

Multimedia Information Extraction and Retrieval SoSe 2010 Exercise Sheet 7

Prof. Dr. Ralf Möller, Sebastian Wandelt

16.06.2010

1. A total of m white and m black balls are distributed among two urns, with each urn containing m balls. At each stage, a ball is randomly selected from each urn and the two selected balls are interchanged. Let X_n denote the number of black balls in urn 1 after the n th interchange.
 - (a) Give the transition probabilities of the Markov chain X_n , $n \geq 0$.
 - (b) Without any computations, what do you think are the limiting probabilities of this chain?

Solution:

see http://s3.amazonaws.com/cramster-resource/10482_n_23725.pdf

2. Assume that we have an index with N pages interlinked among each other. PageRank is computed for that index. The probability that a random walker jumps to an arbitrary node is $1 - q$. A spammer adds his web page `bluepill.com` to the index and wants *bluepill.com* to be ranked as high as possible. The spammer decides to use a simple strategy: add k fake websites to the index and make all of them link to *bluepill.com*. After the spammer's additions, the total number of pages in the index is $N + k + 1$. What is the PageRank value of *bluepill.com*? Assume that there are no links between the N pages and the spammer's $k + 1$ pages and that the PageRank values are normalized such that the sum of all PageRank values equals 1.

Do you think the spammer has chosen a good strategy? Could such spamming work in practice in a large-scale Internet search engine? Discuss.

Solution:

see

[http://lsirwww.epfl.ch/courses/dis/2007ws/exercises/week12/Exercise 9 solution.pdf](http://lsirwww.epfl.ch/courses/dis/2007ws/exercises/week12/Exercise%209%20solution.pdf)

3. The classical PageRank algorithm assumes that the random walker chooses the next hop uniformly randomly from the available outgoing links. However, in practice Google does not treat all links equally and uses several proprietary heuristics to determine the importance of a link. Assume that the weight (a positive real) of each link is known and the random walker chooses the next hop with the probability proportional to the outgoing link's weight. What modifications need to be made to the PageRank random walker model (slide 19 in lecture notes) and the iterative PageRank computation (slide 25) to take the link weights into account?

Solution:

see

[http://lsirwww.epfl.ch/courses/dis/2007ws/exercises/week12/Exercise 9 solution.pdf](http://lsirwww.epfl.ch/courses/dis/2007ws/exercises/week12/Exercise%209%20solution.pdf)