5
Conclusão

Nesta dissertação apresentamos o algoritmo NCE para resolver o problema de extração automática de conteúdo relevante de páginas *web* de notícias. O título e o corpo da notícia são os textos da página considerados conteúdo relevante.

Além do NCE foi implementado o algoritmo apresentado por Zheng et. al (ZSW07), chamado de V-Wrapper. Para aferir a qualidade dos dois algoritmos montamos dois corpus. Um dos corpus foi utilizado para inicializar os parâmetros do algoritmo NCE, chamado de corpus de exploração (CE), e o outro corpus, chamado de corpus de avaliação (CE), foi testado já com os parâmetros do NCE fixados. O corpus CE possui 140 páginas de 95 portais diferentes e o corpus CA possui 339 páginas com 23 portais diferentes.

Através do NCE mostramos uma forma simples de procurar pelo conteúdo da página de notícia em um árvore DOM. Esta forma utiliza a hipótese do nó separador, que afirma que existe um nó u na árvore DOM associada a página HTML e um número real tr para o qual a união dos blocos de textos da floresta $F_{u,tr}$ é uma boa aproximação do conteúdo relevante.

Adicionalmente, também mostramos dois módulos de otimização do NCE. Um módulo busca o título da notícia na árvore e o outro remove comentários das notícias. Os resultados apresentados no Capítulo 4 mostram que o NCE alcançou bons resultados. No corpus de avaliação alcançou uma média na medida $F1$ de 0.88 e para o corpus de exploração a média da medida $F1$ foi de 0.9. Ao remover os módulos de otimização os resultados apresentaram uma queda de até 0.2 na média da medida $F1$.

A maioria dos *sites* apresentou bons resultados ao serem processados pelo NCE. A exceção foi o portal www.watfordobserver.co.uk que apresentou um estilo de quebra de linha fora do comum e que fragmentou a notícia além do esperado. Pelos resultados fica claro que este portal é uma exceção e o NCE consegue generalizar seu processo de extração de conteúdo em páginas de notícias de forma satisfatória.

O tempo gasto pelo algoritmo V-Wrapper para renderizar as páginas supera, consideravelmente, o tempo gasto pelo algoritmo NCE. Mostrando que

a escalabilidade necessária em um algoritmo deste tipo não é alcançada.

Alguns detalhes de implementação do `V-Wrapper` também não ficaram claros. Isso pode ter degradado a qualidade do seu classificador de folhas e portanto comprometendo o resultado final.

Os resultados do algoritmo `V-Wrapper` ficaram bem abaixo do algoritmo do `NCE`. Portanto o algoritmo `NCE` oferece uma solução eficiente e de qualidade para o problema apresentado nesta dissertação.