Stochastic Models for Web Ranking

Y. Volkovich
University of Twente
Enschede, The Netherlands
y.volkovich@ewi.utwente.nl

D. Donato
Yahoo! Research
Barcelona Catalunya, Spain
debora@yahoo-inc.com

N. Litvak
University of Twente
Enschede, The Netherlands
n.litvak@ewi.utwente.nl

Web search engines need to deal with hundreds and thousands of pages which are relevant to a user's query. Listing them in the right order is an important and non-trivial task. Thus Google introduced *PageRank* [1] as a popularity measure for Web pages. Besides its primary application in search engines, PageRank also became a major method for evaluating importance of nodes in different informational networks and database systems.

The definition of PageRank is as follows

$$PR(i) = c\sum_{j\to i}\frac{1}{d_j}PR(j) + \frac{c}{n}\sum_{j\in\mathcal{D}}PR(j) + 1 - c, \quad (1)$$

where $i = 1,\dots n$, $PR(i)$ is the PageRank of page i, d_j is the number of outgoing links of page j, the sum is taken over all pages j that link to page i, $\mathcal{D}$ is the set of pages without outgoing links (*dangling nodes*), n is the number of pages in the Web, and c is the damping factor, which is a constant between 0 and 1.

Most experimental studies of the Web agree that *in-degree*, the number of incoming links of a page, and PageRank follow similar power laws with exponent $\alpha = 1.1$. It is clear from the definition (1) that the PageRank of a page depends on the popularity and the number of pages that link to it. Thus it could be expected that the distribution of PageRank should be related to the distribution of in-degree. It is also clear that PageRank is a *global* characteristic of the Web, which should depend on *out-degrees*, correlations, and other characteristics of the underlying graph. We study the influence of in-degrees, out-degrees and dangling nodes on the PageRank distribution [3, 4]. We model the relation between these variables through a stochastic equation inspired by the definition of PageRank (1). To this end, we view the PageRank of a random page as a random variable R with $\mathbb{E}(R) = 1$. We formally describe the concept of power law in terms of regular varying random variables. Thus, we take a non-negative, integer and regularly varying random variable N for the in-degree of a random page. We consider a random variable D (*effective out-degree*), which represents the out-degree of a page that links to a particular randomly chosen page. We note that D is not the same random variable as the out-degree of a random page. Further, we assume that the fraction of the total PageRank mass concentrated in dangling nodes, equals the fraction of dangling nodes p_0. Then the PageRank R is a solution of the following stochastic equation:

$$R \stackrel{d}{=} c\sum_{j=1}^{N}\frac{1}{D_j}R_j + [1 - c(1 - p_0)].$$

Here N, the R_j's and D_j's are independent; the R_j's are distributed as R, the D_j's are distributed as D. As before, $c \in (0, 1)$ is the damping factor.

We use recent results on regular variation [2] to obtain PageRank asymptotics. To this end, we provide a recurrent stochastic model for the power iteration algorithm commonly used in PageRank computations. We start with initial distribution $R^{(0)}$, satisfying $\mathbb{E}(R^{(0)}) = 1$, and for every $k \geq 1$, we define the result of the kth iteration through the distributional identity

$$R^{(k)} \stackrel{d}{=} c\sum_{j=1}^{N}\frac{1}{D_j}R_j^{(k-1)} + [1 - c(1 - p_0)],$$

where N, $R_j^{(k-1)}$ and D_j, $j \geq 1$, are independent.

Then we obtain the PageRank asymptotics after each iteration.

THEOREM 1. *If* $\mathbb{P}\left(R^{(0)} > x\right) = o(\mathbb{P}(N > x))$, *then for all* $k \geq 1$,

$$\mathbb{P}(R^{(k)} > x) \sim C_k\mathbb{P}(N > x) \text{ as } x \to \infty,$$

where $C_k = \left(\frac{c(1-p_0)}{d}\right)^{\alpha}\sum_{j=0}^{k-1}c^{j\alpha}b^j$, *and* $b = d\mathbb{E}\left(1/D^{\alpha}\right)$.

Thus we clearly show that the power law of in-degree remains a major factor shaping the PageRank distribution. The difference between the power laws is in the multiplicative constant C_k, which depends mainly on the fraction of dangling nodes, the average in-degree, the power law exponent, and the damping factor. As we can quantify from Theorem 1, the out-degree distribution has a minor effect on the PageRank tail behavior.

Our theoretical predictions also show a good agreement with experimental data on the different Web samples [4].

1. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks*, 33:107–117, 1998.

[2] A. H. Jessen and T. Mikosch. Regularly varying functions. *Publications de l'institut mathematique, Nouvelle série*, 79(93), 2006.

[3] N. Litvak, W. R. W. Scheinhardt, and Y. Volkovich. In-degree and PageRank of Web pages: Why do they follow similar power laws? to appear in *Internet Math.*, 2007.

[4] Y. Volkovich, N. Litvak, and D. Donato. Determining factors behind the PageRank log-log plot. Technical Report 1823, University of Twente, Enschede, 2007.