PROCESSAMENTO TEXTUAL EM PÁGINAS DA WEB

Aluno: Pedro Lazéra Cardoso
Orientador: Eduardo Sany Laber

Introdução

Na primeira fase da Iniciação Científica, foi realizado um estudo dos conceitos básicos de algoritmos e programação, visto que o aluno estava dando os primeiros passos na área. Dentre os tópicos estudados, podemos destacar os seguintes: estrutura de dados, algoritmos em grafos e análise de algoritmos. Muitos dos algoritmos estudados foram implementados na linguagem de programação C.

Após essa etapa, o aluno começou a estudar a linguagem de programação Python, visto que esta é bastante prática para processamento de textos. Para praticar a linguagem, o aluno implementou alguns procedimentos simples para extrair conteúdo relevante de tabelas da web.

Na fase atual, estamos trabalhando no sentido de entender e aprimorar a heurística de extração de conteúdo relevante de páginas web, desenvolvida no laboratório em que o aluno realiza sua pesquisa. Os próximos itens dizem respeito a esta fase do projeto.

Objetivos

Desenvolver métodos para identificar conteúdos de texto semelhantes de um conjunto qualquer de árvores DOM. Essa identificação servirá como pós-processamento para a heurística NCE (News Content Extractor) que extrai conteúdo relevante de páginas da web. O objetivo é melhorar a precisão dos resultados dessa heurística.

Metodologia

Dado um conjunto de árvores DOM, é possível localizar os nós de texto de cada árvore através de uma busca em profundidade. Nessa busca, um identificador de cada nó de texto é armazenado numa estrutura de hash. Escolhemos esta estrutura com o objetivo de utilizar pouca memória e ter rápido acesso aos dados armazenados.

Para determinar em que posição da tabela o identificador de um nó é guardado, aplica-se uma função de hash em seu texto. Depois de percorrer todas árvores, é possível verificar se um certo texto está presente em muitas delas analisando a estrutura de hash. A função de hash garante uma alta probabilidade de que elementos associados a uma mesma posição sejam idênticos. Assim, se o identificador de um nó está associado a uma posição da tabela com um grande número de colisões, isso indica que seu texto ocorre em muitas das árvores.

Conjeturando que todo texto presente em muitas árvores é provavelmente conteúdo irrelevante, cada texto com essa característica é removido da árvore. Em seguida, utilizando ferramentas para medir precisão e revocação, podemos descobrir como tal algoritmo afeta o resultado da extração de conteúdo relevante obtido pela heurística NCE.

É importante fazer algumas observações. Primeiro, essas árvores não representam páginas de web inalteradas – cada uma das páginas é parte de uma outra página que o NCE (News Content Extractor) marcou como relevante. Finalmente, esse pós-processamento da heurística foi programado usando o Ptyhon, em razão da facilidade de se escrever códigos com esta tecnologia, de sua eficiência e da experiência que o grupo do laboratório teve com ela em projetos anteriores.

Conclusões

Estudar os conceitos básicos de programação e de algoritmos foi essencial para que o aluno pudesse entender algumas das técnicas utilizadas em processamento textual.

Na fase atual, onde estamos desenvolvendo métodos para filtrar textos que ocorrem muitas vezes em um conjunto de páginas, o aluno está tendo a oportunidade de colaborar diretamente com uma pesquisa na área de extração de conteúdo relevante de páginas.

Ainda não podemos avaliar como a heurística afeta a precisão do NCE, porque estamos na etapa inicial do processo, em que só o esboço do projeto é conhecido. No entanto, é possível afirmar que nossa heurística pode ser estendida para identificar outros objetos diferentes de textos em estruturas diferentes de páginas da web. O que vai determinar o quanto a heurística deve ser modificada são os objetos e as estruturas com os quais ela vai trabalhar.

Referências

1 - KLEINBERG, Jon; TARDOS, Eva. **Algorithms Design**.

2 - TAN, Pang-Ning; KUMAR, Michael Steinbach Vipin. **Introduction to Data Mining**.

3 – TENGLI, Ashwin; YANG, Yiming; MA, Nian Li. Learning **Learning Table Extraction from Examples**. Pittsburgh, PA: Carnegie Mellon University.