ANÁLISE ESTRUTURAL PARA CLASSIFICAÇÃO DE PÁGINAS DA WEB

Aluno: Iam Vita Jabour
Orientador: Raúl Rentería e Eduardo Laber

Introdução

Este trabalho apresenta uma abordagem para a classificação funcional de páginas da Web denominada classificação estrutural, onde apenas informações topológicas são analisadas, criando independência à análise textual e apresentando um novo ferramental para a resolução do problema. É adotada a análise individual de páginas, ao invés da análise de sítios.

Objetivos

Classificar de acordo com a funcionalidade as páginas HTML da Web, utilizando informações obtidas a partir da estrutura desses documentos.

Metodologia

O problema de classificação de páginas da Web, também conhecido como categorização de páginas da Web, é abordado como um problema de classificação supervisionada [1], onde um conjunto de dados previamente rotulados é utilizado para o treino e teste do modelo de classificação. As classes utilizadas são formuladas a partir de escopos como funcional, onde o objetivo da página define as classes, e de assunto, onde o que a página apresenta define sua classe. Neste trabalho é adotado o escopo funcional.

Para o aprendizado e a criação do modelo de classificação são utilizadas as classes notícias, portal de notícias e outros. A classe notícias é constituída por páginas que apresentam informações de forma textual, onde o texto é o elemento principal da página. A classe portal de notícias é constituída por páginas que trazem diversos títulos ou chamadas para notícias, junto ao hiperlink para essas notícias. A classe outro é formada por páginas que não se adéquam a classe notícia ou a classe portal de notícias.

Um conjunto de páginas, denominado corpus, foi obtido da Web e rotulado. Esse é formado por trezentas páginas, sendo cem de cada classe apresentada anteriormente. Esse corpus é dividido em duas partes, onde: o conjunto de treino formado por 80% dos documentos é utilizado para a evolução e aprendizado; e o conjunto de teste, é guardado no início das análises para testar o modelo de classificação obtido a partir do conjunto de treino.

A análise topológica é realizada sobre a estrutura de um documento, por isso é utilizado DOM[1], fornecido pela W3C[2], para a criação das estruturas dos documentos estudados. Essa estrutura é uma árvore, onde seus nós são de tipos específicos, existindo 11 tipos. Os tipos *Element* e *Text* são importantes para este estudo, pois agregam grande quantidade de informação da topologia do documento. Ignorar os outros tipos não prejudica a análise, apenas a simplifica.

Observando essas estruturas os esforços são direcionados em três linhas principais. O estudo de nós busca identificar padrões estruturais, criados pelos autores dos documentos, na tentativa de encontrar semelhanças entre documentos da mesma classe. O estudo de tags procura identificar as tags que são mais utilizadas dentre as classes, assim como, as mais

[1] DOM: Document Object Model – http://www.w3.org/DOM/
[2] W3C: World Wide Consortuim – http://www.w3.org

importantes e como obter, a partir dessas informações, atributos para a classificação dos documentos. O estudo de caracteres identifica padrões no balanceamento do texto dentro da estrutura desses documentos.

Após analisar dados estatísticos para cada estudo apresentado, é possível observar que as informações são significativas para a classificação dos documentos. E a partir disso realizar experimentos para testar a qualidade de modelos de classificação utilizando os atributos obtidos a partir das observações estruturais.

A ferramenta WEKA, desenvolvida pela Universidade de Waikato, foi utilizada para a realização dos experimentos. A técnica de Support Vector Machine [2] (SVM) é escolhida para a criação do modelo de classificação e o algoritmo SMO [3] é adotado, por ser oferecido de forma nativa pela WEKA.

Os experimentos sobre o conjunto de treino foram realizados utilizando a técnica de validação cruzada com k partições [4] executando-a dez vezes. Esses apresentaram os seguintes resultados para duas (notícia x portal de notícias) e três classes (notícia x portal de notícias x outro):

Informação	Nós		Tags		Caracteres		Todos	
Classes	2 classes	3 classes	2 classes	3 classes	2 classes	3 classes	**2 classes**	**3 classes**
Acurácia	71,36%	56,12%	74,68%	58,76%	81,51%	56,87%	**83,99%**	**64,92%**
Precision	0,69	0,56	0,78	0,70	0,78	0,60	**0,88**	**0,78**
Recall	0,80	0,52	0,70	0,55	0,92	0,88	**0,81**	**0,71**
F1	0,73	0,58	0,73	0,60	0,84	0,70	**0,83**	**0,73**

Os experimentos do conjunto de teste foram realizados obtendo-se um modelo a partir do conjunto de treino e aplicando-o no conjunto de teste. Seus resultados para duas e três classes, como no experimento anterior são apresentados na tabela abaixo:

Informação	Nós		Tags		Caracteres		Todos	
Classes	2 classes	3 classes	2 classes	3 classes	2 classes	**3 classes**	**2 classes**	3 classes
Acurácia	65%	40%	67.5%	48.33%	70%	**55%**	**75%**	45%
Precision	0.63	0.37	0.73	0.53	0.67	**0.5**	**0.81**	0.48
Recall	0.75	0.5	0.55	0.45	0.8	**0.8**	**0.65**	0.55
F1	0.68	0.43	0.63	0.49	0.73	**0.62**	**0.72**	0.51

Conclusões

A classificação funcional de páginas da Web utilizando apenas elementos estruturais é possível, sendo apresentados resultados com acurácia de 75% para a classificação com duas classes (notícia x portal de notícia). A abordagem estrutural apresenta grande quantidade de informação que ainda pode ser explorada, apresentando grande potencial para classificação de outras classes.

Referências

1 - Mitchell, T.M. Machine Learning. New York: McGraw-Hill. 1997.

2 - Bores, B.; Guyon, I.; Vapnik, V. A training algorithm for optimal margin classifiers. Annual Workshop on Computational Learning Theory. 1992

3 - Platt, J. Using Sparseness and Analutic QP to Speed Training of Support Vector Machines. Advances in Neural Information Processing Systems 11. 1999

4 - Tan, P.; Steinbach, M.; Kumar, V. Introdution to Data Mining. US Ed edition: Addison Wesley, 2005. P. 146-236