Graph mining
SD212
PageRank

Thomas Bonald

2017 – 2018

These lecture notes introduce PageRank (originally designed for the Web) and related metrics to rank the nodes of a graph in terms of frequency of visits by a random walk.

1 Notation

Let $G = (V, E)$ be a directed graph of n nodes and m edges, with $V = \{1, \ldots, n\}$. Each edge is assigned a positive weight. The particular case where all weights are equal to 1 corresponds to an unweighted graph. We denote by A the weighted adjacency metric: for each $i, j \in V$, A_{ij} is the weight of the edge (i, j), if any, and 0 otherwise. Let $w_i^- = \sum_{j \in V} A_{ji}$ and $w_i^+ = \sum_{j \in V} A_{ij}$ be the in-weights and out-weights of node i. We say that node i is a source if $w_i^- = 0$ and a sink if $w_i^+ = 0$. Unless otherwise specified, we assume that the graph G has no sink.

2 Random walk

Consider a random walk in the graph G with a probability of moving from node i to node j equal to A_{ij}/w_i. Let $X_0, X_1, X_2, \ldots$ be the sequence of nodes visited by the random walk. This defines an irreducible Markov chain on $\{1, \ldots, n\}$ with transition matrix $P = D^{-1}A$, where $D = \mathrm{diag}(w^+)$. We have for all $t \geq 1$:

$$\forall i = 1, \ldots, n, \quad \mathrm{P}(X_t = i) = \sum_{j=1}^{n} \mathrm{P}(X_{t-1} = j) P_{ji}.$$

Denoting the distribution of X_t as a row vector π_t, we get:

$$\pi_t = \pi_{t-1} P, \tag{1}$$

so that $\pi_t = \pi_0 P^t$, where π_0 is the initial distribution. If the graph is strongly connected and aperiodic (that is, the largest common divisor of the cycle lengths is equal to 1), the following limit exists and is unique:

$$\pi = \lim_{t \to +\infty} \pi_t. \tag{2}$$

This is the stationary distribution, which satisfies the balance equations:

$$\pi = \pi P. \tag{3}$$

In particular, π is the unique left eigenvector of P for the eigenvalue 1 such that $\pi 1 = 1$ (observe that $P1 = 1$, that is, 1 is the corresponding right eigenvector). The vector π gives the frequency of visits of the random walk to each node, and as such provides a natural ranking of the nodes. In general, π cannot be computed exactly but, in view of (1), can be approximated by successive matrix-vector multiplications. It is independent of the initial distribution π_0.

Remark 1 *It can be shown that the sequence π_t converges to π at an exponential rate equal to the modulus of the second largest eigenvalue of P.*

Remark 2 *If the graph is undirected, it can be easily verified that $\pi \propto w$, with $w \equiv w^+$: the frequency of visits to a node is proportional to its weight (equal to its degree for unit weights).*

3 PageRank

If the graph G contains sinks, the random walk is no longer defined. A natural approach consists in letting the random walk jump to any node chosen uniformly at random in V (in particular, the random walk stays in the same node with probability $1/n$). The transition matrix becomes:

$$P_{ij} = \begin{cases} \frac{A_{ij}}{w_i^+} & \text{if } w_i^+ > 0, \\ \frac{1}{n} & \text{otherwise.} \end{cases} \tag{4}$$

The random walk is then well defined but not necessarily irreducible (some nodes may still be not accessible from some other nodes). To get irreducibility, one may let the random walk continue with probability α and restart with probability $1-\alpha$, for some $\alpha \in (0,1)$. The corresponding transition matrix is:

$$P^{(\alpha)} = \alpha P + (1-\alpha)\frac{11^T}{n}.$$

The corresponding stationary distribution $\pi^{(\alpha)}$ is called the PageRank vector, as proposed by the founders of Google [3]. We have:

$$\pi^{(\alpha)} = \alpha\pi^{(\alpha)}P + (1-\alpha)\frac{1^T}{n}. \tag{5}$$

The parameter α is known as the damping factor. Observe that the path length of the random walk until restart has a geometric distribution with parameter $1-\alpha$. In particular, the average path length is:

$$\frac{\alpha}{1-\alpha}.$$

The default value $\alpha = 0.85$ corresponds to an average path length of 5.7, which is typical of the distance between two nodes of real graphs (cf. the Small World property).

The PageRank vector can be computed very efficiently by successive matrix-vector multiplications, as described below. With some appropriate data structure for the transition matrix (e.g., the Compressed Sparse Row format), the complexity is in $O(km)$ where k is the number of iterations.

```
PageRank

Input:
$P$, transition matrix, given by (4)
$\alpha$, damping factor
$k$, number of iterations

Do:
$\pi \leftarrow \frac{1}{n}(1,\dots,1)$
For $t = 1,\dots,k$, $\pi \leftarrow \alpha\pi P + (1-\alpha)\frac{1}{n}(1,\dots,1)$

Output:
$\pi$, PageRank vector
```

The following result shows that the PageRank vector is the smoothing average of the distributions π_t of the pure random walk (without damping) at times $t = 0, 1, 2, \ldots$, with π_0 the uniform distribution.

Proposition 1 *We have:*

$$\pi^{(\alpha)} = (1-\alpha) \sum_{t=0}^{+\infty} \alpha^t \pi_t. \tag{6}$$

Proof. It is sufficient to check that the row vector $\pi^{(\alpha)}$ defined by (6) satisfies (5):

$$\begin{aligned} \alpha \pi^{(\alpha)} P + (1-\alpha)\frac{1^T}{n} &= \alpha(1-\alpha) \sum_{t=0}^{+\infty} \alpha^t \pi_t P + (1-\alpha)\pi_0, \\ &= (1-\alpha) \sum_{t=0}^{+\infty} \alpha^{t+1} \pi_{t+1} + (1-\alpha)\pi_0, \\ &= \pi^{(\alpha)}. \end{aligned}$$

□

Observe that $\pi^{(\alpha)} = \pi_0 + \alpha(\pi_1 - \pi_0) + o(\alpha)$. Since π_0 is the uniform distribution, the ranking is that induced by the sampling of a random neighbor when $\alpha \to 0$. When $\alpha \to 1$, the ranking is that induced by the limit of π_t when $t \to +\infty$.

The approximation provided by the first k jumps of the random walk (see the above algorithm) amounts to truncating the sum (6), namely to approximating $\pi^{(\alpha)}$ by

$$\alpha^k \pi_k + (1-\alpha) \sum_{t=0}^{k-1} \alpha^t \pi_t.$$

4 Personalized PageRank

While PageRank provides a global ranking of the nodes, it is interesting in practice to get a local ranking, relative to some target node(s). This is the objective of Personalized PageRank, used by Web search engines to display the most relevant pages relative to some request.

Let $s \in V$ be some target node. The idea of Personalized PageRank is to rank nodes relative to their frequency of visits for a random walk (re)starting from that node. Specifically, the transition matrix of the random walk is such that:

$$\forall j \neq s, \quad P_{ij}^{(\alpha)} = \begin{cases} \alpha \frac{A_{ij}}{w_i^+} & \text{if } w_i^+ > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding stationary distribution $\pi^{(\alpha)}$ is called the Personalized PageRank vector. The expression (6) remains valid, with $\pi_0 = 1_s^T$ (unit row vector on s) and π_t the distribution after t jumps from s (with restart to s from any sink). The parameter α controls the “locality” of the ranking, the neighbors of s being favored when $\alpha \to 0$.

The Personalized PageRank can be generalized to some set $S \subset V$ of target nodes, with relative weights captured by some distribution μ on S. The transition matrix of the random walk becomes:

$$\forall j \notin S, \quad P_{ij}^{(\alpha)} = \begin{cases} \alpha \frac{A_{ij}}{w_i^+} & \text{if } w_i^+ > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\forall j \in S, \quad P_{ij}^{(\alpha)} = \begin{cases} (1-\alpha)\mu_j & \text{if } w_i^+ > 0, \\ \mu_j & \text{otherwise.} \end{cases}$$

We give below the Personalized PageRank algorithm in the particular case where the distribution μ is uniform on S.

```
Personalized PageRank

Input:
$P$, transition matrix, given by (4)
$S$, set of target nodes
$\alpha$, damping factor
$k$, number of iterations

Do:
$\mu \leftarrow 1_S^T/|S|$
$\pi \leftarrow \mu$
For $t = 1, \dots, k$, $\pi \leftarrow \alpha \pi P + (1-\alpha)\mu$

Output:
$\pi$, Personalized PageRank vector
```

It turns out that, in the absence of sinks, the Personalized PageRank vector associated with some distribution μ on S follows from the Personalized PageRank vectors associated with the nodes $s \in S$ taken individually. This is interesting from a mathematical point of view only; for computations, it is preferable to use the algorithm above, whose complexity is in $O(km)$ independently of the cardinality of the set S.

Proposition 2 *In the absence of sinks, we have:*

$$\pi^{(\alpha)} = \sum_{s \in S} \mu_s \pi_s^{(\alpha)}, \tag{7}$$

where $\pi_s^{(\alpha)}$ is the Personalized PageRank vector associated with node s.

Proof. Observing that

$$P^{(\alpha)} = \alpha P + (1-\alpha) \sum_{s \in S} \mu_s 1 1_s^T,$$

with $P = D^{-1}A$, we get

$$\begin{aligned} \pi^{(\alpha)}(\alpha P + (1-\alpha)\sum_{s \in S} \mu_s 1 1_s^T) &= \alpha \pi^{(\alpha)} P + (1-\alpha) \sum_{s \in S} \mu_s 1_s^T, \\ &= \alpha \sum_{s \in S} \mu_s \pi_s^{(\alpha)} P + (1-\alpha) \sum_{s \in S} \mu_s 1_s^T, \\ &= \sum_{s \in S} \mu_s (\alpha \pi_s^{(\alpha)} P + (1-\alpha) 1_s^T), \\ &= \sum_{s \in S} \mu_s \pi_s^{(\alpha)}. \end{aligned}$$

□

Observe that the result is no longer valid in the presence of sinks, due to the restart mechanism.

5 Forward-Backward PageRank

The above PageRank metrics assess the importance of each node through its frequency of visits by a random walk. In Personalized PageRank for instance, the neighbors of the target node s typically get the top ranks (at least for low values of α). In many practical cases, it makes more sense to rank first those nodes having the same successors as s (respectively, the same predecessors). In Wikipedia for instance, two pages having a large number of common links (that is, successors in the corresponding graph) are typically closely related.

This motivates the following *forward-backward* random walk, where edges are followed in forward and backward directions alternately. Starting from any non-sink node i, the random walk first moves to node k with probability A_{ik}/w_i^+, then to node j with probability A_{jk}/w_k^- (observe that $w_k^- > 0$ because k is a successor of i). The probability of moving from node i to node j is then:

$$P_{ij} = \sum_{k \in V:(i,k)\in E} \frac{A_{ik}}{w_i^+}\frac{A_{jk}}{w_k^-}.$$

This random walk is perfectly defined over the set of non-sink nodes: unlike the regular random walk, the forward-backward random walk can never reach a sink (if it reaches a sink in the forward direction, it will escape from it in the backward direction). The probability of moving from i to j is high if these nodes have a large number of common successors k.

Remark 3 *When there are neither sinks nor sources in the graph, the transition matrix is simply:*

$$P = P^+P^-,$$

where P^+ is the transition matrix of the random walk in forward direction (as considered so far) and P^- be the transition matrix of the random walk in backward direction.

Remark 4 *If the graph is undirected, then $P^+ = P^-$ and $P = (P^+)^2$. This corresponds to a 2-hop random walk, leading to a ranking of nodes different from that induced by the 1-hop random walk of PageRank.*

Like the regular random walk, the forward-backward random walk is not irreducible in general. To get a global ranking, we introduce a damping factor $\alpha \in (0,1)$ as in PageRank. Specifically, the random walk goes on with probability α and restarts with probability $1-\alpha$. This ensures that the random walk has a unique stationary distribution $\pi^{(\alpha)}$, which we refer to as the Forward-Backward PageRank vector. In the particular case where there are neither sinks nor sources in the graph (see Remark 3), we get:

$$\pi^{(\alpha)} = \alpha\pi^{(\alpha)}P^+P^- + (1-\alpha)\frac{1^T}{n}. \tag{8}$$

To understand the meaning of this Forward-Backward PageRank, it is useful to introduce the co-citation graph G^{co} associated with the initial graph G. This is the weighted, undirected graph associated with the weighted adjacency matrix:

$$A^{\text{co}}_{ij} = \sum_{k\in V} \frac{A_{ik}A_{jk}}{w_k^-}.$$

Note that the weight of node i in the co-citation graph is the out-weight of i in the original graph:

$$\sum_{j\in V} A^{\text{co}}_{ij} = \sum_{k\in V} A_{ik} \sum_{j\in V}\frac{A_{jk}}{w_k^-} = \sum_{k\in V} A_{ik} = w_i^+.$$

In particular, the forward-backward random walk in the original graph corresponds to a regular random walk in the co-citation graph, since the probability of moving from some non-sink node i to j is:

$$P_{ij} = \frac{A^{\text{co}}_{ij}}{w_i^+}.$$

In principle, the Forward-Backward PageRank can be computed from the PageRank in the co-citation graph. In practice, this is not desirable (or simply not feasible) because the co-citation graph G^{co} is typically much denser than the original graph G. Indeed, any node k of G forms a clique of d_k^- nodes in G^{co}, so that the number of edges in the co-citation graph may be as large as:

$$\sum_{k\in V}(d_k^-)^2.$$

For a power-law distribution of in-degrees, which is typical of real graphs, this may be huge, and in any case much larger than the number of edges m in the original graph, given by

$$m = \sum_{k\in V} d_k^-.$$

We give below the Forward-Backward PageRank algorithm using the original graph (and not the co-citation graph), assuming that the transition matrices P^+ and P^- are well defined (neither sinks nor sources).

```
Forward-Backward PageRank

Input:
$P^+$ and $P^-$, forward and backward transition matrices
$\alpha$, damping factor
$k$, number of iterations

Do:
$\pi \leftarrow \frac{1}{n}(1,\dots,1)$
For $t = 1,\dots,k$,
$\pi \leftarrow \pi P^+$
$\pi \leftarrow \alpha\pi P^- + (1-\alpha)\frac{1}{n}(1,\dots,1)$

Output:
$\pi$, Forward-Backward PageRank vector
```

We may define similarly the Backward-Forward PageRank, giving a different ranking of nodes. These two types of ranking were originally proposed by Kleinberg in the HITS[1] algorithm [1] to differenciate so-called "hubs" and "authorities" in the Web. The Forward-Backward PageRank and Backward-Forward PageRank presented here are described in [2]. Both have their personalized versions enabling local ranking relative to some target nodes.

References

[1] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[2] R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, 19(2):131–160, 2001.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[1] Hyertext-Induced Topic Search.