MAS275 Probability Modelling Chapter 4: Google PageRank

Dimitrios Kiagias

School of Mathematics and Statistics, University of Sheffield

Spring Semester, 2020

Introduction

In this section we introduce the PageRank algorithm developed by Google as a method for ranking web pages in a search.

The aim of PageRank was to find a measure of "quality" for a web page, to try to ensure that the answers returned for a search query were high quality pages as well as relevant to the query.

The details of the algorithm were published in a paper by S. Brin and L. Page (1998), *The anatomy of a large-scale hypertextual web search engine*, Computer Networks and ISDN Systems 30: 107117.

Directed graph

The idea of PageRank is similar to that of a random walk on a graph in section 1.5.

Consider the Web as a graph, where the vertices represent the pages, and edges represent links between them.

In fact, because of the nature of web links, it makes sense to consider the Web as a **directed graph**, so that each edge has an inherent direction and can only be traversed in one direction.

(A pair of pages may, of course, have links going in both directions between them.)

Random walk

Given a directed graph, we can define a random walk on the graph in essentially the same way as we did in section 1.5:

We imagine a particle moving from vertex to vertex in such a way that at each step, it chooses one of the possible edges leaving its current vertex, each with equal probability, and travels along that edge.

In the context of the Web, you can think of a "random surfer":

Random walk

Given a directed graph, we can define a random walk on the graph in essentially the same way as we did in section 1.5:

We imagine a particle moving from vertex to vertex in such a way that at each step, it chooses one of the possible edges leaving its current vertex, each with equal probability, and travels along that edge.

In the context of the Web, you can think of a "random surfer": start at a web page, and click on one of the links at random, then click on one of the links from the new page at random, and keep going.

Transition probabilities

We can write the transition probabilities as

$$p_{ij} = \ell_{ij}/d_i,$$

where ℓ_{ij} is the number of links from page i to page j, and d_i is the total number of links from page i.

Complication

[There is a small complication: it is possible for there to be a page with no links, in which case our random surfer gets stuck.

One way to solve this problem is to assume that if we reach such a dead end then the surfer starts again at a random page.

In what follows, we will assume that such pages do not exist.]

Markov chain theory

This random walk as described is a Markov chain, whose state space is the set of pages of the Web.

If it is irreducible and aperiodic, then by the results of section 3.4 it will have a unique stationary distribution, and this stationary distribution will represent the limiting probabilities that our random surfer is on each page, far into the future.

We can use the values from this stationary distribution as rankings for our pages.

Example

Example

Random surfer on a mini-Web

Damping

A modification to the random surfer Markov chain of the previous section is to say that, from time to time, the random surfer gives up following links and starts again from a page chosen totally at random.

We assume that this happens at each time step with probability $1 - d$; with probability d the surfer behaves as in the previous section.

[In the paper by Brin and Page referred to above, they say that d was usually set to 0.85.]

Transition probabilities

The transition probabilities are now

$$p_{ij} = d\left(\frac{\ell_{ij}}{d_i}\right) + (1-d)\frac{1}{N},$$

Transition probabilities

The transition probabilities are now

$$p_{ij} = d\left(\frac{\ell_{ij}}{d_i}\right) + (1-d)\frac{1}{N},$$

where N is the total number of pages in the system.

Markov chain theory

This modification ensures that the Markov chain is irreducible and aperiodic.

So by the results of section 3.4 it will have a unique stationary distribution $\boldsymbol{\pi}$.

The PageRank of page i is then defined by its probability under the stationary distribution, π_i.

Example

Example

PageRank for a mini-Web