HELSINKI UNIVERSITY OF TECHNOLOGY
LABORATORY OF COMPUTER AND INFORMATION SCIENCE
NEURAL NETWORKS RESEARCH CENTRE

Stability of HITS and PageRank

Arto Klami

based on

Ng, Zheng, Jordan : Stable Algorithms for Link Analysis

Contents

HITS

Hubs y and authorities x solved iteratively

Adjacency matrix L, $L(i,j) = 1$ means that there is a link from page i to page j

$x(t+1) = (L^T L)x(t)$
$y(t+1) = (LL^T)y(t)$

Corresponds to the power method of obtaining the first eigenvector of matrices $L^T L$ and LL^T

PageRank

Simulates a random surfer: at each step the surfer follows one of the links in the current page with total probability α, or with probability $1 - \alpha$ it writes a new totally random URL

Corresponds to a Markov chain, with transition matrix $\alpha M + (1 - \alpha)U$, where M is normalized version of L, and U contains uniform probabilities for all transitions

The stationary distribution p reveals the probabilities of visiting different pages, and p_i is taken as the PageRank score of ith document

Stability

Web crawlers do not find all documents, and the collections are usually not up to date

Small perturbations to the document collection should not lead to significant changes in the page ordering ("the same search should return the same documents also tomorrow")

Simulation : rank pages $\longrightarrow$ remove some documents $\longrightarrow$ rank again

Theoretical results

HITS: Perturbation on a scale of eigengap (difference of the two largest eigenvalues) can cause large changes

PageRank: Difference in ranks is smaller than the sum of probabilities of changed documents divided by $(1-\alpha)/2 \rightarrow$ removing unimportant documents doesn't matter

Smaller α increases stability of PageRank, meaning that more restarts is better

Experimental results

Cora database, several thousand ML papers. The database is perturbed by randomly selecting a subset of 70%

HITS results change dramatically: some documents from the original top ten drop a few hundred positions, and documents with rank 2000 climb to the top ten

PageRank is very robust: the top ten documents in perturbed collections are always in the top twelve of the original collection

Could the stability of HITS be improved?

Randomized HITS

Maybe PageRank is stable because of random restarts $\rightarrow$ add restarts to the HITS algorithm

A random surfer jumps to a new page with probability $1 - \alpha$ and follows a forward (odd steps) or backward (even steps) link with probability α

Stationarity distribution of odd time steps corresponds to the authority weights, and even time steps to the hub weights

An iterative algorithm similar to the original HITS

Subspace HITS

One form of instability is caused by the eigenvectors swapping places. Hence, individual eigenvectors are not stable, but subspaces might be

Find the first k eigenvectors of $L^T L$ (or LL^T for hubs)

Authority scores: $a_j = \sum_{i=1}^{k} f(\lambda_i)(e_j^T x_i)^2$

Weighted length of the projection of the jth basis vector e_j onto the subspace spanned by the eigenvectors x_i

HITS and citation counting as special cases of function $f(\lambda)$. In the presented paper $f(\lambda) = \lambda^2$ is used

More results on stability

Randomized HITS is roughly as stable as PageRank

Subspace HITS is also clearly more stable than HITS, but not quite as stable as the other two methods (?)

Experiments were repeated for a slightly differently perturbed collection, this time for web page queries

The results in general are worse because some authoritative documents are lost in the perturbation process, but again HITS is clearly worse than the other three methods

Diversity of returned pages

The adjacency matrix can be composed of multiple components that are not connected to each other. Experiments show that this is the case with actual web queries

HITS selects only the largest of these components, as the eigenvector has zero values on other components. This leads to the highest scoring documents being similar to each other.

PageRank and Randomized HITS weight the different components and thus return authoritative documents from all components

Subspace HITS can use more components because it uses more than one eigenvector

HITS and LSI

Latent Semantic Indexing handles co-occurrence data of documents and words. The task is to characterize documents by sets of words

$L(i, j) = 1$ if document i contains word j

Leads to computation of all eigenvectors of LL^T and $L^T L$

HITS: concatenate documents and words, and set links from each document to each word that appears in it. Documents become hubs and words become authorities

Mostly a theoretical connection

Conclusions

Need for stability: the ordering of the pages should not change dramatically when the link structure is slightly perturbed

PageRank gives clearly more stable document orderings than HITS

The stability is because of random restarts; HITS with restarts (Randomized HITS) is as stable as PageRank

Stability of HITS is also increased by combining several eigenvectors (Subspace HITS)

HITS also has problems with multiple separate components