Journal of Computational and Applied Mathematics 277 (2015) 192–201

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

The eigenvectors corresponding to the second eigenvalue of the Google matrix and their relation to link spamming

Alex Sangers, Martin B. van Gijzen [*]

Delft University of Technology, Delft Institute of Applied Mathematics, Mekelweg 4, 2628 CD, The Netherlands

ARTICLE INFO

Article history:
Received 10 January 2014
Received in revised form 26 July 2014

Keywords:
Google PageRank
Link spamming
Second eigenvector
Markov chains
Irreducible closed subsets

ABSTRACT

Google uses the PageRank algorithm to determine the relative importance of a website. Link spamming is the name for putting links between websites with no other purpose than to increase the PageRank value of a website. To give a fair result to a search query it is important to detect whether a website is link spammed so that it can be filtered out of the search result.

While the dominant eigenvector of the Google matrix determines the PageRank value, the second eigenvector can be used to detect a certain type of link spamming. We will describe an efficient algorithm for computing a complete set of independent eigenvectors for the second eigenvalue, and explain how this algorithm can be used to detect link spamming. We will illustrate the performance of the algorithm on web crawls of millions of pages.

1. Introduction

Google's PageRank algorithm aims to return the best ranking of websites when searching on the web. The PageRank model assumes that a web surfer randomly follows one of the outgoing hyperlinks at a given website with a probability p or jump to a random website with probability $1-p$. Mathematically this can be modeled by a Markov chain. The PageRank of a website is the probability to be on this website in the stationary distribution of the Markov chain. This stationary distribution is given by the first eigenvector of the transition matrix of the Markov chain.

According to Haveliwala and Kamvar [1] the eigenvectors for the second eigenvalue are also of importance: they can be used to detect link spam. Link spam is the name for putting links between web pages with no other purpose than to increase the PageRank of a website. Specifically, the conclusions of [1] state that "The eigenvectors corresponding to the second eigenvalue $\lambda_2 = p$ are an artifact of certain structures in the web graph. In particular, each pair of leaf nodes in the SCC[1] graph for the chain $\mathbf{P}$ corresponds to an eigenvector of $\mathbf{A}$ with eigenvalue p. These leave nodes in the SCC are those subgraphs in the web link graph which have incoming edges, but have no edges to other components. Link spammers often generate such structures in attempts to hoard rank. Analysis of the nonprincipal eigenvectors of $\mathbf{A}$ may lead to strategies for combating link spam".

In this paper we will explain this remark. We will review the theory about the second eigenvalue of the Google Matrix that is described in [1,2] and extend it with results for the corresponding eigenvectors. We will use our findings to propose an efficient algorithm to detect these structures in the web that may indicate link spamming. We will illustrate the performance of the algorithm on web crawls containing several millions of pages.

[*] Corresponding author. Tel.: +31 152782519.
E-mail addresses: A.Sangers@student.tudelft.nl (A. Sangers), M.B.vanGijzen@tudelft.nl (M.B. van Gijzen).

[1] Strongly Connected Component.

http://dx.doi.org/10.1016/j.cam.2014.09.014
0377-0427/© 2014 Elsevier B.V. All rights reserved.

The structure of this paper is as follows. Section 2 explains the structure of the Google Matrix and gives different methods for computing the PageRank. Section 3 discusses the relation between irreducible closed subsets in a graph and link spamming. Section 4 gives the relevant theory for the second eigenvalue and the corresponding eigenvectors of the Google Matrix. It also explains how the second eigenvectors are related to the irreducible closed subsets. Section 5 describes two algorithms for computing the second eigenvectors. Section 6 compares the performance of the algorithms on web crawls of several millions of pages. Section 7 summarizes our findings and makes some concluding remarks.

Remarks on notation and terminology: The terms 'web sites', 'web pages' and 'nodes' as well as the terms 'hyperlinks' and 'web links' are used interchangeably.
The ith eigenvector is written as $\mathbf{x^{(i)}}$ and the jth element of vector $\mathbf{x}$ is written as x_j. A submatrix of matrix $\mathbf{A}$ will be denoted by $\mathbf{A_{ij}}$ and an element of $\mathbf{A}$ by a_{ij}.

2. The Google matrix

We introduce W, a set of the web pages, that are connected to each other by hyperlinks, i.e., incoming and outgoing links between web pages. The mathematical representation of W is a directed graph, in which a directed link between nodes of the graph represents an incoming or outgoing link between web pages.

Let n be the number of websites. Further, let $\mathbf{G}$ be the n-by-n connectivity matrix with $g_{ij} = 1$ if there is an outgoing hyperlink from page j to i and $g_{ij} = 0$ otherwise. $\mathbf{G}$ is the matrix representation of W. The number of websites n is extremely large, hundreds of millions, while every website only contains a few outgoing links. The matrix $\mathbf{G}$ is therefore large and sparse.

We denote by c_j the column sums of $\mathbf{G}$, that is $c_j = \sum_i g_{ij}$. Note that c_j is the number of outgoing hyperlinks of website j. We will also call this the out-degree of page j.

Surfing the web can be modeled as a Markov process, where one state transitions into another state by following hyperlinks. In order to model this process we introduce the row-stochastic matrix $\mathbf{P}$. The entries p_{ji} of $\mathbf{P}$ are given by

$$p_{ji} = \begin{cases} g_{ij}/c_j & \text{if } c_j \neq 0, \\ 1/n & \text{if } c_j = 0. \end{cases} \tag{2.1}$$

Note that $\mathbf{P^T}$ is the column-stochastic transition probability matrix of the Markov process. Nodes without outgoing hyperlinks are called *dangling nodes*. From (2.1) follows that from a dangling node all pages can be reached with equal probability. Following [3], we assume that self-referencing nodes, i.e., $g_{ii} = 1$ for node i, are not allowed.

The above Markov process does not capture the possibility that a web surfer jumps to another page without following an outlink. To include this behavior, called teleportation, Google's PageRank model assumes that an outlink is followed with probability p and a jump to a random page is made with probability $1 - p$. Typically, p is chosen between 0.85 and 0.99.

Let $\mathbf{A}$ be the n-by-n column-stochastic transition matrix of this Markov process that includes teleportation. The elements a_{ij} of this matrix are given by

$$a_{ij} = \begin{cases} pg_{ij}/c_j + (1-p)/n & \text{if } c_j \neq 0. \\ 1/n & \text{if } c_j = 0. \end{cases} \tag{2.2}$$

In matrix notation this can be written as

$$\mathbf{A} = p\mathbf{P^T} + \frac{(1-p)}{n}\mathbf{ee^T}, \tag{2.3}$$

with $\mathbf{e}$ the n-vector of all ones. Also, recognize that if page j is a dangling node then each page has a probability $1/n$ to be chosen. Thus, if column $\mathbf{a_j} = \mathbf{e}/n$ then page j is a dangling node.

By introducing the diagonal matrix $\mathbf{D}$, of which the main diagonal elements d_{jj} are defined by

$$d_{jj} = \begin{cases} 1/c_j & \text{if } c_j \neq 0 \\ 0 & \text{if } c_j = 0, \end{cases} \tag{2.4}$$

and by defining the vector $\mathbf{z}$ with coefficients z_j given by

$$z_j = \begin{cases} (1-p)/n & \text{if } c_j \neq 0 \\ 1/n & \text{if } c_j = 0, \end{cases} \tag{2.5}$$

the matrix $\mathbf{A}$ can also be written as

$$\mathbf{A} = p\mathbf{GD} + \mathbf{ez^T}. \tag{2.6}$$

The matrix $\mathbf{ez^T}$ accounts for teleportation. Note that as a consequence of this teleportation matrix, $\mathbf{A}$ is positive, meaning that every entry is positive, and is irreducible.

The PageRank is determined as the eigenvector of the dominant eigenvalue of the following system:

$$\mathbf{Ax^{(1)}} = \lambda_1\mathbf{x^{(1)}}. \tag{2.7}$$

Intuitively, when recalling the random web surfer from Section 1, the eigenvector $\mathbf{x^{(1)}}$ is the distribution of the visiting frequency for each node. The more often the surfer passes node j, the higher its PageRank will be.