



Thoran Araguez Rodrigues

**Estudo Comparativo de Estratégias de Classificação
de Páginas Web**

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós-Graduação em
Informática da Pontifícia Universidade Católica do Rio de Janeiro
como requisito parcial para a obtenção do grau de Mestre em
Informática.

Orientador: Prof. Eduardo Sany Laber

Rio de Janeiro,
Março de 2009



Thoran Araguez Rodrigues

Estudo Comparativo de Estratégias de Classificação de Páginas Web

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Eduardo Sany Laber

Orientador

Departamento de Informática – PUC-Rio

Prof. Raul Pierre Renteria

Departamento de Informática – PUC-Rio

Prof. Ruy Luiz Milidui

Departamento de Informática – PUC-Rio

Prof. José Eugenio Leal

Coordenador Setorial do Centro

Técnico Científico – PUC-Rio

Rio de Janeiro, 3 de março de 2009

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização do autor, do orientador e da universidade.

Thoran Araguez Rodrigues

Graduou-se em Engenharia de Computação pela PUC-Rio em 2005.

Ficha Catalográfica

Rodrigues, Thoran Araguez

Estudo comparativo de estratégias de classificação de páginas Web / Thoran Araguez Rodrigues ; orientador: Eduardo Sany Laber. – 2009.

82 f. : il.(color.) ; 30 cm

Dissertação (Mestrado em Informática)– Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui bibliografia

1. Informática – Teses. 2. Aprendizado de Máquina. 3. Classificação. 4. Web. 5. Blogs. 6. News Pages. I. Laber Eduardo Sany. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

A minha esposa, pela paciência e apoio.

Ao meu orientador, Professor Eduardo Sany Laber pelo auxílio, direcionamento, e cobrança nas medidas corretas para a confecção deste trabalho.

Aos meus pais, por me incentivarem a sempre terminar o que eu comecei.

Aos meus colegas de trabalho, pela ajuda quando o tempo ficou curto.

À todos os que de alguma forma me ajudaram, com recomendações e conselhos.

Resumo

Rodrigues, Thoran Araguez; Laber, Eduardo Sany. **Estudo Comparativo de Estratégias de Classificação de Páginas Web**. Rio de Janeiro, 2009. 82p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A quantidade de informações na Internet aumenta a cada dia. Embora esta proliferação aumente as chances de que o tema sendo buscado por um usuário esteja presente na rede, ela também torna encontrar a informação desejada mais difícil. A classificação automática de páginas é, portanto, uma importante ferramenta na organização de conteúdo da Web, com aplicações específicas na melhoria dos resultados retornados por máquinas de busca. Nesta dissertação foi realizado um estudo comparativo de diferentes conjuntos de atributos e métodos de classificação aplicados ao problema da classificação funcional de páginas web, com foco em 4 classes: Blogs, Blog Posts, Portais de Notícias e Notícias. Ao longo dos experimentos, foi possível constatar que a melhor abordagem para esta tarefa é a utilização de atributos tanto da estrutura quanto do texto das páginas. Foi apresentada também uma estratégia nova de construção de conjuntos de atributos de texto, que leva em consideração os diferentes estilos de escrita das classes de páginas.

Palavras-chave

Aprendizado de Máquina; Classificação; Web; Blogs; News Pages

Abstract

Rodrigues, Thoran Araguez; Laber, Eduardo Sany (Advisor). **A Comparative Study of Web Page Classification Strategies**. Rio de Janeiro, 2009. 82p. Msc. Dissertation – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The amount of information on the Internet increases every day. Even though this proliferation increases the chances that the subject being searched for by an user is on the Web, it also makes finding the desired information much harder. The automated classification of pages is, therefore, an important tool for organizing Web content, with specific applications on the improvement of results displayed by search engines. In this dissertation, a comparative study of different attribute sets and classification methods for the functional classification of web pages was made, focusing on 4 classes: Blogs, Blog Posts, News Portals and News. Throughout the experiments, it became evident the best approach for this task is to employ attributes that come both from the structure and the text of the web pages. We also presented a new strategy for extracting and building text attribute sets, that takes into account the different writing styles for each page class.

Keywords

Machine Learning, Classification, Web, Blogs, News Pages

Sumário

1. Introdução	10
2. Conceitos Gerais de Classificação de Documentos na Web	13
2.1. Páginas, Sites e Outras Terminologias	13
2.2. A Evolução de Documentos na Web	14
2.3. Rankeamento	15
2.4. Classificação na Internet	16
3. Classes Funcionais	21
3.1. Blogs	21
3.2. Blog Posts	27
3.3. Portais de Notícias	28
3.4. Notícias	30
4. Coleta de Informações, Extração de Atributos e Métodos de Classificação	33
4.1. Coleta de Informações	33
4.2. Conjuntos de Atributos e Processos de Extração	36
5. Experimentos e Resultados	41
5.1. Classificação Funcional com Atributos Estruturais	42
5.2. Classificação Funcional com Atributos de Texto (80%)	46
5.3. Classificação Funcional com Atributos de Texto (50%)	49
5.4. Classificação Funcional com Seleção de Melhores Atributos	53
5.5. Classificação Funcional com Atributos “Não-Tecnológicos”	57
5.6. Classificação Funcional com Atributos Estruturais Refinados	61
5.7. Classificação Funcional Combinando Atributos Estruturais e de Texto (80%)	65
6. Conclusões	70

6.1.	Seleção do Conjunto de Atributos	70
6.2.	Classificadores	72
6.3.	Classes de Segmentação	73
6.4.	Extração de Atributos	74
6.5.	Limitações	77
6.6.	Trabalhos Futuros	77
7.	Bibliografia	79

Lista de figuras

Figura 1 – Imagem retirada do blog portal “Mister Music Blog”	22
Figura 2 – Imagem retirada do blog portal de filmes do jornal NY Post	23
Figura 3 – Imagem do site da CNN	29
Figura 4 – Imagem do site MSN	30
Figura 5 – Imagem de uma notícia retirada do site “The Local”	31
Figura 6 – Imagem de uma notícia retirada do site do International Herald Tribune	32
Figura 7 – Árvore de decisão construída para o conjunto de atributos de texto 80%	75
Figura 8 – Ramo de árvore de decisão construída para o conjunto de atributos que combina atributos estruturados com atributos de texto	76