ifis
Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval and Web Search Engines
Summer Semester 2010

Prof. Dr. Wolf-Tilo Balke and Joachim Selke

Homework Assignment 6

Due to July 15, 2010
(32 points in total)

Exercise 6.1 (PageRank)

The following MATLAB command generates a 10 × 10 adjacency matrix: `triu(ones(10), 1)`. What is the PageRank vector of its corresponding network graph, for $\lambda = 0.1$, $\lambda = 0.5$, and $\lambda = 1$? (6 points)

Exercise 6.2 (Miscellaneous exam-style questions)

a) What are the differences and similarities between Boolean retrieval and coordination level matching? (2 points)

b) How is TF–IDF defined and what is its underlying rationale? (2 points)

c) What does the Probabilistic Ranking Principle say? (2 points)

d) Recall and fallout are hard to compute. Why? What can you do to resolve this issue? (2 points)

e) How does LSI work, and why is it useful for IR tasks? (2 points)

f) What is the cluster hypothesis? Is it true? (2 points)

g) What is the key idea language models are based upon? (2 points)

h) What is the main advantage and disadvantage of pseudo relevance feedback? (2 points)

i) What are support vector machines and how do they work? What is the kernel trick? (2 points)

j) What are the three most important differences between classical information retrieval and modern Web search? (2 points)

k) What are the major components of a Web search engine? (2 points)

l) How can we estimate the size of the Web? (2 points)

m) What is shingling used for in Web search? Why do we need this complicated randomized approximation algorithm instead of just computing Jaccard coefficients? (2 points)