PROCESSAMENTO TEXTUAL EM PÁGINAS DA WEB

Aluno: Pedro Lazéra Cardoso
Orientador: Eduardo Sany Laber

Introdução

Na primeira fase da Iniciação Científica, o aluno deu continuidade ao projeto que estava desenvolvendo no ano anterior. O projeto envolvia a extração de entidades na web, como a identificação de título, data, fonte e autor em páginas de notícia. Na etapa a corrente, foi dada prioridade ao estudo de Data Mining, no intuito de entender a construção de modelos de classificação para o problema de extração de entidades.

Entre os tópicos estudados, podemos destacar os conceitos básicos de Data Mining, a construção de Arvores de Decisão e a avaliação de modelos de classificação. Além disso, é importante ressaltar que a linguagem Python foi utilizada para realizar tarefas, entre elas a construção de um corpus que serviu para a avaliação dos modelos de classificação propostos. Para a construção de modelos, demos preferência ao Weka.

Na fase atual, estamos trabalhando no sentido de entender e aprimorar a heurística de extração de conteúdo relevante de página da web, o NCE [3], desenvolvida no laboratório em que o aluno fez sua pesquisa. Os próximos itens dizem respeito a esta fase do projeto.

Objetivos

Desenvolver métodos para identificar elementos de páginas de notícia da web, como título, data, autor e fonte. Essa identificação se baseia na exploração da árvore DOM de uma página da web e nos atributos de cada nó dessa árvore. Ela representa uma tentativa de aprimorar o NCE, News Relevant Content Detector, que anteriormente identificava numa página seu conteúdo relevante, sem distinguir título, data etc.

Metodologia e Desenvolvimento

Dada uma página da web em HTML, pode-se construir sua árvore DOM. Explorando a árvore DOM e também as folhas de estilo (CSS) relacionadas à página, é possível associar a cada nó de texto da árvore uma série de atributos.

Esses atributos são como características dos nós de texto e podem ser úteis para rotular esses nós. Entre os atributos que selecionamos, constam a quantidade de caracteres do texto, o tamanho da fonte absoluto e o relativo, o percentual de caracteres numéricos e outros.

Construção do Corpus

Uma das tarefas da iniciação envolveu a construção de um corpus, o RCD4, composto por 200 páginas de notícia de mais de 30 domínios diferentes, como CNN, BBC, Folha de São Paulo e O Globo. Para cada página, foi gerado um documento XML com o conteúdo relevante dessa página separado pelos campos que queríamos ser capazes de identificar com a classificação: corpo, título, data, autor, fonte. Esse processo foi realizado com uma ferramenta de anotação desenvolvida no laboratório.

Construção e avaliação de modelos de classificação

O RCD4 foi utilizado com dois propósitos: (i) construir um modelo de classificação - uma árvore de decisão - capaz de identificar o título, a data, o autor e a fonte de páginas de notícia da web; (ii) avaliar o modelo construído.

Antes da elaboração da árvore de decisão, foram escolhidos os atributos com os quais o modelo seria feito. Para estabelecer quais atributos eram relevantes, primeiro foi selecionado um conjunto de atributos que a priori poderiam ser úteis. A seguir, esses atributos foram utilizados em pares para a construção de árvores de decisão. Através da comparação da performance desses modelos, foi possível ter uma noção da importância de cada atributo e descartar aqueles que de fato não eram úteis à classificação. Esse processo é necessário principalmente porque o NCE tem como uma das suas principais características a velocidade, tornando-se importante utilizar o menor conjunto possível de atributos.

Para construir um modelo de classificação e, com o mesmo Corpus, poder avaliá-lo, o RCD4 foi dividido em duas partes. Uma delas era utilizada para o treino, ou seja, para a construção do modelo que melhor classifica esse subconjunto de páginas. A outra, formada pelas páginas não fornecidas no treino, era utilizada para a avaliação do modelo. Para evitar modelos viciados, o conjunto de treino e o conjunto de teste continham páginas de domínios distintos.

Também foi cogitado usar o "cross-validation" na construção e na avaliação do modelo, visto que o RCD4 não é um corpus muito extenso. No entanto, o software utilizado nessa etapa não permitia que no "cross-validation" o conjunto de treino e o conjunto de teste tivessem páginas de domínios distintos.

Por fim, é importante fazer algumas observações. Primeiro, essas árvores não representam páginas de web inalteradas – cada uma delas é parte de uma outra página que o NCE (News Content Extractor) marcou como relevante. Finalmente, para a construção de modelos de classificação e posteriormente a avaliação do desempenho desses modelos, utilizamos o software WEKA[2]

Conclusões

Dar continuidade ao estudo os conceitos básicos de programação e de algoritmos, além de possibilitar o entendimento do funcionamento do NCE, foi essencial para que o aluno conhecesse melhor algumas das áreas da Engenharia de Computação. Entre elas, vale destacar a construção de modelos de classificação, bem como a avaliação da performance desses modelos. Também foi possível ter uma noção do trabalho de um pesquisador, uma vez que o laboratório onde a iniciação foi realizada faz pesquisas em parceria com empresas.

Na fase atual, onde estamos tentando desenvolver métodos para identificar com mais precisão os elementos de uma página de notícia da web, o aluno está tendo a oportunidade de colaborar diretamente com uma pesquisa na área de extração de informação na web.

Finalmente, podemos afirmar com certa segurança que os resultados da classificação são satisfatórios. Na identificação de títulos, obtivemos precisão e recall de aproximadamente 90%. Em relação às datas, os resultados são um pouco inferiores, girando em torno de 70%. No entanto, esse aprimoramento do NCE se deu em detrimento de parte de sua velocidade, uma vez que o processamento de folhas de estilo é relativamente caro.

Referências

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

2- http://www.cs.waikato.ac.nz/ml/weka

3- E. Laber; C. Souza; I. Jabour; E. Amorim; E.Cardoso; R. Renteria; L. Tinoco; C. Valentim
A fast and simple method for extracting relevant content from news webpages
CIKM, pp. 1685-1688, ACM, 2009.