An Early DiscoWeb Prototype at TREC8

Brian D. Davison, Apostolos Gerasoulis,
Konstantinos Kleisouris, Yingfang Lu, Hyun-ju Seo,
Junyu Tian, Song Wang, Wei Wang, and Baohua Wu
Department of Computer Science
Rutgers, The State University of New Jersey
Piscataway, NJ 08855 USA

Recently the notion of popularity and its generalizations have been investigated as a possible alternative approach to text only analysis to rank web pages in search engines (e.g. [Kle98, BP98, CDR$^+$98, CDDG$^+$98, BH98, HHMN99] among others). We have built a research prototype that incorporates many link analysis algorithms from the literature and also new algorithms to investigate the impact of the popularity on the ranking of the search engines [DGK$^+$99].

Our goal in the TREC8 competition was to investigate the quality of the results using the TREC data and in particular the large web track. Unfortunately we did not have the needed hardware in time to generate results for the large web track. We only participated in the Small Web Track (Text Only and Text and Link Analysis). However, our system was designed for large datasets and the quality of the TREC8 results are not representative of the system. More recently we have experimented with larger datasets and we have come to the conclusion that link analysis can significantly increase the quality of the ranking of search engines, a conclusion that is shared by many others in the literature [BP98, PBMW98, Kle98, CDR$^+$98, CDDG$^+$98, CDG$^+$99]. We will report these new results in a future publication.

References

[BH98] Krishna Bharat and Monika R. Henzinger. Improved Algorithms for Topic Distillation in Hyperlinked Environments. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 104–111, August 1998.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[CDDG$^+$98] Soumen Chakrabarti, Byron E. Dom, Prabhakar Raghavan David Gibson, Ravi Kumar, Sridhar Rajagopalan, and Andrew Tomkins. Experiments in Topic Distillation. In *ACM SIGIR Workshop on Hypertext Information Retrieval on the Web*, Melbourne, Australia, 1998.

[CDG$^+$99] Soumen Chakrabarti, Byron E. Dom, David Gibson, Jon M. Kleinberg, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Hypersearching the Web. *Scientific American*, June 1999.

[CDR+98] Soumen Chakrabarti, Byron E. Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon M. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[DGK+99] Brian D. Davison, Apostolos Gerasoulis, Konstantinos Kleisouris, Yingfang Lu, Hyun-ju Seo, Wei Wang, and Baohua Wu. DiscoWeb: Applying Link Analysis to Web Search. In *Poster proceedings of the Eighth International World Wide Web Conference*, pages 148–149, Toronto, Canada, May 1999.

[HHMN99] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. Measuring index quality using random walks on the web. In *Proceedings of the Eighth International World Wide Web Conference*, pages 213–225, Toronto, Canada, May 1999.

[Kle98] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA-98)*, pages 668–677, San Francisco, CA, January 1998. Expanded version at http://www.cs.cornell.edu/home/kleinber/.

[PBMW98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Unpublished draft (no longer available online)., 1998.