6
Conclusões

As tarefas de segmentação de documentos HTML já mostraram que podem proporcionar melhoras na extração de informação. Além disso, a cada ano são apresentados mais trabalhos direcionados à identificação de segmentos, sendo discutidos novos tipos de segmentos e aprimoradas as técnicas existentes, melhorando os resultados. Acreditamos que esse seja um dos caminhos para a segmentação dos documentos HTML, uma vez que a tarefa estará concluída assim que todos os segmentos de um documento forem identificados. Essa visão vai em direção contrária a de segmentar totalmente um documento HTML, para depois classificar cada segmento. A segmentação por meio de identificação é, aparentemente, um processo menos genérico. Porém, esse tipo de abordagem faz com que os segmentos mais importantes na Web sejam catalogados. Nesta dissertação, apresentamos uma abordagem estrutural para a identificação de um segmento conhecido, que é o segmento de tabela. Também apresentamos uma abordagem para identificar segmentos de listas de produtos, que é importante no domínio de comércio eletrônico.

No Capítulo 2, apresentamos como as informações visuais de um documento HTML exigem um longo fluxo de processamento para que possam ser utilizadas. Destacamos como diversos trabalhos que utilizam informações visuais, muitas vezes, não destacam o custo de obter essas informações. Durante nosso estudo, podemos perceber que a necessidade de utilizar estruturas mais simples é eminente, pois a necessidade de processar diversos documentos é uma realidade na área. Acreditamos que a utilização de atributos obtidos a partir de estruturas mais básicas, como a árvore DOM, seja o caminho natural para os trabalhos de identificação de segmentos.

Em seguida, no Capítulo 3, apresentamos três funções de distância em árvores. Também são apresentados dois algoritmos para realizar as comparações dentro de uma mesma árvore, buscando por similaridades estruturais (isomorfismo). Tanto os algoritmos de distância quanto os de busca de similaridade apresentaram bons resultados durante as fases de experimentação apresentadas nos Capítulos 4 e 5. Diminuir a complexidade desses algoritmos e testar outras maneiras de buscar por similaridades são tarefas que podem ser

exploradas por futuros trabalhos. No Capítulo 3, também discutimos as dificuldades em utilizar os documentos HTML da Web, constatando a existência de diversos problemas de codificação de caracteres e erros de formação desses documentos. Essas dificuldades foram enfrentadas durante o desenvolvimento da ferramenta utilizada para experimentação de nossos algoritmos. Essa ferramenta foi projetada de maneira que seja possível aproveitá-la para realizar outras tarefas de identificação de elementos em documento HTML, sendo uma das contribuições desta dissertação.

Na tarefa de identificação de tabelas, apresentada no Capítulo 4, foi possível observar como a estrutura do documento ajuda na identificação das tabelas. A utilização dos atributos apresentados em (Wang e Hu, 2002) foi um ponto interessante em nossa experimentação, já que demostra a utilização de nossa abordagem como um pré-processamento. Acreditamos que utilizando a estrutura para obter informações iniciais, que facilitam a identificação de segmentos, para depois executar um algoritmo específico, seja uma boa forma de utilizar a estrutura do documento HTML. Infelizmente, não foi possível obter resultados melhores do que o estado da arte nessa área. Porém, nossos resultados são competitivos, pois utilizam atributos que podem ser obtidos com menos esforço computacional e mostram a existência de padrões na estrutura do documento HTML que podem ajudar na identificação de outros tipos de segmentos.

No Capítulo 5, a tarefa de identificação de listas de produtos é apresentada e resolvida utilizando apenas informações presentes na estrutura do documento. Foi interessante observar o comportamento dos algoritmos de similaridade em árvore para essa tarefa, já que as estruturas de listas são variadas. A existência de diversas listas em um documento HTML foi um fato que despertou uma questão interessante: as estruturas de listas conseguem segmentar um documento de maneira adequada? Infelizmente, essa questão não foi investigada, já que, para isso seria necessário um conjunto de experimentação com documentos segmentados. Essa é mais uma pergunta que fica em aberto para trabalhos futuros.