Francisco Benjamim Filho

XHITS: Estendendo o Algoritmo HITS para Extração de Tópicos na WWW

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE INFORMÁTICA
Programa de Pós-Graduação em Informática

Rio de Janeiro, abril de 2005

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Francisco Benjamim Filho

XHITS: Estendendo o Algoritmo *HITS* para Extração de Tópicos na *WWW*

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio.

Orientadores: Ruy Luiz Milidiú
Raúl Rentería

Rio de Janeiro, abril de 2005

Francisco Benjamim Filho

XHITS: Estendendo o Algoritmo *HITS* para Extração de Tópicos na WWW

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ruy Luiz Milidiú
Orientador
Departamento de Informática - PUC - Rio

Prof. Raúl Pierre Renteria
Departamento de Informática - PUC - Rio

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

Prof. Daniel Schwabe
Departamento de Informática - PUC - Rio

Prof. José Eugenio Leal
Coordenador(a) Setorial do Centro
Técnico Científico - PUC-Rio

Rio de Janeiro, 04 de abril de 2005

Francisco Benjamim Filho

Graduou-se em Engenharia de Computação no Instituto Militar de Engenharia.

Ficha Catalográfica

Benjamim Filho, Francisco

XHITS: estendendo o algoritmo HITS para extração de tópicos WWW / Francisco Benjamim Filho ; orientadores: Ruy Luiz Milidiu, Raúl Rentería. – Rio de Janeiro : PUC-Rio, Departamento de Informática, 2005.

76 f. ; 30 cm

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

Inclui referências bibliográficas

1. Informática – Teses. 2. Análise de hyperlinks. 2. Busca na WWW. 3. Hub. 4. Autoridade. 5. Portal. 6. Novidade. 7. HITS. 8. XHITS. I. Milidiu, Ruy Luiz. II. Rentería, Raúl. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

À minha esposa Nathália.

Agradecimentos

À minha esposa Nathália, pelo apoio incondicional e compreensão.

À família, pelo apoio.

Ao meu orientador, Prof. Ruy Luiz Milidiú, pelas correções e sugestões a esse trabalho.

Ao Prof. Raúl Rentería, pelas correções e sugestões a esse trabalho.

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

Resumo

Benjamim Filho, Francisco. ***XHITS*: Estendendo o Algoritmo HITS para Extração de Tópicos na *WWW*.** Rio de Janeiro, 2005. 076p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O ambiente baseado em *hyperlink* possui na sua topologia informações substanciais sobre o seu conteúdo. Baseado nesse tipo de ambiente, *Jon Kleingerg* desenvolveu um conjunto de algoritmos, popularmente conhecido como *HITS (Hyperlink Induced Topic Search)*, que utiliza a estrutura de *hyperlinks* na *WWW* para extrair essas informações. O foco central desses algoritmos é a classificação de tópicos de busca de caráter geral na *WWW*, através da descoberta de páginas que representam autoridade sobre tais tópicos. Para tanto, os algoritmos formulam a noção de autoridade considerando o relacionamento, decorrente da estrutura de *hyperlink*, entre o conjunto de páginas que são autoridades relevantes e o conjunto de páginas que apontam para essas, denominadas de *hubs*. *Jon Kleingerg* definiu, portanto, uma relação de interdependência entre os conjuntos anteriormente citados: uma boa autoridade será uma página apontada por bons *hubs* e um bom *hub* será uma página que aponta para boas autoridades. Neste trabalho, propomos a extensão do modelo formulado por *Jon Kleingerg*, através da inserção de novos conceitos nas relações de interdependência entre autoridades e *hubs*. Assim, formulamos um algoritmo estendido, *XHITS (Extended Hyperlink Induced Topic Search)*, que visa melhorar a classificação das autoridades do ambiente. Nessa extensão as autoridades são apontadas por bons *hubs*, às vezes apontadas por bons portais e também apontam para boas novidades. Os bons *hubs* são páginas que apontam para boas autoridades e novidades, e são apontados por bons portais. As boas novidades são páginas que são apontadas pelas boas autoridades, pelos bons *hubs* e pelos bons portais e bons portais são páginas que apontam para as boas autoridades, para bons *hubs* e para boas novidades. Adicionalmente, mostramos que o algoritmo proposto converge e também os diversos resultados experimentais que indicam a melhoria na precisão dos hiperdocumentos recuperados.

Palavras-chave

Análise de *hyperlinks*; busca na *WWW*; *hub*; autoridade; portal; novidade; *HITS*; *XHITS*.

Abstract

Benjamim Filho, Francisco. **XHITS: Extending the HITS Algorithm for distillation of broad search topic on *WWW*.** Rio de Janeiro, 2005. 076p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The network structure of a hyperlinked environment can be a rich source of information about the content of this environment. Jon Kleinberg developed a set of algorithms, called HITS (Hyperlink Induced Topic Search), for extracting information from the hyperlink structures of those environments. The aim of these algorithms is the distillation of broad search topics, through the discovery of related authoritative information sources. The notion of authority is based on the hyperlink structure relationship between a set of relevant authoritative pages and the set of hubs. Thus, hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed by many good hubs. In this work, we present the XHITS (Extended Hyperlink Induced Topic Search) algorithm, an extension of the HITS algorithm by introducing new concepts on the mutually reinforcing relationship. In XHITS, a good authority is a page that is pointed by many good hubs, some good portals and points to good novels; a good hub is a page that points to many good authorities, some good novels and is pointed by some good portals; and a good novel is a page that is pointed by good authorities, some good hubs and some good portals; a good portal is a page that points to some good authorities, some good hubs and some good novels. In addition, we show that XHITS converges and, through some experiments, the improved quality of the hyper documents retrieved.

Keywords

link analysis; ranking; Web searching; hubs; authorities; novel; portal; HITS; XHITS.

Sumário

Lista de figuras

Lista de tabelas