PROCESSAMENTO TEXTUAL EM PÁGINAS DA WEB

Aluno: Pedro Lazéra Cardoso
Orientador: Eduardo Sany Laber

Antecedentes

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], News Relevant Content Detector, desenvolvida no laboratório em que o aluno fez sua pesquisa.

Introdução

A popularização da Internet na última década, através do aumento do número de páginas e principalmente do número de usuários, fez com que a linguagem HTML evoluísse e se tornasse mais complexa. Essa evolução, marcada pela maior versatilidade de que atualmente a linguagem dispõe, pode ser evidenciada pela presença maciça de elementos de apresentação em páginas da web, como menus dinâmicos, conteúdo em flash, banners comerciais etc. Esse fato foi determinante para que a maneira como os sites são construídos se tornasse bastante heterogênea. Além disso, em função do aumento do número de usuários, os elementos de apresentação do layout ganharam mais importância. De fato, Gibson [4] estima que entre 40% e 50% de todo o conteúdo da web seja relacionado a esses elementos de apresentação de layout.

Figura 1. Exemplo de página com alta proporção de conteúdo irrelevante

Na figura 1.1, podemos ver que uma página da Folha de São Paulo tem boa parte de seu conteúdo relacionado a temas que não pertencem à notícia em destaque.

Essa gama de novos recursos tornou a internet acessível a qualquer pessoa. Por outro lado, esse fenômeno também causou um efeito indesejado: aumentou em muito a dificuldade dos mecanismos de busca operarem. Isso ocorre porque esses mecanismos, para estabelecer a lista de páginas relacionadas a um assunto, usam como base o conjunto de palavras das páginas. Parte desse conjunto, como menus e anúncios, em geral não tem relação direta com o assunto procurado. Dessa forma, as tarefas de busca e de ranking podem ficar comprometidas. *Tokens, frases* e *named-entites* pertencentes à sessão de anúncios são muitas vezes indistinguíveis daqueles presentes no corpo e no título de uma notícia, por exemplo.

Uma maneira de contornar essa dificuldade é extrair do conjunto de palavras de uma página seu conteúdo relevante. Por relevante, entendemos o texto que melhor representa o conteúdo da página. Para páginas de notícia, como exemplo, o conteúdo relevante seria o corpo, o título e a data da notíca.

O problema de extração pode ser solucionado com a renderização de uma página, de modo parecido com o que um navegador faria para um ser humano, explorando os atributos visuais obtidos a partir dela. Esse processo não é o único conhecido, mas é o mais utilizado, considerando que possibilita uma alta qualidade de extração. O grande revés dessa solução é o seu elevado custo computacional.

A necessidade de soluções com baixo custo computacional motivou a criação do NCE, uma heurística capaz de desempenhar a tarefa de extração de forma eficiente. No projeto de iniciação científica, estamos trabalhando no sentido de melhorar os resultados apresentados por essa heurística. Nas sessões a seguir, a maneira como o algoritmo funciona e os detalhes do projeto serão explicados com mais detalhes.

Detalhes sobre o NCE

O NCE é uma heurística que, a partir da análise da árvore DOM de uma página da web, desempenha a tarefa de extrair do seu conteúdo de texto da parte considerada relevante. O projeto foi focado no conteúdo de páginas de notícias e um de seus objetivos era garantir eficiência computacional.

Além disso, buscou-se tornar o código reutilizável - as únicas tags analisadas são as de link <a>, parágrafo <p> e título <title>. A utilização exagerada de tags poderia tornar o NCE obsoleto no futuro.

Não nos preocuparemos em explicar como o NCE funciona em detalhes – basta apenas entender que seu objetivo é extrair conteúdo relevante em páginas de notíciase e que isso é feito marcando numa árvore DOM os nós cujo conteúdo é relevante. Por fim, também é importante ressaltar que seu código foi escrito inicialmente em Python, mas depois migrado para C Sharp.

A primeira versão do projeto separava o conteúdo de uma página em duas partes: a relevante e a irrelevante. O objetivo mais concreto do projeto de iniciação era contribuir para a evolução do NCE, tornando-o capaz de distinguir entidades (título, data, autor, fonte da notícia) dentro do que é considerado relevante. Para isso, além das tags, a heurística passou a explorar os atributos dos nós da árvore DOM e as folhas de estilo (CCS) relacionadas a uma página.

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, entre 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 como parte do projeto de iniciação.

Seleção de atributos

O corpus 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 contrúido. 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.

Entre os atributos que selecionamos como candidatos estão:.

1- Bold - tem como valores possíveis “yes” e “no” e informa se um nó de texto está ou não em negrito.
2- FontRelativeSize – tem como valores possíveis os racionais positivos. Informa o tamanho da fonte do texto normalizado (o maior valor entre todos os nós de texto passa a ser 1).
3- Len – informa a quantidade de caracteres de um nó.
4- Unique – um nó de texto tem unique igual a x se existem x nós iguais a ele. Dois nós são iguais se têm (att1, att2, att3) iguais.
5- NumericPorcentage – informa o percentual de caracteres numéricos do texto de um nó. É usado para datas.

Construção e avaliação de modelos de classificação

Na construção do modelo de classificação, algumas medidas foram tomadas para evitar problemas de overfiting, underfiting e também para contornar a desvantagem de se trabalhar com um corpus pouco extenso, com 200 exemplos.

O RCD4 foi dividido em três partes – aqui, P1, P2 e P3. Essas três partições tinham sites de domínios diferentes para evitar um modelo de classificação viciado. Com elas, foram

realizados três experimentos. Cada experimento era realizado da seguinte forma: duas partições eram selecionadas para o treino, ou seja, para a construção do modelo que melhor classifica esse subconjunto de páginas. A partição restante era usada para o teste. A diferença entre cada experimento estava na escolha da partição cujos exemplos serviriam para o teste.

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. De fato, o procedimento que utilizamos, descrito acima, é uma espécie de cross-validation, onde garantimos que os conjuntos de treino e de teste não tivessem sites de um mesmo domínio.

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, marcou como relevante. Finalmente, para a construção de modelos de classificação e posteriormente a avaliação do desempenho desses modelos, utilizamos o software WEKA[2].

Resultados

Até a data em que esse relatório foi escrito, a heurística do NCE estava produzindo resultados satisfatórios apenas na detecção de título e de data. Assim, abaixo seguem os resultados referentes a essas duas instâncias.

Nas medições, utilizamos o *recall* e o *precision*. O *F1-score* representa a média harmônica dos resultados.

1- Recall: quantidade de palavras que o NCE selecionou corretamente / quantidade de palavras corretas.

2- Precision: quantidade de palavras que o NCE selecionou corretamente / quantidade de palavras que o NCE selecionou, corretamente ou não.

Títulos

A investigação a respeito dos títulos ficou concluída, e os resultados da extração foram bastante satisfatórios.

Tabela 1 – Extração de título

Experimento	Recall	Precision	F1-score
Exp.1 – P1 no teste	0,918	1,000	0,957
Exp.2 – P2 no teste	0,898	0,964	0,930
Exp.3 – P3 no teste	0,933	0,966	0,949
Média	**0,916**	**0,977**	**0,945**

Datas

A investigação a respeito das datas ainda não tinha sido concluída até o fechamento desse relatório, mas a heurística já estava produzindo os seguintes resultados.

Tabela 2 – Extração de data

Experimento	Recall	Precision	F1-score
Exp.1 – P1 no teste	0,486	0,927	0,638
Exp.2 – P2 no teste	0,674	0,821	0,740
Exp.3 – P3 no teste	0,663	0,846	0,743
Média	**0,608**	**0,865**	**0,707**

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.

4- Gibson, D. ; Punera, K. ; Tomkins, A. **The volume and evolution of web pag templates**. Em WWW ’05: Special interest tracks and posters of the 14th international conference on World Wide Web, páginas 830–839, New York, NY, USA, 2005. ACM.