Chapter 21 Link analysis

1. PageRank

Consider a random surfer who begins at a web page (a node of the web graph) and executes a random walk on the Web as follows. At each time step, the surfer proceeds from his current page A to a randomly chosen web page that A hyperlinks to. As the surfer proceeds in this random walk from node to node, he visits some nodes more often than others; intuitively, these are nodes with many links coming in from other frequently visited nodes. The idea behind PageRank is that pages visited more often in this walk are more important.

What if the current location of the surfer, node A, has no out-links? To address this we introduce an additional operation for our random surfer: the teleport operation. In the teleport operation, the surfer jumps from a node to any other node in the web graph.

In assigning a PageRank score to each node of the web graph, we use the teleport operation in two ways:

- When at a node with no out-links, the surfer invokes the teleport operation.
- At any node that has outgoing links, the surfer invokes the teleport operation with probability $0 < \alpha < 1$ and the standard random walk with probability $1-\alpha$, where α is a fixed parameter chosen in advance. Typically, α might be 0.1.

1.1 Markov chains

A Markov chain is a discrete-time stochastic process, a process that occurs in a series of time steps in each of which a random choice is made.

A Markov chain is characterized by an $N \times N$ transition probability matrix P each of whose entries is in the interval $[0, 1]$; the entries in each row of P add up to 1. The Markov chain can be in one of the N states at any given time-step; then, the entry P_{ij} tells us the probability that the state at the next time-step is j, conditioned on the current state is i. Each entry P_{ij} is known as a transition probability and depends only on the current state i; this is known as the Markov property. Thus, by the Markov property,

$$\forall i, j,\ P_{ij} \in [0, 1]$$

and

$$\forall i,\ \sum_{j=1}^{N} P_{ij} = 1 \tag{1}$$

A matrix with non-negative entities that satisfies Equation (1) is known as a stochastic matrix.

In a Markov chain, the probability distribution of next states for a Markov chain depends only on the current state, and not on how the Markov chain arrived at the current state.

We can view a random surfer on the web graph as a Markov chain, with one state for each web page, and each transition probability representing the probability of moving from one web page to another. The teleport operation contributes to these transition probabilities. The adjacency matrix A of the web graph is defined as follows: if there is a hyperlink from page i to page j, then $A_{ij} = 1$, otherwise $A_{ij} = 0$. If a row of A has no 1's, then divide each element by $1/N$.

Then we could construct the transition matrix as follows:

- Divide each 1 in A by the number of 1s in its row. Thus, if there is a row with three 1s, then each of them is replaced by 1/3.
- Multiply the resulting matrix by $1-\alpha$.
- Add α/N to every entry of the resulting matrix, to obtain P.

A Markov chain is said to be *ergodic* if there exists a positive integer T_0 such that for all pairs of states i, j in the Markov chain if it is started at time 0 in state i then for all $t > T_0$, the probability of being in state j at time t is greater than 0.

For a Markov chain to be ergodic, two technical conditions are required of its states and the nonzero transition probabilities; these conditions are known as *irreducibility* and *aperiodicity*.

- *Irreducibility* ensures that there is a sequence of transitions of nonzero probability from any state to any other.
- *Aperiodicity* ensures that the states are not partitioned into sets such that all state transitions occur cyclically from one set to another.

For any ergodic Markov chain, there is a unique steady-state probability vector $\overrightarrow{\pi}$ that is the principal left eigenvector of P, such that if $\eta(i, t)$ is the number of visits to state i in t steps, then

$$\lim_{t \to \infty} \frac{\eta(i, t)}{t} = \pi(i) \tag{2}$$

where $\pi(i) > 0$ is the steady-state probability for state i.

Consequently, the random walk with teleporting results in a unique distribution of steady-state probabilities over the states of the induced Markov chain. This steady-state probability for a state is the PageRank of the corresponding web page.

Begin at a state and run the walk for a large number of steps t, keeping track of the visit frequencies for each of the states. After a large number of steps t, these frequencies "settle down" so that the variation in the computed frequencies is below some predetermined threshold. We declare these tabulated frequencies to be the PageRank values.

1.2 Topic-specific PageRank

PageRank supposes the surfer jumps from one web to any other linked web is equality. However, it may not be true in practice. For instance, a sports-centered web may have a higher probability and frequency to jump to another sports relevant page, rather than other webs. Therefore, our random surfer teleports to a random web page on the topic of sports instead of teleporting to a uniformly chosen random web page.

Provided the set S of sports-related pages is nonempty, it follows that there is a nonempty set of web pages $Y \supseteq S$ over which the random walk has a steady-state distribution; let us denote this sports PageRank distribution by $\overrightarrow{\pi}_s$. For web pages not in Y, we set the PageRank values to 0. We call $\overrightarrow{\pi}_s$ the topic-specific PageRank for sports.

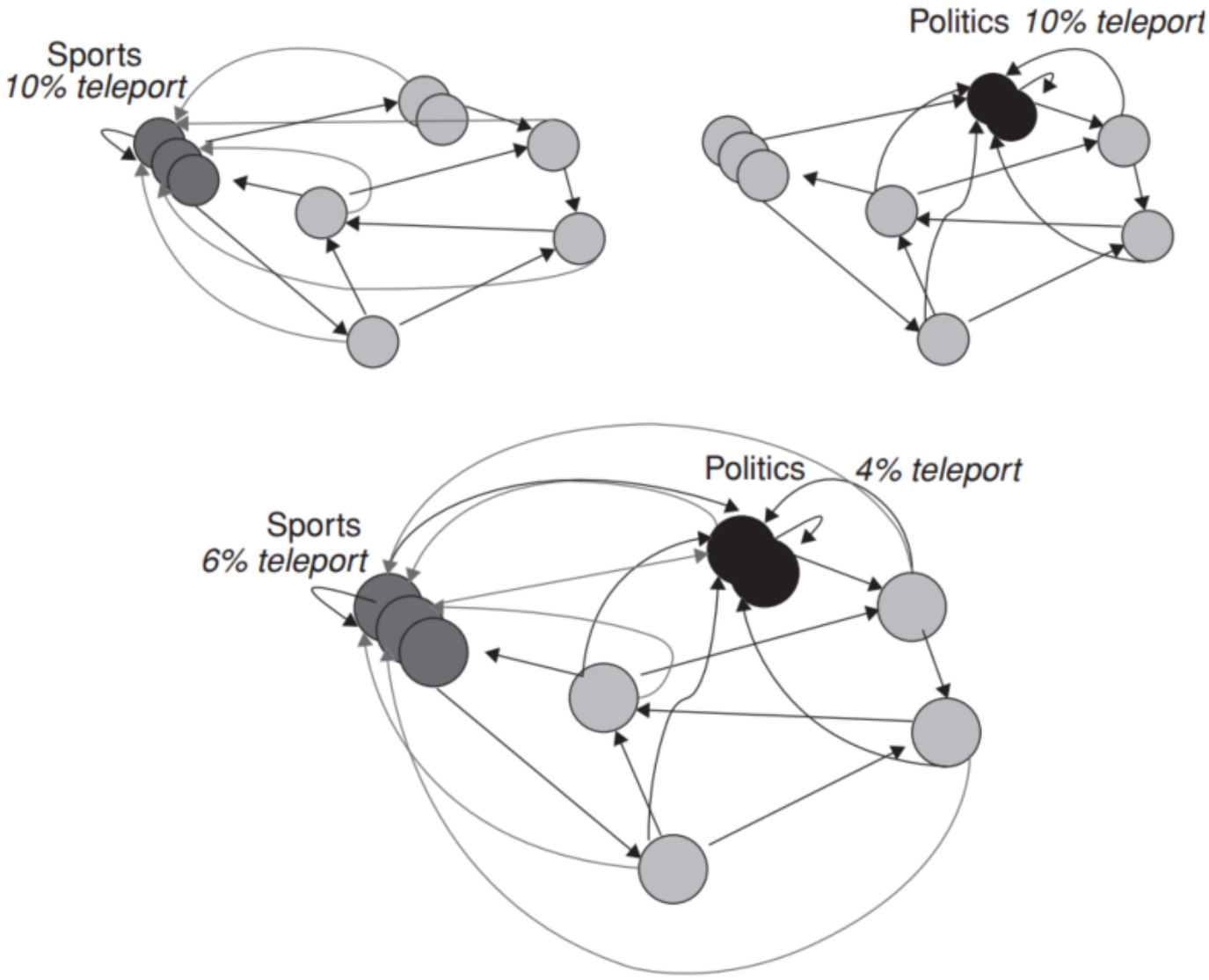

Figure 21.5 Topic-specific PageRank. In this example we consider a user whose interests are 60% sports and 40% politics. If the teleportation probability is 10%, this user is modeled as teleporting 6% to sports pages and 4% to politics pages.

The personalized PageRank vector for any user can be expressed as a linear combination of the underlying topic-specific PageRanks. For instance, the personalized PageRank vector for the user whose interests are 60% sports and 40% politics can be computed as

$$0.6\overrightarrow{\pi}_s + 0.4\overrightarrow{\pi}_p$$

where $\overrightarrow{\pi}_s$ and $\overrightarrow{\pi}_p$ are the topic-specific PageRank vectors for sports and for politics, respectively.

2. Hubs and authorities

We now develop a scheme in which, given a query, every web page is assigned two scores. One is called its hub score and the other its authority score. For any query, we compute two ranked lists of results rather than one. The ranking of one list is induced by the hub scores and that of the other by the authority scores.

- **Authoritative page** The web page process authoritative sources of information on the topic.
- **Hub page** There are many pages on the Web that are hand compiled lists of links to authoritative web pages on a specific topic. These hub pages are not in themselves authoritative sources of topic-specific information, but rather compilations that someone with an interest in the topic has spent time putting together.

A good hub page is one that points to many good authorities; a good authority page is one that is pointed to by many good hub pages.

For a web page v in our subset of the web, we use $h(v)$ to denote its hub score and $a(v)$ its authority score. Initially, we set $h(v) = a(v) = 1$ for all nodes v. We also denote by $v \longmapsto y$ the existence of a hyperlink from v to y.

Thus,

$$\begin{aligned} h(v) &\leftarrow \sum_{v \longmapsto y} a(y) \\ a(v) &\leftarrow \sum_{y \longmapsto v} h(y) \end{aligned} \tag{3}$$

Let $\overrightarrow{h}$ and $\overrightarrow{a}$ denote the vectors of all hub and all authority scores respectively, for the pages in our subset of the web graph. Let A denote the adjacency matrix of the subset of the web graph that we are dealing with: A is a square matrix with one row and one column for each page in the subset. The entry A_{ij} is 1 if there is a hyperlink from page i to page j, and 0 otherwise. Then, we may write Equation (3)

$$\begin{aligned} \overrightarrow{h} &\leftarrow A\overrightarrow{a} \\ \overrightarrow{a} &\leftarrow A^T\overrightarrow{h} \end{aligned} \tag{4}$$

and,

$$\begin{aligned} \overrightarrow{h} &\leftarrow AA^T\overrightarrow{h} \\ \overrightarrow{a} &\leftarrow A^TA\overrightarrow{a} \end{aligned} \tag{5}$$

Given any matrix A, and denote x and λ as the eigenvector and eigenvalue of A. We have,

$$\begin{aligned} Ax &= \lambda x \\ x &= A^T \lambda x \end{aligned} \tag{6}$$

Introduce equation (6) to Equation(5), we have,

$$\begin{aligned} \overrightarrow{h} &= (1/\lambda_h) A A^T \overrightarrow{h} \\ \overrightarrow{a} &= (1/\lambda_a) A^T A \overrightarrow{a} \end{aligned} \tag{5}$$

where, $(1/\lambda_h)$ and $(1/\lambda_a)$ are eigenvales, and AA^T, $A^T A$ are eigenvectors.

The resulting computation thus takes the following form:

- Step 1. Assemble the target subset of web pages, form the graph induced by their hyperlinks and compute AA^T and $A^T A$.
- Step 2. Compute the principal eigenvectors of AA^T and $A^T A$ to form the vector of hub scores $\overrightarrow{h}$ and authority scores $\overrightarrow{a}$.
- Step 3. Output the top-scoring hubs and the top-scoring authorities.

It is possible that a small number of iterations of the power iteration method yields the relative ordering of the top hubs and authorities. Experiments have suggested that in practice, about five iterations of Equation (3) yield fairly good results.

Conclusions

- Link analysis for web search has intellectual antecedents in the field of citation analysis, aspects of which overlap with an area known as bibliometrics.
- Because the iterative updates captured the intuition of good hubs and good authorities, the high-scoring pages we output would give us good hubs and authorities from the target subset of web pages.
- Ng et al. (2001b) suggest that the PageRank score assignment is more robust than HITS in the sense that scores are less sensitive to small changes in a graph topology. However, it has also been noted that the teleport operation contributes significantly to PageRank's robustness in this sense.
- Ng et al. (2001b) introduce a notion of stability for link analysis, arguing that small changes to link topology should not lead to significant changes in the ranked list of results for a query.