Pontifícia Universidade Católica
do Rio de Janeiro

Francisco Benjamim Filho

Classificação de páginas Web por aprendizagem de múltiplas categorias latentes

Tese de Doutorado

Tese apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática da PUC–Rio como requisito parcial para obtenção do título de Doutor em Informática

Orientador: Prof. Ruy Luiz Milidiú

Rio de Janeiro
Março de 2012

Pontifícia Universidade Católica
do Rio de Janeiro

Francisco Benjamim Filho

Classificação de páginas Web por aprendizagem de múltiplas categorias latentes

Tese apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do título de Doutor em Informática. Aprovada pela comissão examinadora abaixo assinada.

Prof. Ruy Luiz Milidiú
Orientador
Departamento de Informática — PUC–Rio

Prof. Marcus Vinicius Soledade Poggi de Aragão
Departamento de Informática — PUC-RIO

Prof. Raúl Pierre Rentería
Departamento de Informática — PUC-RIO

Prof. Bianca Zadrozny
UFF

Prof. Julio Cesar Duarte
IME

Prof. José Eugenio Leal
Coordenador do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 27 de Março de 2012

Francisco Benjamim Filho

Graduou-se no ano de 2001 em Engenharia da Computação no Instituto Militar de Engenharia (Rio de Janeiro). Em 2005 obteve o título de Mestre em Informática pela Pontifícia Universidade Católica do Rio de Janeiro. Trabalha desde 2001 no Centro Tecnológico do Exército (Rio de Janeiro) integrando atualmente a equipe de desenvolvimento de Radio Definido por Software.

Ficha Catalográfica

Benjamim Filho, Francisco

Classificação de páginas Web por aprendizagem de múltiplas categorias latentes / Francisco Benjamim Filho; orientador: Ruy Luiz Milidiú. — Rio de Janeiro : PUC–Rio, Departamento de Informática, 2012.

v., 76 f: il. ; 29,7 cm

1. Tese (Doutorado em Informática) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Tese. 2. WWW. 3. Algoritmo. 4. Relevância. 5. Máquinas de busca. 6. Classificação baseada em hiperlink. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

À minha esposa e ao meu filho.

Agradecimentos

À minha esposa Nathália por todo apoio, amor, carinho, dedicação e paciência, e pelo incentivo a seguir em frente, principalmente, nos momentos mais difíceis.

Ao meu filho Benício que me apresentou uma alegria ímpar e um amor incomensurável.

Aos demais familiares, pelo apoio.

Ao meu orientador, Professor Ruy Luiz Milidiú, pelas sugestões e correções a esse trabalho e pela atenção, incentivo e compreensão ao longo do curso.

Aos amigos que sempre estiveram presente, mesmo que de maneira virtual, me apoiando, incentivando.

À direção do CTEx e à chefia da Divisão de Tecnologia da Informação, por permitirem a realização do curso de doutorado em tempo parcial.

À banca examinadora pelas críticas positivas ao presente trabalho.

Agradeço à PUC-Rio por me oferecer um ambiente tão propício à minha formação.

Aos meus professores pelo exemplo de seriedade, profissionalismo e por apontarem bons caminhos a serem trilhados.

Resumo

Benjamim Filho, Francisco; Milidiú, Ruy Luiz. **Classificação de páginas Web por aprendizagem de múltiplas categorias latentes**. Rio de Janeiro, 2012. 76p. Tese de Doutorado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O crescimento explosivo e a acessibilidade generalizada da *World Wide Web* (*WWW*) levaram ao aumento da atividade de pesquisa na área da recuperação de informação para páginas *Web*. A *WWW* é um rico e imenso ambiente em que as páginas se assemelham a uma comunidade grande de elementos conectada através de hiperlinks em razão da semelhança entre o conteúdo das páginas, a popularidade da página, a autoridade sobre o assunto e assim por diante, sabendo-se que, em verdade, quando um autor de uma página a vincula à outra, está concebendo-a como importante para si. Por isso, a estrutura de hiperlink da *WWW* é conhecida por melhorar significativamente o desempenho das pesquisas para além do uso de estatísticas de distribuição simples de texto. Nesse sentido, a abordagem *Hyperlink Induced Topic Search* (*HITS*) introduz duas categorias básicas de páginas *Web*, *hubs* e autoridades, que revelam algumas informações semânticas ocultas a partir da estrutura de hiperlink. Em 2005, fizemos uma primeira extensão do *HITS*, denominada de *Extended Hyperlink Induced Topic Search* (*XHITS*), que inseriu duas novas categorias de páginas *Web*, quais sejam, novidades e portais. Na presente tese, revisamos o *XHITS*, transformando-o em uma generalização do *HITS*, ampliando o modelo de duas categorias para várias e apresentando um algoritmo eficiente de aprendizagem de máquina para calibrar o modelo proposto valendo-se de múltiplas categorias latentes. As descobertas aqui expostas indicam que a nova abordagem de aprendizagem fornece um modelo *XHITS* mais preciso. É importante registrar, por fim, que os experimentos realizados com a coleção *ClueWeb09* 25TB de páginas da *WWW*, baixadas em 2009, mostram que o *XHITS* pode melhorar significativamente a eficácia da pesquisa *Web* e produzir resultados comparáveis aos do *TREC 2009/2010 Web Track*, colocando-o na sexta posição, conforme os resultados publicados.

Palavras–chave

WWW ; Algoritmo ; Relevância ; Máquinas de busca ; Classificação baseada em hiperlink.

Abstract

Benjamim Filho, Francisco; Milidiú, Ruy Luiz (advisor). **Ranking of Web pages by learning multiple latent categories**. Rio de Janeiro, 2012. 76p. DSc. Thesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The rapid growth and generalized accessibility of the World Wide Web (WWW) have led to an increase in research in the field of the information retrieval for Web pages. The WWW is an immense and prodigious environment in which Web pages resemble a huge community of elements. These elements are connected via hyperlinks on the basis of similarity between the content of the pages, the popularity of a given page, the extent to which the information provided is authoritative in relation to a given field etc. In fact, when the author of a Web page links it to another, s/he is acknowledging the importance of the linked page to his/her information. As such the hyperlink structure of the WWW significantly improves research performance beyond the use of simple text distribution statistics. To this effect, the HITS approach introduces two basic categories of Web pages, hubs and authorities which uncover certain hidden semantic information using the hyperlink structure. In 2005, we made a first extension of HITS, called Extended Hyperlink Induced Topic Search (XHITS), which inserted two new categories of Web pages, which are novelties and portals. In this thesis, we revised the XHITS, transforming it into a generalization of HITS, broadening the model from two categories to various and presenting an efficient machine learning algorithm to calibrate the proposed model using multiple latent categories. The findings we set out here indicate that the new learning approach provides a more precise XHITS model. It is important to note, in closing, that experiments with the ClueWeb09 25TB collection of Web pages, downloaded in 2009, demonstrated that the XHITS is capable of significantly improving Web research efficiency and producing results comparable to those of the TREC 2009/2010 Web Track.

Keywords

WWW ; Algorithm ; Relevance ; search engines ; Classification based on hyperlink.

Sumário

Lista de figuras

Lista de tabelas