HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL

Comparing Methods for Near-Uniform URL Sampling

Name der Diplomandin / des Diplomanden	Sebastian de Castelberg Markus Kinzler	Department of Computer Science HSR
Examinator / Experte	Prof. Stefan F. Keller	CC integis HSR
Industriepartner	Prof. Dr. Monika Henzinger	Swiss Federal Institute of Technology Lausanne
Diplomausstellungs-Raum	1.262	

This diploma thesis investigated the problem of sampling URLs uniformly at random from the web. Such a method for sampling URLs can be used to estimate various properties of web pages.

For example, one could estimate:

- The fraction of web pages written in various languages
- The coverage of various search engines
- The distribution of web pages in top-level domains

In the literature there are two approaches about sampling web pages based on random walks, namely by Henzinger et al. [1] and Bar-Yossef et al [2]. These two methods crawl up to 10 million web pages in their described approaches. The path of the crawl is captured as a directed graph whereas pages represent nodes and links edges.

Both methods have only been tested individually. The goal of this project was to compare them by performing random walks and by evaluating the generated random samples (i.e. the results). If these two methods provide nearly equal samples, there is high evidence that both generate good random samples; this means one can make accurate statistics about the internet in a few days given current hardware.

Because there are no good statistics about the internet, the two methods have been tested on a known web graph, which is based on a large crawl from august 2002 [3]. This web graph contains 95 million URLs and 500 million links. The distribution of the top-level domains of this web graph is known (47.5% .com, 17.2 % .edu, 10.22% .org, 0.42% .ch). By adopting the methods to crawl this web graph, one has the possibility to compare the distribution of the top-level domains from the random samples with those from the whole web graph. If the results are nearly equal, there is high evidence that they will also give good results on the actual internet.

Up to the moment of writing of this abstract (november 2004) there is not yet any reliable data available which would allow an analysis of the quality of the sampling methods.

More information: http://versus.integis.ch

References:

- [1] Monika Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork (2000): *On Near-Uniform URL Sampling*. In: Proceedings of the 9th International World Wide Web Conference (May 2000), p. 295-308.
- [2] Ziv Bar-Yossef, Alexander Berg, Steve Chien, Jittat Fakcharoenphol, and Dror Weitz (2000): *Approximating Aggregate Queries about Web Pages via Random Walks*. In: Proceedings of the 26th International Conference on Very Large Databases, p. 535-544.
- [3] Web graph resource: http://retrieve.computing.dcu.ie/