Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Projeto de Final de Curso
Projeto e Análise de Algoritmos

Proposta de Trabalho

Alberto Hideki Ueda

Orientador: Berthier Ribeiro Neto

Belo Horizonte
6 de outubro de 2014

Coletores de páginas da *Web* constituem o primeiro passo para a implementação de máquinas de busca modernas. De forma geral, um coletor - em inglês, *crawler* - é um sistema que faz requisições a servidores da *Web* de forma planejada e automática, coleta parte ou todo o conteúdo das páginas devolvidas pelas requisições e utiliza este novo conteúdo para realizar novas requisições [1]. Estima-se que hoje mais de 10% das visitas a *Websites* sejam feitas por coletores [8].

O primeiro coletor *Web* conhecido foi criado em 1993 por Matthew Gray - então graduando do MIT - e chamava-se WWWW (*World Wide Web Wanderer).* Comprovando a forte relação com a história dos sistemas de busca na *Web*, no mesmo ano foi lançada também a primeira máquina de busca conhecida, ALIWEB, criada por Martijn Koster [1]. Nesta época, um número razoável de servidores para se obter uma boa cobertura da rede girava em torno de apenas alguns milhares. Desde então, o número de *hosts* tem aumentado em alta velocidade (chegando a praticamente dobrar a cada ano, de 1993 a 1996 [4]), tornando as máquinas de busca ainda mais necessárias e, consequentemente, também os coletores de dados.

Hoje, porém, mesmo as principais máquinas de busca cobrem apenas uma parte da *Web* atual. Em 2005, foi demonstrado que o nível de cobertura das principais máquinas de buscas existentes está entre 58% e 76% da *Web* [5]. Além disso, o custo da utilização da rede também deve ser considerado. Em 2004, tal custo foi estimado em US$ 1,5 milhão para coletores de larga escala [3].

Portanto, tal problema pode ser considerado de difícil resolução, dado o tamanho da entrada necessária para uma solução exata. A ideia deste trabalho é modificar um algoritmo existente de um coletor genérico (*General Crawler*) e transformá-lo em um coletor temático (*Focused Crawler*) - que concentra-se em um único tópico de interesse - visando aumentar tanto a qualidade quanto a cobertura das páginas coletadas em relação ao algoritmo original.

Mais especificamente, este trabalho consistirá em alterações no escalonamento de longo prazo de um *General Crawler*, direcionando as requisições de *downloads* para páginas relevantes a um tópico pré-estabelecido, por meio de consultas de referência (*driving queries*) ou documentos de exemplo (como um conjunto de páginas *Web*). As páginas da *Web* consideradas para este trabalho serão tanto públicas (não são protegidas por autenticação de usuário) quanto estáticas (não são criadas dinamicamente pela entrada do usuário). Considerando as páginas da *Web* como vértices e os diversos *links* encontrados como arestas, pode-se aplicar diferentes algoritmos para a seleção *on-line* (em tempo de execução) dos *links* que o coletor irá visitar em um nova coleta, destacando-se estratégias *Breadh-first* [7], utilização de *PageRank* [2] e até

mesmo o uso de algoritmos genéticos [6].

Referências

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval: The concepts and technology behind search*. Pearson Education Limited, 2nd edition, 2011.

[2] Junghoo Cho, Hector Garcia-Molina, and Lawrence Page. Efficient crawling through url ordering. In *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, pages 161–172, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[3] Nick Craswell, Francis Crimmins, David Hawking, and Alistair Moffat. Performance and cost tradeoffs in web search. In *Proceedings of the Australasian Database Conference ADC2004*, pages 161–170, Dunedin, New Zealand, January 2004. `http://research.microsoft.com/users/nickcr/pubs/craswell_adc04.pdf`.

[4] M. Gray. Web growth, 1996.

[5] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*, WWW '05, pages 902–903, New York, NY, USA, 2005. ACM.

[6] Judy Johnson, Kostas Tsioutsiouliklis, and C. Lee Giles. Evolving strategies for focused web crawling. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 298–305, 2003.

[7] Marc Najork and Janet L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the Tenth Conference on World Wide Web*, pages 114–118, Hong Kong, May 2001. Elsevier Science.

[8] J. Nielsen. Statistics for traffic referred by search engines and navigation directories to useit. `http://www.useit.com/about/searchreferrals.html`, 2004.