APLICANDO TÉCNICAS DE DETECÇÃO DE SPAMDEXING BASEADAS EM LINKS

Autor: Thiago Winkler Alves *
Orientadora: Luciana Salete Buriol
Co-orientadora: Viviane Pereira Moreira

Este trabalho foi parcialmente financiado pelo projeto ALGOWEB

*** Contato:twalves@inf.ufrgs.br**

OBJETIVOS: O objetivo deste trabalho é estudar duas técnicas, já propostas na literatura, de detecção de *spamdexing* baseadas em *links* diferentes. Posteriormente, analisar e estudar as técnicas estudadas, além de avaliar e comparar os resultados de ambas. Por fim, combinar os resultados das técnicas e analisar o resultado da combinação.

APLICABILIDADE: Para garantir a qualidade de seus resultados, os motores de busca mais modernos precisam utilizar, constantemente, técnicas de detecção de *spamdexing*.

O PROJETO

Técnicas de *spamdexing* têm "assombrado" os motores de busca por mais de uma década e ainda são um problema hoje em dia. Muitas técnicas baseadas em conteúdo para detectar esses métodos já foram propostas na literatura, mas a pesquisa sobre técnicas de detecção de *spamdexing* baseadas em *links* é recente. Este trabalho tem como foco o estudo e implementação de duas dessas técnicas baseadas em *links*, medindo suas qualidades com métricas de Recuperação de Informações bem conhecidas, e comparando seus resultados.

DESENVOLVIMENTO

As técnicas de detecção de *spamdexing* estudadas foram o *Truncated PageRank* (BECCHETTI et al., 2006) e o *TrustRank* (GYONGYI; GARCIA-MOLINA; PEDERSEN, 2004):

- A ideia geral por de trás de cada uma dessas técnicas é criar um algoritmo de *ranking* que avalia as páginas em um grafo Web de acordo com algumas propriedades;
- Após, um classificador é criado, se utilizando dos resultados destes algoritmos junto com um pequeno conjunto de *hosts* pré-rotulados por humanos como sendo *spam* ou não *spam*;
- Com este classificador, o restante das páginas presentes no grafo são classificadas e, com as métricas de RI, esta classificação é então avaliada. Além disso, experimentos com a combinação de ambas as técnicas também foram realizados.

ano

BECCHETTI, L. et al. Using Rank Propagation and Probabilistic Counting for Link- Based Spam Detection. In: WORKSHOP ON WEB MINING AND WEB USAGE ANALYSIS (WEBKDD). Proceedings. . . Citeseer, 2006.

GYONGYI, Z.; GARCIA-MOLINA, H.; PEDERSEN, J. Combating Web Spam with TrustRank. In: THIRTIETH INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES. Proceedings. . . [S.l.: s.n.], 2004. p.587.

RESULTADOS

Os resultados obtidos foram, para todos os casos, muito satisfátorios, principalmente quando os resultados de ambos os algoritmos são combinados em um único classificador. Com isso, podemos concluir que, embora o problema gerado pelos métodos de *spamdexing* ainda não possa ser cem por cento eliminado, muito já pode ser feito para melhorar a resposta dos motores de busca para as consultas submetidas por usuários.

	Truncated PageRank	*TrustRank*	*Truncated PageRank + TrustRank*
Precisão	0.70	0.77	0.77
Revocação	0.64	0.70	0.73
F-Measure	0.67	0.73	0.75
Falsos Positivos	6.9%	5.3%	5.6%
Falso Negativos	36%	30%	27%

CONCLUSÕES

- A qualidade dos resultados obtidos pelos algoritmos de detecção de *spamdexing* estudados não se compara com a de filtros anti-*spam* encontrados nas soluções de e-mail modernas, mas já é bem aceitável, mesmo quando consideramos as porcentagens de Falso positivos e Falso negativos;

- Quando adicionamos mais *features* aos classificadores, os resultados melhoraram;

- Trabalhos Futuros: modificar outros algoritmos de *ranking* da mesma maneira em que o *PageRank* já foi modificado.

Av. Bento Gonçalves, 9500 - Bloco IV - Campus do Vale - Caixa Postal 15064 - 91501-970 - Porto Alegre/RS
Fone: (51) 3308 6168 - Fax: (51) 3308 7308 - www.inf.ufrgs.br