Text Technologies for Data Science Assignment 4

s1140740

November 13, 2014

Introduction

PageRank Algorithm

People are represented as nodes with a directed edge from a person sending an email to a person receiving the email. The weight (w) of this edge represents the number of emails the person sent. PageRanks (PR) are initialised to $1/N$ where N is the total number of people. The first 10 iterations of the algorithm are then executed. Each iteration S, the sum of PageRanks of sink nodes, is computed and people's PageRanks are updated using the PageRank formula.[1]

Note that every iteration PageRanks are computed from PageRanks from the previous iteration.

[1]
$$PR(x) = \frac{1-\lambda+\lambda S}{N} + \lambda \sum_{y \to x} \frac{w \cdot PR(y)}{out(y)}$$

HITS Algorithm

Same representation as for PageRank is used. However, hub and authority values are initialised to $1/\sqrt{N}$. Again, first 10 iterations of the algorithm are executed. Each iteration the hub value of a node is updated using authority values of nodes it is pointing to it[2] and the authority value is updated using hub values pointing to it.[3]

Both hub and authority values are then normalized[4].

[2]
$$H(x) = \sum_{y \leftarrow x} A(y)$$

[3]
$$A(x) = \sum_{y \to x} H(y)$$

[4]
$$\sum_x H(x)^2 = 1 = \sum_x A(x)^2$$