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

Web Dynamics (SS 10)
Assignment 2

Handout on: May 6, 2010

Due on: May 20, 2010

Exercise 2.1: PageRank and MapReduce

Discuss the differences between PageRank and HITS. What are the advantages of the PageRank and those of HITS? Give examples when the each of the two algorithms can be used.

Exercise 2.2: Seed selection

Seed URLs are the starting points for crawling. They can also be referred to as entry-point URLs. Suggest how to select a k number of seeds, given a currently known portion of the web so that as many new high-quality pages as possible will get crawled, and as many currently crawled high-quality pages as possible will be retained.

Exercise 2.3: MapReduce

Sketch a possible computation of HITS using MapReduce. You can start with the MapReduce-based algorithm for computing PageRank discussed in the lecture. What are the main differences?