# Web Information Retrieval

## Crash notes on Personalized Pagerank

### Iterative computation of Personalized Pagerank

Assume you want to compute personalized pagerank with respect to some personalization vector $\mathbf{p}$. For example, $\mathbf{p}$ might be:

$$\mathbf{p_i} = \begin{cases} \frac{1}{|S|}, & i \in S, \\ 0 & otherwise \end{cases}$$

where $S$ denotes a seed set. Note that $\mathbf{p}$ might in general be any probability distribution over the $n$ pages in of the Web graph, not necessarily uniform over a given seed set.

For any personalization vector $\mathbf{p}$, we know that for a generic page $i$ we have:

$$\boxed{\pi_i(t) = (1-\alpha)\mathbf{p}_i + \alpha \sum_{j \to i} \frac{\pi_j(t-1)}{d_j},}$$

where $1-\alpha$ is the teleportation probability and $d_j$ denotes $j$'s out-degree (we assume all sinks were previously removed by linking each of them to each node in $S$).

A compact and equivalent way of writing this equation is:

$$\boxed{\pi(t)^T = \pi(t-1)^T \left((1-\alpha)\mathbf{1}\mathbf{p}^\mathbf{T} + \alpha M\right),}$$

where $M$ is the transition matrix of the underlying graph (after removal of sinks). Of course, this is exactly the original, iterative formulation of pagerank, with the only difference that we changed the vector of teleportation probabilities from being uniform over the entire set of Web pages to being defined by some distribution $\mathbf{p}$. As a consequence, the very same version of the power method using *pagerank leaking* can be applied, thus exploiting the sparseness of $M$. Note that, in general, the following happens: i) the personalized pagerank exists, since the Markov chain we obtain is still irreducible and aperiodic; ii) the stationary distribution will in general differ from the original pagerank. *You should elaborate on why this is the case*.

## Composability of Personalized Pagerank

Next, let

$$P = (1-\alpha)\mathbf{1}\mathbf{p}^\mathbf{T} + \alpha M$$

Of course we have $\pi(t)^T = \mathbf{p}^T P^t$, given that $\pi(0) = \mathbf{p}$.

### Building personalization vectors for single users

Assume we have personalization vectors $\mathbf{q}^1, \ldots, \mathbf{q}^{(k)}$ for some number $k$ of "broad" topics. These might for example be "sports", "news", "movies" etc. Note that each $\mathbf{q}^{(r)}$ is a probability distribution over the entire set pages of Web pages, possibly placing $0$ mass on some (or even the majority) of them. So, each such vector has $n$ entries, with $n$ the number of Web pages. For the sake of exposition, we call *broad personalization vectors* the $\mathbf{q}^{(r)}$'s and we call *broad personalized pagerank vectors* the $k$ corresponding pagerank vectors that correspond to them in the remainder of these notes. Assume now that some user's interests are succintly described by a distribution $\mathbf{s}$

over the set of $k$ topics. The $\ell$-th entry of vector $\mathbf{s}$ can for instance be estimated from the fraction of times the user clicked on pages that are relevant for the $\ell$-th topic when returned result sets corresponding to her queries. It should be emphasized that $\mathbf{s}$ can be computed in many ways, more or less refined, depending on users' profile information available to the search engine. Next, let

$$Q = (\mathbf{q}^1 \ \ldots \mathbf{q}^{(k)})$$

I.e., $Q$ is an $n \times k$ matrix whose columns are the broad personalization vectors. A personalization vector $\mathbf{p}$ for the user under consideration can then be obtained as $\mathbf{p} = Q\mathbf{s} = \sum_{r=1}^{k} \mathbf{s}_r \mathbf{q}^{(r)}$. Note that $\mathbf{p}$ is still a distribution over the set of the $n$ Web pages (*you are warmly invited to check this yourself*). On the other hand, if we consider the pagerank vector $\pi(t)$ after $t$ steps with personalization vector $\mathbf{p}$ we obtain:

$$\pi(t)^T = \mathbf{p}^T P^t = \left( \sum_{r=1}^{k} \mathbf{s}_r \mathbf{q}^{(r)} \right)^T P^t = \sum_{r=1}^{k} \mathbf{s}_r (\mathbf{q}^{(r)})^T P^t = \sum_{r=1}^{k} \mathbf{s}_r (\pi^{(r)}(t))^T,$$

where $\pi^{(r)}$ denotes the personalized pagerank corresponding to personalization vector $\mathbf{q}^{(r)}$ after $t$ steps, while the last equality follows immediately from the linearity of matrix transposition. Since this equality holds for every value of $t$, it also holds for the limiting, stationary distribution, i.e., the personalized pagerank corresponding to personalization vector $\mathbf{p}$ (the argument is *almost* formal and this will be enough for these notes).

## Meaning and advantages

In practice, the derivation above implies that we do not need to explicitely compute personalized pagerank vectors (using the power method) for every user, whenever personalization vectors reflect users' interests with respect to a set of $k$ given "broad" topics. Rather, it is enough to compute, once and for all, $k$ personalized pagerank vectors, one for each of the aforementioned topics. At this point, if the personalization vector of a user is expressed as a linear combination of the $k$ broad personalization vectors, his/her personalized pagerank is a linear combination of the corresponding broad personalized pagerank vectors, according to the same coefficients. This makes it possible to compute personalized pagerank vectors for different user profiles (e.g., reflecting different user categories) on the fly, without having to run the power method every time.