The T-PageRank: A Model of Self-Validating Effects of Web Surfing

Marianne Akian[1], Stéphane Gaubert[1], and Laure Ninove[2]

[1] INRIA, Domaine du Voluceau, BP105, 78153 Le Chesnay Cédex, France.
{Marianne.Akian,Stephane.Gaubert}@inria.fr

[2] Department of Mathematical Engineering, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium.
ninove@inma.ucl.ac.be

Abstract. We model the behaviour of a surfer who makes a walk on the webgraph favouring webpages with a high ranking.

With the growth of the World Wide Web, algorithms are needed to search and to rank webpages. PageRank [3] is a well known ranking of the webpages, which uses the graph structure of the web. Like other information retrieval methods, it considers the web as a graph. Basically, the PageRank score attributed to the webpages measures how often a given page would be visited by a random walker on the webgraph.

Formally, let $G = (V, E)$ denote a directed graph representing the web or a portion of the web: $V = \{1, \ldots, n\}$ represents the set of webpages, and $E \subset V \times V$ represents the set of hyperlinks, meaning that $(i, j) \in E$ if there is an hyperlink in page i pointing to page j. We assume for simplicity that G is strongly connected. Let C denote the adjacency matrix of G, so that $C_{ij} = 1$ if $(i, j) \in E$ and $C_{ij} = 0$ otherwise. Imagine that, when visiting page i, a websurfer chooses randomly the next webpage he will visit, among the pages pointed from page i, with the uniform distribution. Then, the trajectory of the websurfer is a Markov chain with transition matrix $M = [M_{ij}]$, given by

$$M_{ij} = \frac{C_{ij}}{\sum_k C_{ik}} \quad .$$

In its most basic version, the PageRank $\mathbf{r}$ is defined as the stationary distribution of this random walk. Thus, $\mathbf{r}$ is the invariant measure of the matrix M (which is unique because we assumed G to be strongly connected), i.e. the unique stochastic vector such that

$$\mathbf{r} = \mathbf{r}M \quad .$$

The assumption that the websurfer makes uniform draws may seem unrealistic: a websurfer may have an a priori idea of the value of pages, favouring

C. Commault and N. Marchand (Eds.): Positive Systems, LNCIS 341, pp. 239–246, 2006.

pages from "reputed" sites. Since the webrank influences the "reputation" of a site, it may influence the behaviour of websurfers.

In this paper, we present a simple mathematical model of the behaviour of a surfer who makes a walk on the webgraph, favouring webpages with a high ranking. The following process is iterated. Let $\mathbf{r}(s)$ denote the stochastic vector giving the webrank at step $s \in \mathbf{N}$. The websurfer moves from page i to page j with probability proportional to $C_{ij}e^{\mathbf{r}(s)_j/T}$, where $T > 0$ is a fixed positive parameter, that we call the *temperature*. Hence, the trajectory of this websurfer is a Markov chain with transition matrix $M(\mathbf{r}(s))$, where for all vectors $\mathbf{x}$,

$$M(\mathbf{x})_{ij} = \frac{C_{ij}e^{\mathbf{x}_j/T}}{\sum_k C_{ik}e^{\mathbf{x}_k/T}} \ .$$

The temperature T measures the randomness of the process. If T is small, with overwhelming probability, the websurfer shall move from page i to one of the pages j referenced by page i of best rank, i.e. maximizing $\mathbf{r}(s)_j$, whereas if $T = \infty$, the websurfer shall draw the next page among the pages j referenced by page i, with the uniform distribution, as in the standard webrank definition.

The updated webranking $\mathbf{r}(s+1)$ is the invariant measure of the matrix $M(\mathbf{r}(s))$. Thus, $\mathbf{r}(s+1)$ is the unique stochastic vector such that

$$\mathbf{r}(s+1) = \mathbf{r}(s+1)M(\mathbf{r}(s)) \ .$$

Note that if $\mathbf{r}(0)$ is the uniform distribution, then $\mathbf{r}(1)$ is the classical PageRank. We call *T-PageRank* the limit of $\mathbf{r}(s)$ when s tends to infinity, if it exists. For $T = \infty$, the T-PageRank coincides with the classical PageRank, because $M(\mathbf{r}(s)) = M$.

We show that, for large enough values of T, the T-PageRank is independent of the initial ranking, whereas for small values of T, several T-PageRanks exist, depending on the choice of the initial ranking. In some cases, the T-PageRank does nothing but validating the initial "belief" in the interest of pages given by the initial ranking. This suggests that webusers should not rely too much on PageRank type measures to assess the quality of pages.

We also give a simple iterative algorithm to compute the T-PageRank, at least when the temperature T is large enough and when the matrix C is primitive. We show that the T-PageRank can be obtained as the limit of the sequence $\tilde{\mathbf{r}}(s)$ defined by

$$\tilde{\mathbf{r}}(s+1) = \tilde{\mathbf{r}}(s)M(\tilde{\mathbf{r}}(s)) \ , \tag{1}$$

the initial condition $\tilde{\mathbf{r}}(0)$ being an arbitrary stochastic vector. This is similar to applying the standard power algorithm.

This paper is organized as follows. We briefly introduce some preliminaries in Section 1. Then, in Section 2, we analyse the convergence and the limit of the algorithms described above. In Section 3, we show that even for small or regular graphs, the T-PageRank can have a complex behaviour. In Section 4, we consider a variant of the model, inspired by the standard PageRank model, which allows us to deal with non strongly connected graphs. And finally, in

Section 5, we experiment the T-PageRank algorithm on a "real-world example".

In this short version of the paper, proofs are omitted. They will be presented in a forthcoming paper. The results are obtained by studying the behaviour of the iterates of some nonlinear selfmaps on the positive cone. Finally, we note that the general iteration (1) has been studied in the setting of inhomogeneous products of nonnegative matrices (see for instance [1] and the references therein). However, the main results of the present paper can not be derived from these works.

1 Preliminaries

We denote by $\mathbf{R}_{\geq 0}$ the set of nonnegative numbers and by $\mathbf{R}_{>0}$ the set of positive numbers. The simplex $\Sigma = \{\mathbf{x} \in \mathbf{R}^n_{\geq 0} \colon \sum_i \mathbf{x}_i = 1\}$ is the set of stochastic vectors. Its relative interior is denoted by $\mathrm{int}(\Sigma) = \Sigma \cap \mathbf{R}^n_{>0}$. *Hilbert's projective metric*

$$d_H \colon \mathbf{R}^n_{>0} \times \mathbf{R}^n_{>0} \to \mathbf{R}_{\geq 0} \colon (\mathbf{x}, \mathbf{y}) \mapsto \max_{i,j} \ln \frac{\mathbf{x}_i \mathbf{y}_j}{\mathbf{y}_i \mathbf{x}_j} \ ,$$

defines a distance on $\mathrm{int}(\Sigma)$ and is very useful for studying selfmaps on the positive cone. The *coefficient of ergodicity* τ_B, also known as Birkhoff's contraction coefficient, is defined for a matrix A having no zero row as

$$\tau_B(A) = \sup_{\mathbf{x},\mathbf{y} \in \mathbf{R}^n_{>0},\, \mathbf{x} \neq \lambda \mathbf{y}} \frac{d_H(\mathbf{x}A, \mathbf{y}A)}{d_H(\mathbf{x}, \mathbf{y})} \ .$$

The coefficient $\tau_B(A)$ can be written explicitly as a function of the entries of the matrix A, and it can be shown that $\tau_B(A) < 1$ if and only if A is a positive matrix. For background about iterated maps on the positive cone, nonnegative matrices and Markov chains, see respectively [5], [2] and [6].

2 Fixed Points and Convergence

Let C be an irreducible $n \times n$ nonnegative matrix. For any $T > 0$, and any $\mathbf{x} \in \Sigma$, let $M_T(\mathbf{x})$ be the irreducible stochastic matrix such that

$$M_T(\mathbf{x})_{ij} = \frac{C_{ij} e^{\mathbf{x}_j/T}}{\sum_k C_{ik} e^{\mathbf{x}_k/T}} \ .$$

2.1 Fixed Points and Convergence of $\mathbf{u}_T$

Since C is irreducible, we can define the map

$$\mathbf{u}_T \colon \mathbf{R}^n_{\geq 0} \to \Sigma \colon \mathbf{x} \mapsto \mathbf{u}_T(\mathbf{x}) \ ,$$

which sends $\mathbf{x}$ on the unique invariant measure $\mathbf{u}_T(\mathbf{x})$ of $M_T(\mathbf{x})$. Tutte's Matrix Tree Theorem [7, 4] enables us to express explicitly the invariant measure $\mathbf{u}_T(\mathbf{x})$ in terms of the entries of $M_T(\mathbf{x})$:

$$\mathbf{u}_T(\mathbf{x})_r = \mu_T(\mathbf{x}) \sum_{R \in S(r)} \prod_{(i,j) \in R} M_T(\mathbf{x})_{ij} \ ,$$

where $\mu_T(\mathbf{x}) > 0$ is a normalization factor such that $\mathbf{u}_T(\mathbf{x}) \in \Sigma$, and $S(r)$ is the set of directed subtrees of G in which all nodes except r have an outdegree equal to one. The existence of fixed points for $\mathbf{u}_T$ is then proved using Brouwer's Fixed Point Theorem.

Proposition 1. *The map $\mathbf{u}_T$ has at least one fixed point in $\operatorname{int}(\Sigma)$. Moreover, every fixed point of $\mathbf{u}_T$ is in $\operatorname{int}(\Sigma)$.*

The uniqueness of the fixed point and the convergence of the orbits of $\mathbf{u}_T$ for a sufficiently large temperature T are proved using Theorems 2.5 and 2.7 of Nussbaum in [5].

Theorem 1. *If $T \geq n$, the map $\mathbf{u}_T$ has a unique fixed point $\mathbf{x}_T$, which belongs to $\operatorname{int}(\Sigma)$. Moreover, if C is primitive, all the orbits of $\mathbf{u}_T$ converge to this fixed point.*

2.2 Fixed Points and Convergence of $\mathbf{f}_T$

We now consider the map

$$\mathbf{f}_T \colon \Sigma \to \Sigma \colon \mathbf{x} \mapsto \mathbf{x}\, M_T(\mathbf{x}) \ .$$

Theorem 2 shows that the iterates of $\mathbf{f}_T$ can be used in order to compute the fixed point of the map $\mathbf{u}_T$, if T is sufficiently large.

Theorem 2. *The fixed points of $\mathbf{u}_T$ and $\mathbf{f}_T$ are the same. Moreover, if C is primitive, then, for T sufficiently large, all the orbits of $\mathbf{f}_T$ converge to the fixed point $\mathbf{x}_T$ of $\mathbf{u}_T$.*

For positive matrices C, another convergence criterion can be derived, depending on Birkhoff's coefficient of ergodicity.

Proposition 2. *Assume that C is positive. If $T > 2(1-\tau_B(C))^{-1}$, then $\mathbf{f}_T$ has a unique fixed point $\mathbf{x}_T \in \operatorname{int}(\Sigma)$ and all the orbits of $\mathbf{f}_T$ converge to this fixed point.*

2.3 Existence of Multiple Fxed Points of $\mathbf{u}_T$ and $\mathbf{f}_T$

Theorems 1 and 2 show that for a temperature T sufficiently large, the maps $\mathbf{u}_T$ and $\mathbf{f}_T$ have a unique fixed point. We can naturally wonder about the uniqueness of the fixed point for small T: we show that, at least when C is positive, multiple fixed points always exist.

Proposition 3. *Assume that the first column of C is positive. Then, there exists a map $\mathbf{x}\colon \mathbf{R}_{>0} \to \mathrm{int}(\Sigma)\colon T \mapsto \mathbf{x}(T)$ such that for all $T > 0$, $\mathbf{x}(T)$ is a fixed point of $\mathbf{f}_T$ and $\lim_{T\to 0} \mathbf{x}(T) = \mathbf{e_1}$, where $\mathbf{e_1}$ is the first basis vector.*

It follows from Proposition 3 that if the matrix C is positive, self-validating effects appear for small values of T: arbitrary close initial rankings can induce totally different final rankings. For instance, if $C = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, the point $(\frac{1}{2}\ \frac{1}{2})$ is a fixed point of $\mathbf{u}_T$ for all T. But this fixed point is unstable if $T < \frac{1}{2}$, and two orbits starting respectively from $(\frac{1}{2}\ \frac{1}{2}) \pm (\varepsilon\ \ -\varepsilon)$ will converge to each of the two other fixed points of $\mathbf{u}_T$, which are stable, for any $\varepsilon \neq 0$.

3 Estimating the Critical Temperature in Particular Cases

In this section, we consider some particular simple cases. We are interested in temperatures for which the number of fixed points of $\mathbf{u}_T$ changes. Such a temperature is called *critical temperature*. The *first critical temperature* is the largest one, and corresponds to a loss of uniqueness of the fixed point.

3.1 Matrices of Dimension 2

The case of a graph of two nodes can be studied as a one-dimensional problem. For such graphs, the first critical temperature is always less than 1.

Proposition 4. *Suppose that C is an irreducible 2×2 matrix and $T > 1$. Then $\mathbf{u}_T$ has a unique fixed point and all its orbits converge to this fixed point.*

However, for any $T < 1$, there exist 2×2 matrices such that the fixed point of $\mathbf{u}_T$ is *not* unique. Consider for instance $C = \left(\begin{smallmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{smallmatrix}\right)$, with $0 < \varepsilon < T^{-1} - 1$. Then $\mathbf{u}_T$ has an unstable fixed point $(\frac{1}{2}\ \frac{1}{2})$ and two stable fixed points.

Depending on the parameters $\alpha = C_{11}/C_{12}$ and $\beta = C_{22}/C_{21}$, the map $\mathbf{u}_T$ can have either 1 fixed point for all T, or 1 fixed point for large T and 3 for small T, or even 1, 3 and up to 5 fixed points for some values of T. In this latter case, $\mathbf{u}_T$ can have up to three critical temperatures. Positive parameters α and β such that $\mathbf{u}_T$ has 5 fixed points for some T are depicted in Figure 1, which has been obtained experimentally.

3.2 All Ones Matrix

Let us now consider the particular case where the graph is complete. Then, C is the $n \times n$ matrix of all ones. For this matrix, $\frac{1}{n}(1 \ldots 1)$ is a fixed point of $\mathbf{u}_T$ for all T. We are interested in the existence of other fixed points, depending on the temperature T. We know from Theorem 1 that the first critical temperature is at most n. The following result shows that it may be of order $(\ln n)^{-1}$.

Proposition 5. *Suppose that C is the $n \times n$ matrix of all ones. If $T \geq 2(\ln n)^{-1}$, then $\mathbf{u}_T$ has a unique fixed point. But if $T \leq (2\ln(n+1))^{-1}$, then the map $\mathbf{u}_T$ has at least two fixed points.*

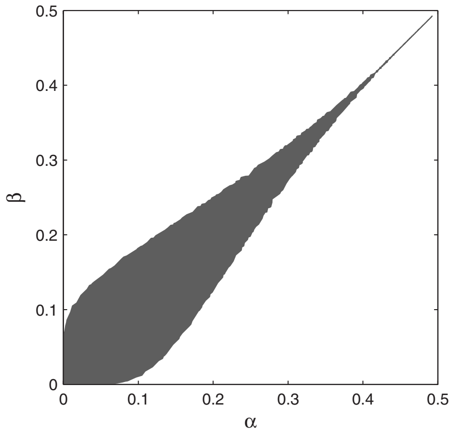

Fig. 1. For $(\alpha, \beta) \in \mathbf{R}^2_{>0}$ belonging to the coloured region, there exists some temperature T such that the corresponding map $\mathbf{u}_T$ has 5 fixed points.

4 Variant of the Model

Since the graphs considered in the applications are not necessary strongly connected, there is a need to cope with reducible matrices C. In their PageRank algorithm, Brin and Page [3] propose to add a damping factor $0 < \gamma < 1$. Now, at each step of his walk, either, with probability γ, the websurfer follows as usual the edges of the graph to move to one of the adjacent nodes, either, with probability $1 - \gamma$, he moves with uniform probability to any node of the graph. The matrix C is supposed to have no zero row, and the PageRank vector $\mathbf{r}$ is now defined by $\mathbf{r} = \mathbf{r} M_\gamma$, where $M_\gamma \in \mathbf{R}^{n\times n}_{>0}$ is given by

$$(M_\gamma)_{ij} = \gamma \, \frac{C_{ij}}{\sum_k C_{ik}} + (1 - \gamma) \, \frac{1}{n} \ .$$

We consider a similar variant for the T-PageRank, weighting the jump probability according to the ranking vector. The positive transition matrix $M_{T,\gamma}(\mathbf{x})$ is then defined as

$$M_{T,\gamma}(\mathbf{x})_{ij} = \gamma \, \frac{C_{ij} e^{\mathbf{x}_j / T}}{\sum_k C_{ik} e^{\mathbf{x}_k / T}} + (1 - \gamma) \, \frac{e^{\mathbf{x}_j / T}}{\sum_k e^{\mathbf{x}_k / T}} \ ,$$

with $T > 0$ and $0 < \gamma < 1$. The maps $\mathbf{u}_{T,\gamma}$ and $\mathbf{f}_{T,\gamma}$ are defined as previously: $\mathbf{u}_{T,\gamma}(\mathbf{x})$ is the unique invariant measure of $M_{T,\gamma}(\mathbf{x})$ and $\mathbf{f}_{T,\gamma}(\mathbf{x}) = \mathbf{x} \, M_{T,\gamma}(\mathbf{x})$. Theorem 1 can be immediately adapted in the following way.

Proposition 6. *For $T \geq 2n$, the map $\mathbf{u}_{T,\gamma}$ has a unique fixed point $\mathbf{x}_{T,\gamma}$ in Σ. Moreover, all the orbits of $\mathbf{u}_{T,\gamma}$ converge to this fixed point.*

5 Application to a Subgraph of the Web

In this section, we briefly present our experiments of the T-PageRank on a large-scale example. We consider a subgraph of the web with about 280000 nodes which has been obtained by S. Kamvar from a crawl on the Stanford web[3]. The chosen damping factor is $\gamma = 0.85$. We have computed the T-PageRank from the recurrence (1) for various temperatures T and initial rankings. As expected, when the temperature T is large, the T-PageRank is very close to the classical PageRank, and when T approaches zero, arbitrary close initial rankings can induce totally different T-PageRanks. The first critical temperature experimentally seems to be about 0.03. Figure 2 compares the T-PageRank with the classical PageRank for a temperature $T = 0.018$, the PageRank vector being taken as initial ranking. The two rankings are globally similar, see Figure 2a. However, as we can see on Figure 2b, the two best nodes are exchanged.

Acknowledgments. Part of this work was done while the third author was visiting INRIA. The third author was supported by a FNRS research fellowship and partially by research programs IAP V/22 (Belgian Federal Science Policy) and ARC (Communauté Française de Belgique). The scientific responsibility rests with its authors.

References

1. M. Artzrouni and O. Gavart, *Nonlinear matrix iterative processes and generalized coefficients of ergodicity*, SIAM J. Matrix Anal. Appl. **21** (2000), no. 4, 1343–1353.
2. A. Berman and R. J. Plemmons, *Nonnegative matrices in the mathematical sciences*, Classics in Applied Mathematics, vol. 9, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994, Revised reprint of the 1979 original.
3. S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, Proceedings of the Seventh International World-Wide Web Conference (WWW7), 1998, Brisbane, Australia.
4. R. A. Brualdi and H. J. Ryser, *Combinatorial matrix theory*, Encyclopedia of Mathematics and its Applications, vol. 39, Cambridge University Press, Cambridge, 1991.
5. R. D. Nussbaum, *Hilbert's projective metric and iterated nonlinear maps*, Mem. Amer. Math. Soc. **75** (1988), no. 391, iv+137.
6. E. Seneta, *Nonnegative matrices and Markov chains*, second ed., Springer Series in Statistics, Springer-Verlag, New York, 1981.
7. W. T. Tutte, *Graph theory*, Encyclopedia of Mathematics and its Applications, vol. 21, Addison-Wesley Publishing Company Advanced Book Program, Reading, MA, 1984.

[3] The adjacency matrix of this graph can be found on Kamvar's webpage `http://www.stanford.edu/~sdkamvar/research.html`.

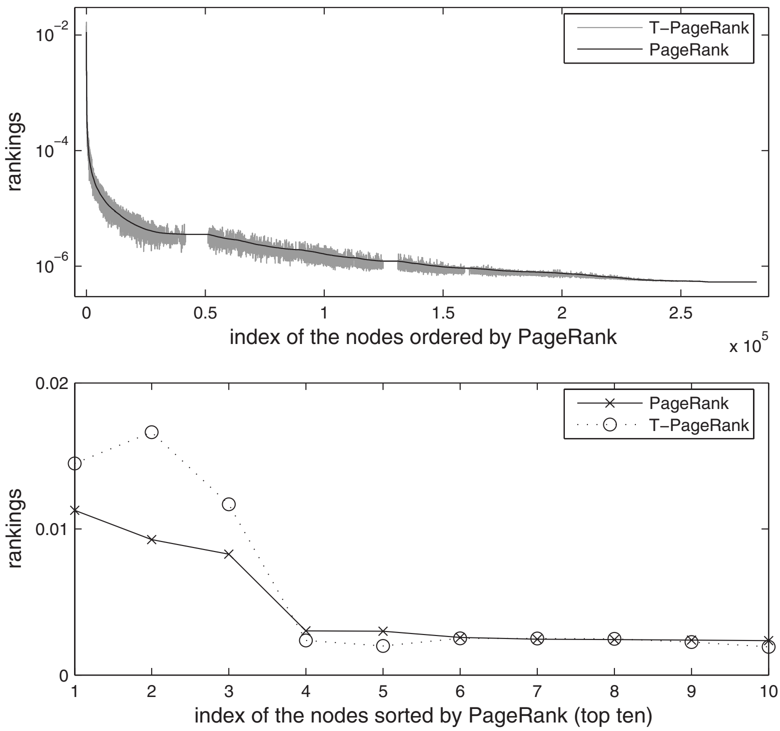

Fig. 2. Comparison of PageRank and T-PageRank ($T = 0.018$, PageRank taken as initial ranking). The nodes are sorted according to the PageRank. **a.** Rankings for every nodes. **b.** Rankings for the PageRank's top ten nodes.