abordagem proposta em (Zhai et. al., 2005) que apresenta uma busca por estruturas similares (isomórficas) em árvores. Essa abordagem tem como objetivo encontrar os mapeamentos possíveis em uma árvore, identificando as sub-estruturas que apresentam alguma similaridade, chamadas pelo autor de data regions. No entanto, Zhai *et. al.* não aplicam sua técnica sobre um domínio específico, identificando apenas as estruturas definidas como *data regions* que basicamente são listas (qualquer estrutura que apresenta sub-estruturas repetidamente).

Utilizamos algoritmos seguindo a abordagem de Zhai *et. al.*, que buscam semelhança em árvores, para identificar tabelas e listas de produtos. Dividimos a tarefa de identificar subárvores semelhantes em duas etapas. A primeira é chamada de etapa de busca e consiste em encontrar quais subárvores devem ser comparadas de modo a diminuir o número de vezes que o cálculo de distância é realizado. A segunda etapa, chamada de cálculo de distância, tem como objetivo calcular a distância entre duas subárvores.

Para a etapa de busca foi testada inicialmente uma abordagem bastante simples, chamada de casamento simples, onde assumimos que as estruturas similares são sempre enraizadas por um nó. Em seguida, testamos uma abordagem que assume estruturas mais complexas, chamada de casamento de árvores. Foram três as funções de distância utilizadas. Na etapa de cálculo da distância, a primeira função de distância é uma modelagem simples para aplicar a função de distância de Levenshtein em uma árvore. A segunda é uma adaptação da primeira modelagem, onde o objetivo foi balancear o cálculo da distância que apresenta os melhores resultados. A terceira e última função de distância foi proposta em (Yang, 1991) e calcula o maior mapeamento entre duas árvores.

Todas as abordagens diminuiram sempre em mais de 50% o domínio de busca (número de nós). Foi possível observar que a utilização da estrutura do documento HTML para a diminuição do domínio de busca é eficiente. Com isso, realizamos diversos testes para avaliar a necessidade de aplicar algum tipo de pós-processamentos para cada tarefa, já que os algoritmos de similaridade estrutural são gerais e não focam em uma estrutura particular. A criação de pós-processamentos específicos para identificar as tabelas e lista de produtos foi extremamente simples, já que o domínio de busca foi reduzido pelas funções de semelhança estrutural. Com regras simples, como a razão do uso de marcações ou a contagem das imagens na sub-estrutura, foi possível obter bons resultados para as duas tarefas.

Na tarefa de identificação de tabelas, os resultados foram de 90.40% de F1, ficando próximo dos melhores resultados reportados na área (Wang e Hu, 2002), (Gatterbauer e Bohunsky, 2006). Já na tarefa de identificação de listas de produtos em sites de comércio eletrônico, acreditamos que os resultados foram motivadores, chegando à marca de 94.95% de F1. Para a realização dos testes, foram utilizados dois corpora, um para a identificação de tabelas disponibilizado por (Wang e Hu, 2002) e outro para a identificação de listas, criado durante este trabalho.

Concluímos que a abordagem estrutural apresentou bons resultados em ambas as tarefas, sendo os algoritmos propostos de baixo custo computacional e também de fácil codificação. Um ponto positivo para a abordagem estrutural é a capacidade de atender duas tarefas de identificação com o mesmo conjunto de algoritmos, sendo necessário apenas fases de pós-processamento para obter resultados competitivos com os trabalhos existentes na área.

1.3 Organização da dissertação

Esta dissertação está organizada da seguinte forma: no Capítulo 2, discutimos conceitos importantes para a leitura e o entendimento da dissertação. Mais especificamente, discutimos a linguagem HTML e estruturas que são utilizadas para a representação de documentos HTML em memória. Em seguida, no Capítulo 3, são apresentados os procedimentos utilizados para a identificação de tabelas e listas de sites de comércio eletrônico. Nesse capítulo, também é apresentado o *framework* construído para facilitar a resolução das tarefas e os experimentos. No Capítulo 4, apresentamos a tarefa de identificação de tabelas, o estado da arte nessa tarefa e o resultado da aplicação da abordagem proposta. O Capítulo 5 é dedicado à tarefa de identificação de listas de produtos em sites de comércio eletrônico. Finalmente, no Capítulo 6, concluímos a dissertação apresentando algumas análises sobre a abordagem proposta, os desafios encontrados durante a experimentação e também os desafios para melhorar as técnicas apresentadas.