

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



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



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

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

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 Milidui, 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. Milidui, Ruy Luiz. II. Rentería, Raúl. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

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 CTEEx 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

1 Introdução	13
2 Modelo de Hubs e Authorities	16
2.1. Algoritmo <i>HITS</i>	16
3 Extensão do Modelo de <i>Hubs e Authorities</i>	20
3.1. O Algoritmo <i>XHITS</i>	20
3.2. Convergência do <i>XHITS</i>	23
4 Ambiente de Experimentação	29
4.1. A Ferramenta	29
4.1.1. SubGrafo da <i>WWW</i>	29
4.1.2. Algoritmo <i>HITS</i>	31
4.1.3. Algoritmo <i>XHITS</i>	33
4.1.4. Estatísticas e Gráficos	34
4.1.5. Algoritmo <i>HITS</i>	35
4.1.6. Algoritmo <i>XHITS</i>	36
4.1.7. Calibração	38
4.2. Benchmark	39
4.2.1. Consultas	39
4.2.2. Avaliação dos Especialistas	42
4.2.3. Relevância	42
5 Experimentos	43
5.1. Soluções por Histograma Média e Modal	44
5.2. Resultados Experimentais	45
5.3. Tópico Fome	52
5.3.1. Análise com a classificação do <i>Yahoo</i>	52
5.3.2. Análise com a classificação do <i>Google</i>	57
5.4. Tópico <i>Harvard</i>	60
5.4.1. Análise com a classificação do <i>Yahoo</i>	60

5.4.2. Análise com a classificação do <i>Google</i>	65
6 Conclusões e Trabalhos Futuros	69
6.1. Conclusões	69
6.2. Trabalhos Futuros	72
7 Referências Bibliográficas	74

Lista de figuras

Figura 1: Expansão de R_G .	17
Figura 2: <i>Hubs</i> e autoridades.	18
Figura 3: Módulo de construção do sub-grafo direcionado.	31
Figura 4: Classificação das páginas do grafo utilizando o algoritmo <i>HITS</i> .	32
Figura 5: Classificação das páginas do grafo utilizando o algoritmo <i>XHITS</i> .	34
Figura 6: Gráficos comuns aos dois algoritmos.	35
Figura 7: Acompanhamento da classificação pela variação das iterações.	36
Figura 8: Acompanhamento da classificação pela variação do parâmetro teta.	37
Figura 9: Calibração dos parâmetros do <i>XHITS</i> .	38
Figura 10: Gráfico de acertos por consulta nas dez primeiras páginas.	50
Figura 11: Gráfico de acertos por consulta nas vinte primeiras páginas.	50
Figura 12: Gráfico de acertos por consulta nas cinquenta primeiras páginas.	51
Figura 13: Gráfico variando o parâmetro ϕ de 0 a 1 com $\alpha = -1$, $\beta = 0$, $\gamma = 0$ e $\theta = 0$ pelo <i>XHITS</i> .	53
Figura 14: Gráfico variando o parâmetro de iteração do algoritmo entre 20 e 40 pelo <i>HITS</i> .	53
Figura 15: Número de acertos nas cinquenta primeiras páginas pelo <i>XHITS</i> .	54
Figura 16: Número de acertos nas vinte primeiras páginas pelo <i>XHITS</i> .	54
Figura 17: Número de acertos nas dez primeiras páginas pelo <i>XHITS</i> .	54
Figura 18: Número de acertos nas dez primeiras páginas pelo <i>HITS</i> .	55
Figura 19: Número de acertos nas cinquenta primeiras páginas pelo <i>HITS</i> .	55
Figura 20: Gráfico variando o parâmetro ϕ de 0 a 1 com $\alpha = -1$, $\beta = 0$, $\gamma = 0$ e $\theta = 0$ pelo <i>XHITS</i> .	57
Figura 21: Gráfico variando o parâmetro de iteração do algoritmo entre 20 e 40 pelo <i>HITS</i> .	58
Figura 22: Número de acertos nas cinquenta primeiras páginas pelo <i>XHITS</i> .	58
Figura 23: Número de acertos nas vinte primeiras páginas pelo <i>XHITS</i> .	59
Figura 24: Número de acertos nas dez primeiras páginas pelo <i>XHITS</i> .	59
Figura 25: Número de acertos nas dez primeiras páginas pelo <i>HITS</i> .	59
Figura 26: Número de acertos nas cinquenta primeiras páginas pelo <i>HITS</i> .	60
Figura 27: Gráfico variando o parâmetro θ de 0 a 1 com $\alpha = -1$, $\beta = -1$, $\gamma = 0$ e $\phi = 0$ pelo <i>XHITS</i> .	61

Figura 28: Gráfico variando o parâmetro de iteração do algoritmo entre 10 e 20 pelo <i>HITS</i> .	61
Figura 29: Número de acertos nas cinquenta primeiras páginas pelo <i>XHITS</i> .	62
Figura 30: Número de acertos nas vinte primeiras páginas pelo <i>XHITS</i> .	62
Figura 31: Número de acertos nas dez primeiras páginas pelo <i>XHITS</i> .	62
Figura 32: Número de acertos nas dez primeiras páginas pelo <i>HITS</i> .	63
Figura 33: Número de acertos nas cem primeiras páginas pelo <i>HITS</i> .	63
Figura 34: Gráfico variando o parâmetro θ de 0 a 1 com $\alpha = 0.3$, $\beta = 0$, $\gamma = 0$ e $\varphi = 1$ pelo <i>XHITS</i> .	66
Figura 35: Gráfico variando o parâmetro de iteração do algoritmo entre 10 e 20 pelo <i>HITS</i> .	66
Figura 36: Número de acertos nas cinquenta primeiras páginas pelo <i>XHITS</i> .	67
Figura 37: Número de acertos nas trinta primeiras páginas pelo <i>XHITS</i> .	67
Figura 38: Número de acertos nas dez primeiras páginas pelo <i>XHITS</i> .	67
Figura 39: Número de acertos nas dez primeiras páginas pelo <i>HITS</i> .	68
Figura 40: Número de acertos nas cem primeiras páginas pelo <i>HITS</i> .	68

Lista de tabelas

Tabela 1: Soluções por Histograma.	45
Tabela 2: Comparações dos diversos resultados obtidos com o <i>Yahoo</i> .	46
Tabela 3: Comparações dos diversos resultados obtidos com o <i>Google</i> .	47
Tabela 4: Os dez primeiros resultados do <i>XHITS</i> para o tópico Lula.	48
Tabela 5: Os dez primeiros resultados do <i>XHITS</i> para o tópico Pelé.	48
Tabela 6: Os dez primeiros resultados do <i>XHITS</i> para o tópico Rio de Janeiro.	49
Tabela 7: Média de acertos por intervalo de páginas para cada solução.	51
Tabela 8: Classificação do <i>Yahoo</i> .	52
Tabela 9: Vinte primeiras páginas classificadas com $\alpha = -1$, $\beta = 0$, $\gamma = 0$, $\varphi = 1$ e $\theta = 0$ pelo <i>XHITS</i> .	56
Tabela 10: Vinte primeiras páginas classificadas pelo <i>HITS</i> .	56
Tabela 11: Classificação do <i>Google</i>	57
Tabela 12: Classificação do <i>Yahoo</i> .	60
Tabela 13: Vinte primeiras páginas classificadas com $\alpha = 0.3$, $\beta = 0$, $\gamma = 0$, $\varphi = 1$ e $\theta = 0.6$ pelo <i>XHITS</i> .	64
Tabela 14: Vinte primeiras páginas classificadas pelo <i>HITS</i> .	64
Tabela 15: Classificação do <i>Google</i> .	65