Max-Planck-Institut für Informatik

Databases and Information Systems Group (AG 5)

Prof. Dr.-Ing. G. Weikum Dipl.-Inform. Hanglin Pan
Dipl.-Inform. Sergej Sizov
Dipl.-Inform. Stefan Siersdorfer

Selected Topics in Web Information Retrieval and Mining (WS 03/04)

Assignment 1

Handout on: Thursday, October 30, 2003

Due on: solutions will be discussed on Friday, Nov 7, 2003

Exercise 1.1:

a) What is the influence of nodes with indegree zero in the Page-Rank computation?
Can such nodes lead to anomalous behaviour?
If so, how can the Page-Rank method be modified so as to avoid this behaviour?
Can you make any statements, ideally theorems with proofs, about the ranks of nodes with indegree zero?

b) What kinds of optimizations regarding nodes with indegree zero can you add to the Page-Rank computation?

c) Give analogous consideration to nodes with outdegree zero.

d) Give analogous consideration to the HITS method.

Exercise 1.2:

Design a topic-specific variant of the HITS method.

Exercise 1.3:

Design an interactive GUI (graphical user interface - don't implement it, however) for presenting multiple result rankings for the same query to the user on one screen (e.g., topic-specific Page-Rank-based rankings for 2 or 3 different topics). What kind of interaction possibilities (e.g., clickable links) would you support with such a GUI? How can such a GUI lead the user to better search result quality? What is your final assessment of such a GUI? Is it useful at all? Is it better than traditional GUIs with one ranking (which in turn may be based on a weighted sum of multiple criteria)?

Exercise 1.4:

Assume that in a Web link graph every node and every edge has a timestamp that tells you
a) when a node (i.e., Web page) or edge (i.e., link) was created, or
b) when a node or link was last modified, or
c) when a node or link was last visited/traversed by some Web user, or
d) when a node or link was last visited/traversed by you, or
e) the most recent time since which a node or link was visited/traversed by n different Web users (with n being 100 or 1000).

How could you exploit this additional information in link-analysis-based authority ranking?
Does it help to answer queries of the kind "*What happened recently and is important?*"?

How do your answers differ when we change from a Web setting (with href hyperlinks) to a setting where "links" denote content associations? For example, in a news archive there would be a link between two articles when both mention the same person, city, historic event, etc.

Exercise 1.5:

Do link-analysis-based authority ranking methods make sense on an undirected document graph?