1
Introdução

1.1
Motivação

A internet é a principal fonte de informação de diversas áreas do conhecimento humano. Alguém que deseja pesquisar sobre determinado assunto na internet pode recorrer às populares máquinas de busca, dentre as mais conhecidas podemos citar Google e Yahoo.

As máquinas de busca normalmente recebem uma consulta do usuário e retornam um conjunto de páginas listadas de acordo com a relevância. Para calcular a relevância de uma página muitos fatores são levados em consideração. Um deles é a ocorrência das palavras da consulta na página *web*. No entanto este mecanismo pode retornar um resultado que não seja relevante para o usuário. Um exemplo é quando uma das palavras da consulta está no menu e a outra no corpo informativo da página, portanto a página retornada apresenta uma das palavras em uma seção que auxilia apenas na navegação do usuário, não lhe dando nenhuma informação.

Uma forma de evitar este problema é localizar o conteúdo informativo e remover o conteúdo restante da página. Uma outra forma é extrair os *templates* da página `HTML`. O *template* de uma página é o conteúdo que aparece repetidamente nas páginas de um *site* (GPT05) e que normalmente auxilia o usuário na navegação do *site*. Exemplos de *templates* são cabeçalhos da página, logo da empresa ou menu.

Os autores de (GPT05) estimam que os *templates* representam entre 40% e 50% do conteúdo da Internet. Além disso, os autores também reportam um crescimento no volume de *templates* de aproximadamente 6% ao ano. Assim se o sistema de busca armazenar apenas o conteúdo relevante das páginas ou as páginas com seus *templates* extraídos, a quantidade de dados armazenada reduz de forma significativa.

Páginas com o mesmo conteúdo informativo, porém com estilos de apresentação diferentes podem ocorrer na *web*. A extração do conteúdo relevante ajuda a tarefa de detecção destas páginas, que na verdade são cópias. O ar-

mazenamento de cópias para máquinas de busca desperdiça o recurso físico e também de processamento, visto que ela precisará ser analisada para as consultas dos usuários.

Além do refinamento do *ranking* das páginas, detecção de cópia e economia do armazenamento de dados, a detecção de conteúdo relevante também pode favorecer o agrupamento e a classificação de páginas (YLL03).

No entanto, desenvolver um algoritmo para detecção de conteúdo relevante para diferentes tipos de domínio, como comércio eletrônico ou portal de um *blog*, não é uma tarefa fácil. A internet é composta de diversas páginas com conteúdo e estilos de apresentação variados. Ao considerar apenas uma classe de páginas a tarefa de detectar o conteúdo relevante é mais viável, neste trabalho o foco são as páginas de notícias. Esta classe de páginas tornou-se a fonte de notícias para pessoas no mundo todo, prova disso é que jornais nacionais e internacionais disponibilizam seu conteúdo na internet.

1.2 Definição do Problema

Como foi dito na Seção 1.1 *templates* são o conteúdo que aparece repetidamente nas páginas de um *site*. Veja um exemplo na Figura 1.1, onde os retângulos tracejados indicam exemplos de *templates*. Na região superior existe um cabeçalho, propaganda, o logo do jornal e um menu. Estes elementos são muito comuns em *templates* na internet.

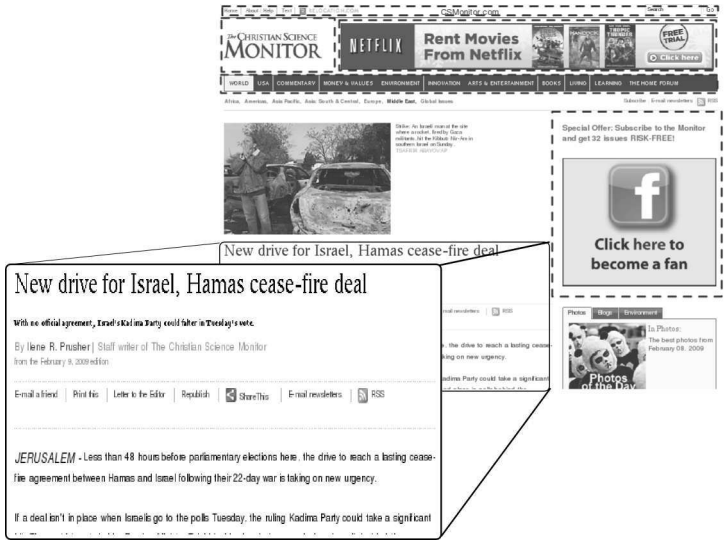

Figura 1.1: Exemplo de template em um página de notícia

Para o nosso estudo, o conteúdo relevante da página é a seção da página que apresenta a notícia, ou seja, o título e o corpo da notícia.

O nosso problema consiste em extrair automaticamente o conteúdo relevante de páginas `HTML` de notícias. No exemplo da Figura 1.1 o segmento ampliado à esquerda é o que gostaríamos de extrair da página.

1.3 Resultados

Nessa dissertação apresentamos o algoritmo `NCE` (*News Content Extractor*). Neste método cada página HTML do conjunto de dados é transformada em uma árvore DOM. Esta árvore tem como nós internos as *tags* e as folhas são textos, imagens ou *links*.

O `NCE` então procura por uma subárvore que contenha uma boa aproximação do conteúdo relevante. Essa procura é guiada por atributos dos nós da árvore DOM, como a razão entre texto e *link*, o tamanho do texto, o número de palavras do título, dentre outros. Após este passo o `NCE` remove os comentários que uma notícia possa ter. Em seguida o algoritmo busca pelo título da notícia. Esta é uma etapa importante, visto que o título constitui uma informação importante e por ser pequeno muitas vezes pode não ser localizado.

Algumas técnicas existentes na literatura utilizam atributos que não foram utilizados no `NCE`, dentre estes podemos destacar o conjunto de atributos visuais dos nós DOM. Estes atributos podem ser capturados quando a página é renderizada.

A técnica implementada por Zheng et. al (ZSW07), chamada de `V-Wrapper`, utiliza os atributos visuais para poder detectar o conteúdo relevante de páginas de notícias. O algoritmo `V-Wrapper` foi implementado para esta dissertação com o objetivo de comparar com o `NCE`. Os resultados mostram que o `NCE` obteve um melhor desempenho na qualidade do conteúdo relevante extraído e no tempo de execução.

1.4 Organização da Dissertação

O próximo capítulo faz um apanhado de alguns trabalhos na área de extração de conteúdo relevante. Eles são descritos resumidamente e agrupados por categoria. O algoritmo apresentado pelos autores de (ZSW07) é descrito em mais detalhes, pois ele foi implementado para comparar com o `NCE`.

No Capítulo 3 o algoritmo `NCE` é apresentado em detalhes. Os atributos dos nós DOM utilizados para extrair o conteúdo relevante são listados. Apresentamos também os parâmetros e seus valores utilizados no algoritmo. Então os passos principais do algoritmo são expostos minuciosamente.

O Capítulo 4 apresenta os dados utilizados para os testes com o `V-Wrapper` e o `NCE`. Em seguida detalhamos os experimentos realizados com os algoritmos e apresentamos os resultados obtidos, assim como uma análise destes.

Por fim, no Capítulo 5, a conclusão deste trabalho é exposta.