UNIVERSITÄT DES SAARLANDES
MAX-PLANCK-INSTITUT INFORMATIK
Dr.-Ing. Ralf Schenkel
Dr.-Ing. Marc Spaniol

Web Dynamics (SS 10)
Assignment 3

Handout on: May 20, 2010

Due on: May 27, 2010

Exercise 3.1: Change rates of Web pages

Given are three web pages p_1, p_2, p_3 and 12 time points $t_1, \ldots, t_{12}$ such that the time intervals $[t_i; t_{i+1}]$ are of length 1. The web pages are accessed as follows:

Page	Accesses	Monitored updates
p_1	$t_1, t_4, t_6, t_{10}, t_{12}$	t_4, t_6, t_{10}
p_2	t_2, t_5, t_7, t_9	t_5, t_9
p_3	t_3, t_8, t_{11}	t_8

Estimate the change rates λ_i using both the simple estimator $\bar{\lambda}_i = X_i/T_i$ and the improved estimator

$$\bar{\lambda}_i = -f_i \cdot \log \frac{Y_i + 0.5}{n_i + 0.5}$$

where T_i is the length of the time interval during which p_i was monitored, n_i is the number of accesses of p_i, X_i is the number of monitored updates, Y_i is the number of times p_i did not change, and f_i is the access frequency.

What are the advantages of the second estimator over the first one?

Exercise 3.2: Parallel crawling

The Web is way too large and dynamic to crawl it with a single-threaded crawler running on a single machine. Real crawlers run on multiple, distributed servers, using multiple parallel processes and/or threads per machine. Discuss how the load can be spread in a way that maximizes coverage while minimizing unecessary work (like pages being accessed by two different processes within a short time). Think of static solutions (where everything is fixed when the crawlers start) and dynamic solutions (where load is distributed at runtime). What other optimizations could be done to improve throughput?

Exercise 3.3: Focused Crawling

For some applications, it is not necessary to crawl the complete Web, but just pages that deal with a predefined topic (like pages on canoeing, geocaching, or barbecueing). Such focused crawlers can be built by extending a standard crawler with a classifier component that categorizes a Web page into one of several predefined categories. Discuss how such a focused classifier could be implemented, what input is needed to build the classifier, and how seed documents for a topic could be found.

Exercise 3.4: Freshness

Given are three web pages p_1, p_2, p_3 with change rates $\lambda_1 = 1/2, \lambda_2 = 1/3, \lambda_3 = 1/4$ and 12 time points $1, \ldots, 12$. At each time point we can download at most one page. Give both a uniform and a proportional download schedule for the pages and the 12 time points. What is the expected freshness of each of the pages at time point 12? Which of the strategies gives better overall freshness?