CS6200 Information Retrieval

Homework4: Web graph computation

Objective

Compute link graph measures for each page crawled using the adjacency matrix. While you have to use the merged team index, this assignment is individual (can compare with teammates the results)

Page Rank - crawl

Compute the PageRank of every page in your crawl (merged team index). You can use any of the methods described in class: random walks (slow), transition matrix, algebraic solution etc. List the top 500 pages by the PageRank score. You can take a look at this PageRank pseudocode (http://www.ccs.neu.edu/course/cs6200f13/proj1.html) (for basic iteration method) to get an idea

Page Rank - other graph

Get the graph linked by the in-links in file resources/wt2g_inlinks.txt.zip
Compute the PageRank of every page. List the top 500 pages by the PageRank score; also display inlink and outlink counts for each page
Explain in few sentences why some pages have a higher PageRank but a smaller inlink count. In particular for finding the explanation: pick such case pages and look at other pages that point to them.

HITS- crawl

Compute Hubs and Authority score for the pages in the crawl (merged team index)

A. Create a root set: Obtain the root set of about 1000 documents by ranking all pages using an IR function (e.g. BM25, ES Search). You will need to use your topic as your query

B. Repeat few two or three time this expansion to get a base set of about 10,000 pages:

- For each page in the set, add all pages that the page points to
- For each page in the set, obtain a set of pages that pointing to the page
  - if the size of the set is less than or equal to d, add all pages in the set to the root set
  - if the size of the set is greater than d, add an RANDOM (must be random) set of d pages from the set to the root set
  - Note: The constant d can be 200. The idea of it is trying to include more possibly strong hubs into the root set while constraining the size of the root size.

C. Compute HITS. For each web page, initialize its authority and hub scores to 1. Update hub and authority scores for each page in the base set until convergence

- Authority Score Update: Set each web page's authority score in the root set to the sum of the hub score of each web page that points to it
- Hub Score Update: Set each web pages's hub score in the base set to the sum of the authority score of each web page that it is pointing to
- After every iteration, it is necessary to normalize the hub and authority scores. Please see the lecture note for detail.

Create one file for top 500 hub webpages, and one file for top 500 authority webpages. The format for both files should be:

- [webpageurl][tab][hub/authority score]\n

EC1

Implement a Topical PageRank by designing categories appropriate for your crawl (merged team index)

EC2

Implement SALSA scoring on your crawl (merged team index) and compare with HITS

Rubric

15 points
Page Rank on wt2g_inlinks data
30 points
PageRank on crawled data
30 points
HITS on side data