TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Exercise 10

Information Retrieval

13. Web Search Basics
14. PageRank and HITS

Warm up

Exercise 13/14.1

- Are the following statements true or false? Give reasons for your answer.

a) The size of the web can easily be determined by crawling the web and counting the pages.

b) In the context of web search, information needs are the only subclass of user needs. other subsets？

c) Shingling is a technique for detecting near duplicates of web pages.

？？

d) The average out-degree of all web pages is higher than their average in-degree. 入度更大？

e) The PageRank algorithm ranks web pages by the number of occurrences of the query terms.

f) The PageRank of a web page is query-dependent.

g) The HITS-algorithm provides two scores per web page: an authority score and a hub score.

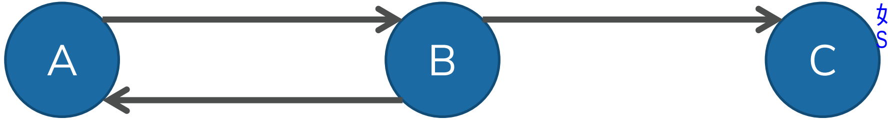
h) The HITS-scores of a web page are query-dependent.

PageRank: Transition Probability Matrices

Dresden Database Systems Group

Exercise 13/14.2

- Consider the web graph shown below



！！条件1：有出度。c用的是另一个公式

！！条件2：有路。如果没有路，算0.
如果有，算出度倒数（==0）
Sackgasse

也可以另行定义

！！验证：：相加等于1

alpha严格大于零
等于0是另一种情况：：0

- Write down the transition probability matrices for the random surfer’s walk with teleporting
- Consider the following three values of the teleport probability

$\boxed{\alpha = 0}$ $\boxed{\alpha = ½}$ $\boxed{\alpha = 1}$

很好的题目

PageRank and HITS

Exercise 13/14.3

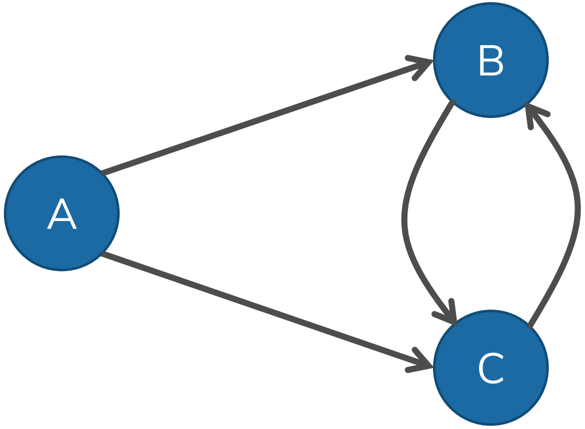
- For the web graph shown on the right, compute the following for each of the three pages A, B, and C

PageRank | authority score | hub score

- Also give the relative ordering of the three nodes for each of these scores, indicating any ties

计算收敛时候的向量，而后再排名

1.写出矩阵
2.写出爹带公式



Hints

- PageRank: Assume that at each step of the random walk, we teleport to a random page with probability 0.1, with a uniform distribution over which particular page we teleport to

- Hubs/Authorities: Normalize the hub (authority) scores so that the maximum hub (authority) score is 1 标准化？　？？ 不是很清楚

Minimum PageRank

Exercise 13/14.4

a) Show that the PageRank of every page is at least ${}^{\alpha}/_{n}$ 考虑各种情况

where α is the teleport probability and n is the total number of web pages

b) What does this imply regarding the difference in PageRank values (over the various pages) as α becomes close to 1? random i think