TotalRank: Ranking Without Damping

Paolo Boldi
DSI, Università degli Studi di Milano
boldi@acm.org

ABSTRACT

PageRank is defined as the stationary state of a Markov chain obtained by perturbing the transition matrix of a web graph with a damping factor α that spreads part of the rank. The choice of α is eminently empirical, but most applications use $\alpha = 0.85$; nonetheless, the selection of α is critical, and some believe that link farms may use this choice adversarially. Recent results [1] prove that the PageRank of a page is a rational function of α, and that this function can be approximated quite efficiently: this fact can be used to define a new form of ranking, TotalRank, that averages PageRanks over all possible α's. We show how this rank can be computed efficiently, and provide some preliminary experimental results on its quality and comparisons with PageRank.

Categories and Subject Descriptors: G.2 [Discrete Mathematics]: Graph Theory; G.3 [Probability and Statistics].

General Terms: Algorithms, Experimentation, Measurement.

Keywords: PageRank, link farms, ranking, Kendall's τ.

1. INTRODUCTION AND MOTIVATIONS

PageRank [5] is one of the most important ranking techniques used in today's search engines: it is simple, robust, reliable and it can be computed in a quite efficient manner. PageRank is defined formally as the stationary distribution of a stochastic process whose states are the nodes of a web graph (a.k.a. web pages). The process itself is obtained by combining (a row-normalised version of) the adjacency matrix of the web graph with a trivial uniform process that is needed to make the combination irreducible and aperiodic, so that the stationary distribution is well defined. The combination depends on a *damping factor* $\alpha \in [0, 1)$.

Let $\boldsymbol{r}(\alpha)$ denote the PageRank vector for a given α; $\boldsymbol{r}(0)$ is actually a uniform vector[1], whereas $\lim_{\alpha\to 1} \boldsymbol{r}(\alpha)$ tends to concentrate all the rank in few pages, sometimes called "rank sinks".

In [5], Brin and Page suggested using $\boldsymbol{r}(0.85)$, and indeed this choice has remained by far the most common; there are empirical evidences that this value gives good ranks, and some *a posteriori* reasons for choosing a damping factor around .80. Nonetheless, some authors recently observed that studying how the PageRank of a given page changes with α can be used to detect link-spam [6]; in principle, this observation may be exploited to use a specific value of α adversarially to build link farms.

[1]For sake of simplicity, in this abstract we assume that the preference vector is uniform; all results carry over to the more general case.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

A way to avoid this danger would be averaging the PageRank value over all possible α's: this new form of ranking, that we call *TotalRank*, is made possible by recent results [1] that show how PageRank can be approximated *as a function*. This abstract introduces TotalRank and presents some preliminary results about its quality and some comparisons with PageRank.

2. DEFINITION AND ALGORITHM

In the following, let $\boldsymbol{r}(\alpha)$ denote the PageRank vector as a function of α. Recall that $\boldsymbol{r}(\alpha)$ is a rational vector function of α with no singularities in $[0, 1)$, so the following definition makes sense:

DEFINITION 1. *The TotalRank vector $\boldsymbol{T}$ is defined as follows: $\boldsymbol{T} = \int_0^1 \boldsymbol{r}(\alpha)d\alpha$ (where the integral is interpreted componentwise).*

In other words, the TotalRank of a page is the area behind the PageRank curve for that page, as α ranges on $[0, 1)$. To explain how TotalRank can be efficiently computed, recall that the simplest algorithm to approximate PageRank is based on the Power Method [5] (and, indeed, most known algorithms to compute PageRank are just variants of the Power Method); let $\boldsymbol{R}_k$ (for $k = 0, 1, \ldots$) be the k-th approximation computed by the Power Method for some (fixed but arbitrary) value α_0 of α, and set $\boldsymbol{R}_{-1} = \boldsymbol{0}$ by convention.

THEOREM 1 ([1]). *The Maclaurin expansion of PageRank is $\boldsymbol{r}(\alpha) = \sum_{k=0}^{\infty} \boldsymbol{c}_k\alpha^k$ where $\boldsymbol{c}_k = (\boldsymbol{R}_k - \boldsymbol{R}_{k-1})/\alpha_0^k$ for all $k = 0, 1, \ldots$.*

As an easy consequence, we have:

THEOREM 2. *With the same notation as above,*

$$\boldsymbol{T} = \sum_{k=0}^{\infty} (\boldsymbol{R}_k - \boldsymbol{R}_{k-1})/((k+1)\alpha_0^k).$$

This result allows one to compute TotalRank using essentially the same classical Power Method algorithm commonly employed for PageRank, with an extra vector to accumulate total ranks; an implementation is distributed under the Gnu Public License at the website `http://law.dsi.unimi.it`.

3. COMPARISONS WITH PAGERANK

In this section, we want to provide quantitative comparisons between TotalRank and PageRank based on experimental results. Experiments have been performed on three different datasets: a snapshot of the Italian web `.it` (41 Mpages), a partial snapshot of the British web `.uk` (18 Mpages) and the public 2001 crawl performed by the WebBase crawler (118 Mpages); all data are publicly available from `http://law.dsi.unimi.it`: graphs are

compressed in the WebGraph format [2], and the first two datasets where gathered using UbiCrawler.

Comparing quantitatively different ranking techniques is a difficult task, and its results are open to interpretations: establishing that two rankings are different is not sufficient *per se* to determine which one is better. We are going to provide some measures that should convince the reader that TotalRank and PageRank provide quite different rankings; then, we show some hints that TotalRank gives results of good quality, and probably better than those obtained with *standard PageRank* (i.e., PageRank with $\alpha = 0.85$).

Comparison using Kendall's τ. As a first comparative evaluation, we computed PageRank for different values of α and compared the resulting ranks with the ones obtained by TotalRank on the same graph; the comparison was performed using Kendall's τ [4], a non-parametric correlation index that measures similarity between two rankings. Figure 1 shows the results obtained for our datasets: as the reader can see, TotalRank is maximally similar with PageRank when α is about 0.7, but even in that case τ never exceeds 0.97.

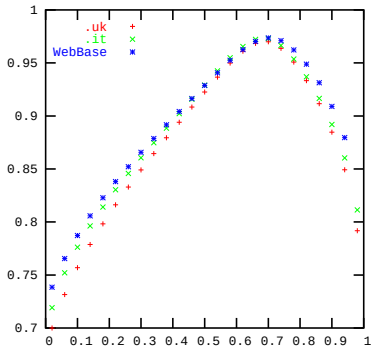

Figure 1: Kendall's τ comparison between TotalRank and PageRank, as α ranges in $[0, 1)$.

Comparing the top-k fraction. One objection to the usage of Kendall's τ as a measure of (dis)similarity is that it weights in the same way all nodes; alternative proposals advocate limiting the comparison to the "most important" pages. This idea can be implemented in different ways [3]; one possible measure (called *intersection metric*) is computed as follows: let A_t and B_t be the set of the top-t pages according to the rankings we want to compare, and let $\delta'(t) = |A_t \triangle B_t|/(2t)$ where $\triangle$ denotes the symmetric set difference; let also $\delta(k)$ be the average of all $\delta'(t)$ when $t \leq k$. Note that $\delta(k) = 0$ if the top-k lists coincide, whereas $\delta(k) = 1$ if they are completely disjoint. Figure 2 shows the results obtained comparing TotalRank with standard PageRank: as you can see, the dissimilarity tends to increase up to 20% in the first few thousandths of pages, and then it goes down to zero, as it should.

Number of homepages. Table 1 presents the number of homepages (URLs whose file portion is either empty or `index|home.*`) that are ranked among the top-k by TotalRank and standard PageRank; the reader may observe that TotalRank finds in all cases a larger number of homepages. Since usually homepages are more important than internal pages, these data indicate that the rankings obtained have high quality.

Top movers. Table 2 shows the 10 pages that moved further up or down in TotalRank order with respect to standard PageRank order. Observe, for example, that `http://tukids.tucows.com/` is advanced to position 99 (it was ranked 281th according to PageRank), as well as the homepage of the University of Melbourne (advanced

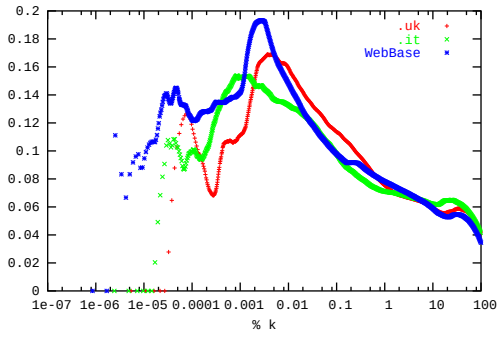

Figure 2: Comparing top-k lists of TotalRank and standard PageRank using intersection metric.

	.uk	.it	WebBase
$k = 100$	40/34 (+17.6%)	62/55 (+12.7%)	53/49 (+8.2%)
$k = 1000$	395/375 (+5.3%)	446/414 (+7.7%)	411/342 (+20.2%)
$k = 10000$	2252/2128 (+5.8%)	2361/2250 (+4.9%)	2901/2639 (+9.9%)

Table 1: Home pages found by TotalRank/standard PageRank among the top-k.

by TotalRank from 181th to 84th). On the contrary, the FAQ page `http://www.worldwidemart.com/scripts/faq/` is demoted from position 90 to 132.

URL	(TR)	(PR)	var.
http://www.worldwidemart.com/scripts/faq/	132	90	↓ 42
http://www.blakeschool.org/...	131	99	↓ 32
http://www.acme.com/	83	54	↓ 29
http://www.gendesigner.com/index.html	87	59	↓ 28
http://www.perl.com/pub	109	83	↓ 26
http://netwinsite.com/	72	139	↑ 67
http://www.scripps.com/	94	162	↑ 68
http://www.unimelb.edu.au/...	84	181	↑ 97
http://www.bef.net/161605.htm	48	147	↑ 99
http://tukids.tucows.com/	99	281	↑ 182

Table 2: Pages (... indicates a long path) with largest position change among the top 100: (TR) position according to TotalRank; (PR) position according to standard PageRank.

4. REFERENCES

[1] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. PageRank as a function of the damping factor. In *Proceedings of the Fourteenth International World–Wide Web Conference*, 2005. To appear.

[2] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. In *Proc. of the Thirteenth International World Wide Web Conference*, pages 595–601, Manhattan, USA, 2004. ACM Press.

[3] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 28–36. Society for Industrial and Applied Mathematics, 2003.

[4] Maurice G. Kendall. *Rank Correlation Methods*. Hafner Publishing Co., New York, 1955.

[5] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, Stanford University, Stanford, CA, USA, 1998.

[6] Hui Zhang, Ashish Goel, Ramesh Govindan, Kahn Mason, and Benjamin Van Roy. Making eigenvector-based reputation systems robust to collision. In Stefano Leonardi, editor, *Proceedings WAW 2004*, number 3243 in LNCS, pages 92–104. Springer-Verlag, 2004.