An Analytical Comparison of Approaches to Personalizing PageRank

Taher Haveliwala, Sepandar Kamvar and Glen Jeh

Stanford University
{taherh,sdkamvar,glenj}@cs.stanford.edu

Abstract. PageRank, the popular link-analysis algorithm for ranking web pages, assigns a query and user independent estimate of "importance" to web pages. Query and user sensitive extensions of PageRank, which use a *basis set* of biased PageRank vectors, have been proposed in order to personalize the ranking function in a tractable way. We analytically compare three recent approaches to personalizing PageRank and discuss the tradeoffs of each one.

1 Preliminaries

In this section we summarize the definition of PageRank [7] and introduce the notation we will use subsequently.

Underlying the definition of PageRank is the following basic assumption. A link from a page $u \in Web$ to a page $v \in Web$ can be viewed as evidence that v is an "important" page. In particular, the amount of importance conferred on v by u is proportional to the importance of u and inversely proportional to the number of pages u points to. Since the importance of u is itself not known, determining the importance for every page $i \in Web$ requires an iterative fixed-point computation.

We next describe an equivalent formulation in terms of a random walk on the directed Web graph G. Let $u \to v$ denote the existence of an edge from u to v in G. Let $\deg(u)$ be the outdegree of page u in G. Consider a random surfer visiting page u at time k. In the next time step, the surfer chooses a node v_i from among u's out-neighbors $\{v|u \to v\}$ uniformly at random. In other words, at time $k+1$, the surfer lands at node $v_i \in \{v|u \to v\}$ with probability $1/\deg(u)$.

The PageRank of a page i is defined as the probability that at some particular time step $k > K$, the surfer is at page i. For sufficiently large K, and with minor modifications to the random walk, this probability is unique, illustrated as follows. Consider the Markov chain induced by the random walk on G, where the states are given by the nodes in G, and the stochastic transition matrix describing the transition from i to j is given by P with $P_{ij} = 1/\deg(i)$. If P is aperiodic and irreducible, then the Ergodic Theorem guarantees that the stationary distribution of the random walk is unique [6]. In the context of computing PageRank, the standard way of ensuring that P is irreducible is to add a new set of complete outgoing transitions, with small transition probabilities, to *all* nodes, creating a complete (and thus an aperiodic and strongly connected) transition graph.[1] Let E be the $n \times n$ rank-one row-stochastic matrix $E = \boldsymbol{e}\boldsymbol{v}^T$, where $\boldsymbol{e}$ is

[1] We ignore here the issue of *dangling nodes*, e.g., nodes with outdegree 0. See [5] for a standard way of dealing with this issue.

the n-vector whose elements are all $e_i = 1$ and $\boldsymbol{v}$ is an n-vector whose elements are all non-negative and sum to 1. We define a new, irreducible Markov chain A^T as follows:[2]

$$A = [cP + (1-c)E]^T \tag{1}$$

In terms of the random walk, the effect of E is as follows. At each time step, with probability $(1-c)$, a surfer visiting any node will jump to a random Web page (rather than following an outlink). The destination of the random jump is chosen according to the probability distribution given in $\boldsymbol{v}$. Artificial jumps taken because of E are referred to as *teleportation*.

When the vector $\boldsymbol{v}$ is nonuniform, so that E adds artificial transitions with nonuniform probabilities, the resultant PageRank vector can be biased to prefer certain kinds of pages. For this reason, we refer to $\boldsymbol{v}$ as the *personalization* vector.

2 Approaches to Personalizing PageRank

Let n be the number of pages on the web. Let $\boldsymbol{x}(\boldsymbol{v})$ denote the n-dimensional personalized PageRank vector corresponding to the n-dimensional personalization vector $\boldsymbol{v}$. $\boldsymbol{x}(\boldsymbol{v})$ can be computed by solving the following eigenvalue problem, where $A = cP^T + (1-c)\boldsymbol{v}\boldsymbol{e}^T$:

$$\boldsymbol{x} = A\boldsymbol{x} \tag{2}$$

Rewriting the above, we see that

$$\boldsymbol{x} = cP^T\boldsymbol{x} + (1-c)\boldsymbol{v} \tag{3}$$

$$\boldsymbol{x} - cP^T\boldsymbol{x} = (1-c)\boldsymbol{v} \tag{4}$$

$$(I - cP^T)\boldsymbol{x} = (1-c)\boldsymbol{v} \tag{5}$$

$I - cP$ is strictly diagonally dominant, so that $I - cP$ is invertible. Therefore, $(I - cP)^T = I - cP^T$ is also invertible. Thus, we get that

$$\boldsymbol{x} = (1-c)(I - cP^T)^{-1}\boldsymbol{v} \tag{6}$$

Let $Q = (1-c)(I - cP^T)^{-1}$. By letting $\boldsymbol{v} = \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the ith elementary vector[3] we see that the ith column of the matrix Q is $\boldsymbol{x}(\boldsymbol{e}_i)$, i.e., the personalized PageRank vector corresponding to the personalization vector $\boldsymbol{e}_i$.

The columns of Q comprise a complete basis for personalized PageRank vectors, as any personalized PageRank vector can be expressed as a convex combination of the columns of Q. For any personalization vector $\boldsymbol{v}$, the corresponding personalized PageRank vector is given by $Q\boldsymbol{v}$. This formulation corresponds to the original approach to personalizing PageRank suggested by Page et al. [7] that allows for personalization on arbitrary sets of pages.

[2] We define the chain in terms of the transpose so that we can discuss right (rather than left) eigenvectors.

[3] i.e., $\boldsymbol{e}_i$ has a 1 in the ith component, and zeros elsewhere

Unfortunately, this first approach, which uses the complete basis for personalized PageRank, is infeasible in practice. Computing the dense matrix Q offline is impractical, as is computing $\boldsymbol{x}(\boldsymbol{v})$ at query time using the Power Method.

However, we can compute low-rank approximations of Q, denoted as $\hat{Q}$, that still allow us to achieve a part of the benefit of fully personalized PageRank. Rather than using a full basis (i.e., the columns of Q), we can choose to use a reduced basis, e.g., using only $k \leq n$ personalized PageRank vectors, each of which is a column (or more generally, a convex combination of the columns) of Q. In this case, we cannot express all personalized PageRank vectors, but only those corresponding to convex combinations of the PageRank vectors in the reduced basis set:

$$\boldsymbol{x}(\boldsymbol{w}) = \hat{Q}\boldsymbol{w} \tag{7}$$

where w is a stochastic k-vector representing weights over the k basis vectors.

The following three approaches each approximate Q with some approximation $\hat{Q}$, although they differ substantially in their computational requirements and in the granularity of personalization achieved.

Topic-Sensitive PageRank. The Topic-Sensitive PageRank scheme proposed by Haveliwala [2] computes an $n \times k$ approximation to Q using k topics, e.g., the 16 top level topics of the Open Directory [1]. Column j of $\hat{Q}$ is given by $\boldsymbol{x}(\boldsymbol{v}_j)$, where $\boldsymbol{v}_j$ is a dense vector generated using a classifier for topic T_j; $(v_j)_i$ represents the (normalized) degree of membership of page i to topic j. Note that in this scheme, each column of $\hat{Q}$ must be generated independently, so that k must be kept fairly small (e.g., $k = 16$). This scheme uses a fairly coarse basis set, making it more suitable for modulating the rankings based on the topic of the query and query context, rather than for truly "personalizing" the rankings to a specific individual. The use of a good set of representative basis topics ensures that the approximation $\hat{Q}$ will be useful.

In Topic-Sensitive PageRank, $\hat{Q}$ is generated completely offline. Convex combinations are taken at query time, using the context of the query to compute the appropriate topic weights.

In terms of the random surfer model of PageRank, this scheme restricts the choice of teleportation transitions so that the random surfer can teleport to a topic T_j with some probability w_j, followed by a teleport to a particular page i with probability $(v_j)_i$.

Modular PageRank. The Modular PageRank approach proposed by Jeh and Widom [3] computes an $n \times k$ matrix using the k columns of Q corresponding to highly ranked pages. In addition, that work provides an efficient scheme for computing these k vectors, in which *partial* vectors are computed offline and then composed at query time, making it feasible to have $k \geq 10^4$.

In terms of the random surfer model of PageRank, this scheme restricts the choice of teleportation transitions so that the random surfer can teleport to certain highly ranked pages, rather than to arbitrarily chosen sets of pages.

A direct comparison of the relative granularity of this approach to the topic-sensitive approach is difficult; although the basis set of personalized PageRank vectors is much larger in this scenario, they must come from personalization vectors $\boldsymbol{v}$ with singleton nonzero entries corresponding to highly ranked pages. However, the larger size of the basis set does allow for finer grained modulation of rankings.

BlockRank. The BlockRank algorithm proposed by Kamvar et al. [4] computes an $n \times k$ matrix corresponding to k "blocks". E.g, in that work, each block corresponds to a host, such as www-db.stanford.edu or nlp.stanford.edu. That work computes a matrix $\hat{Q}$ in which column j corresponds to $\boldsymbol{x}(\boldsymbol{v_j})$, where $\boldsymbol{v_j}$ represents the *local PageRank* of the pages in block j. The BlockRank algorithm is able to exploit the Web's inherent block structure to efficiently compute many of these block-oriented basis vectors, so that $k \geq 10^3$ is feasible.

In terms of the random surfer model of PageRank, this scheme restricts the choice of teleportation transitions so that the random surfer can teleport to block B_j with probability w_j, followed by a teleport to a particular page i in block B_j with probability $(v_j)_i$, rather than to arbitrary sets of pages.

Again, a direct comparison of the granularity of this approach with the previous two is difficult. However, the BlockRank approach allows for a large number of basis vectors without the restriction that the underlying personalization vectors be derived from highly ranked pages.

Acknowledgements

This paper is based on work supported in part by the National Science Foundation under Grant No. IIS-0085896 and Grant No. CCR-9971010, and in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University (research project on Concept Bases for Lexical Acquisition and Intelligently Reasoning with Meaning).

References

1. The Open Directory Project: Web directory for over 2.5 million URLs. http://www.dmoz.org/.
2. T. H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.
3. G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.
4. S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Exploiting the block structure of the web for computing PageRank. *Stanford University Technical Report*, 2003.
5. S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Extrapolation methods for accelerating PageRank computations. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.
6. R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, United Kingdom, 1995.
7. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. *Stanford Digital Libraries Working Paper*, 1998.