Using Polynomial Chaos to Compute the Influence of Multiple Random Surfers in the PageRank Model

Paul G. Constantine and David F. Gleich

Stanford University
Institute for Computational and Mathematical Engineering
{paul.constantine,dgleich}@stanford.edu

Abstract. The PageRank equation computes the importance of pages in a web graph relative to a single random surfer with a constant teleportation coefficient. To be globally relevant, the teleportation coefficient should account for the influence of all users. Therefore, we correct the PageRank formulation by modeling the teleportation coefficient as a random variable distributed according to user behavior. With this correction, the PageRank values themselves become random. We present two methods to quantify the uncertainty in the random PageRank: a Monte Carlo sampling algorithm and an algorithm based the truncated polynomial chaos expansion of the random quantities. With each of these methods, we compute the expectation and standard deviation of the PageRanks. Our statistical analysis shows that the standard deviation of the PageRanks are uncorrelated with the PageRank vector.

1 Introduction

In its purest form, the PageRank model ignores the text underlying pages on the web and creates an irreducible, aperiod Markov chain model for a hypothetical random surfer on the link structure of the web [1]. Each entry of the stationary distribution measures the global importance of a page.

The PageRank model, however, is not unique. A PageRank value depends upon a parameter α which controls how the putative random surfer "teleports" around the web. Upon visiting a website, the random surfer chooses an outlink uniformly at random with probability α and chooses a page according to a prior distribution with probability $1 - \alpha$. This paper focuses on the modeling assumptions for the value of α and suggests a new model for PageRank that fixes a modeling error in the original PageRank formulation.

To continue our discussion, we must define the PageRank model and establish some notation. Let $\mathbf{W}$ be an adjacency matrix for a web graph, $w_{i,j} = 1$ when node i links to node j. We set $\mathbf{P}$ to be a fully row-stochastic random walk transition matrix on $\mathbf{W}$. The matrix $\mathbf{P}$ has dangling nodes corrected in an arbitrary way (for example, see [2,3]) such that $\mathbf{Pe} = \mathbf{e}$ where $\mathbf{e}$ is the vector of all ones. Let $1 - \alpha$ be the teleportation probability and $\mathbf{v}$ be the personalization

A. Bonato and F.R.K. Chung (Eds.): WAW 2007, LNCS 4863, pp. 82–95, 2007.
© Springer-Verlag Berlin Heidelberg 2007

distribution. The PageRank model requires that $0 \leq \alpha < 1$, $v_i \geq 0$, and $\mathbf{v}^T\mathbf{e} = 1$. With these definitions, the PageRank vector $\mathbf{x}(\alpha)$ is the unique eigenvector with $||\mathbf{x}(\alpha)||_1 = 1$ satisfying

$$\left[\alpha\mathbf{P}^T + (1-\alpha)\mathbf{v}\mathbf{e}^T\right]\mathbf{x}(\alpha) = \mathbf{x}(\alpha) \tag{1}$$

or equivalently [4,5] the solution of the linear system

$$\left(\mathbf{I} - \alpha\mathbf{P}^T\right)\mathbf{x}(\alpha) = (1-\alpha)\mathbf{v}. \tag{2}$$

The key error in the PageRank model is that it only accounts for a single surfer because it only permits a single value of α. The choice of α is quite mysterious. Most researchers take $\alpha = 0.85$ [6]. Recently, Avrachenkov et al. suggested choosing $\alpha = 1/2$ [7]. Their suggestion follows from graph theoretic properties of the PageRank solution vector as a function of α. If we believe the PageRank random surfer model, then α should be estimated from Internet usage logs, so that $\alpha = \mathrm{E}[A]$ where A is a random variable representing the teleportation parameter for each user. We are not aware of any studies that attempt to determine α using this methodology.

However, assuming $\alpha = \mathrm{E}[A]$ does not yield the "correct" PageRank vector This fact follows because in general $\mathrm{E}[\mathbf{x}(A)] \neq \mathbf{x}(\mathrm{E}[A])$. Intuitively, this issue arises because the PageRank model consolidates everyone into a single user. Appendix A demonstrates a formal counterexample. A more realistic model would consider that each user should have a small contribution to the final PageRank values.

To reiterate, computing $\mathbf{x}(\mathrm{E}[A])$ does not yield a PageRank vector that expresses all of the users. Instead, we propose using $\mathrm{E}[\mathbf{x}(A)]$ as a new PageRank vector that accurately models the underlying user population.

While our model for PageRank using a random parameter better represents the reality of random surfers, we would not expect the rankings generated by the model to be qualitatively different from those generated by the approximation of using $\alpha = \mathrm{E}[A]$. We expect $\mathrm{E}[\mathbf{x}(A)] \approx \mathbf{x}(\mathrm{E}[A])$ for "reasonable" distributions of A. Our results confirm this expectation, which justifies use of the PageRank vector as a global ranking for all users.

However, by modeling each component of the PageRank vector as a random variable, we gain a distinct advantage when quantifying the importance of a page. Namely, we can compute the *standard deviation* of each PageRank value with respect to the distribution of A. The standard deviation is a key tool in uncertainty quantification and allows us to examine the pages *most sensitive* to changes in PageRank based on the underlying distribution of A. In the results section, we employ the standard deviation of the PageRank vector to generate rankings that are uncorrelated with the original PageRank vector. Uncorrelated vectors are important because they provide additional useful input to a machine learning framework for generating a web search ranking function.