file/after/tracklang.stypackage/after/bidipackage/after/luabidi

Data Mining Homework 3 Report

Introduction

Finding the most important pages on the internet has been an important part of the success of many companies such as Google and Yahoo!. Even today, many companies spend time making sure that their site appears in search engine's results. Deciding the importance of a webiste individually presents a difficult challenge. For this reason, the field of link analysis, which analyzes the importance of different nodes in a graph, takes a look at the children and parents of each node. For a given page, the pages that link to it and the pages it links to can inform us about the relative importance of a site. The graph of the internet can be seen as a directed graph where each page, represented as a node, is connected to other pages via incoming and outgoing links.

This assignment involved implementing influential link analysis algorithms on a series of graphs, calculating the values relevant to that algorithm. The three algorithms investigated were HITS, PageRank, and SimRank. Following is a discussion of the algorithms, their implementation and results, and further discussion on the strenghts and weaknesses of each.

Algorithm Analysis

HITS

The HITS algorithm was one of the first algorithms to analyze pages. This algorithm relies on a few key observations: some pages are not authoritative in themselves, but they link to many important webpages, i.e. their outlinks are to important pages. The HITS algorithm considers two factors for each page: its **authoritativeness** and **hubness**. The former is the measure of importance from its inlinks, while the former is the importance of the site's outlinks.

These two factors rely on a mutual recursion

$$\text{auth}(p) = \sum_{c \in ch[p]} \text{hub}(c) \tag{1}$$

and

$$\text{hub}(p) = \sum_{a \in par[p]} \text{auth}(c) \tag{2}$$

where p is the page in question. We implement the algorithm in an iterative manner rather than recursive. We start of by considering a group of nodes, and each iteration we examine the authority and hubness of its children and parents, respectively. The algorithm stops when the sum of the difference between the previous hubs and authorities drops beneath a threhsold, which in this assignment is set to 1.

In our code, we initialize the authorities and hubs as an array of ones. Then, we loop through the vertices in our graph, and update the hubness and authority for each node v. Node v's authorities and hubness is the sum of parent's hubs and children's authorities, respectively. Since we update the authorities and hubs every iteration, over time this means we are taking into account nodes farther away. The values of nodes farther away in the graph are propagated to the parents and children every iteration. The algorithm stops when the difference between iteration falls below a threshold ϵ.

PageRank

The PageRank algorithm became one of the most recognized link analysis algorithms due to its use in Google. PageRank does not have the idea of hubs and authorities; rather, it defines PageRank as a function of the parent nodes' PageRank. Specifically, for a page p, its PageRank r is defined as

$$(p) = \sum_{w \in pa[p]} \frac{r(w)}{\|ch[p]\|} \tag{3}$$

We first initialize a matrix `M` by dividing each node's row by the number of children. Thus if a node i has 4 outgoing links, all the non-zero values of row i will be $\frac{1}{4}$. We then initialize $\mathbf{x}$ as the vector containing the PageRank values for each node and normalize it.

The PageRank for a page p_i is given by

$$PR(p_i) = \frac{d}{n} + (1-d) \sum_{l_{j,i} \in M} \frac{PR(P_j)}{Out(p_j)} \tag{4}$$

where $l_{i,j}$ indicates if there is a link between pages i and j. Matrix `M` already contains information about links of each node, and $\mathbf{x}$ contains the PageRank of the previous iteration. Thus Equation 4 can be interpreted as taking the product of `M` and $\mathbf{x}$. The algorithm iterates until the difference in $\mathbf{x}$ between iterations drops beneath a threhsold, which we interpret as converging.