5 Web Search

Outline:

1. *Page rank*, for discovering the most "important" pages on the Web, as used in Google.
2. *Hubs and authorities*, a more detailed evaluation of the importance of Web pages using a variant of the eigenvector calculation used for Page rank.

5.1 Page Rank

Intuitively, we solve the recursive definition of "importance": a page is important if important pages link to it.

Create a stochastic matrix of the Web; that is:

1. Each page i corresponds to row i and column i of the matrix.
2. If page j has n successors (links), then the ijth entry is $1/n$ if page i is one of these n successors of page j, and 0 otherwise.

The intuition behind this matrix is:

- Imagine that initially each page has one unit of importance. At each round, each page shares whatever importance it has among its successors, and receives new importance from its predecessors.
- Eventually, the importance of each page reaches a limit, which happens to be its component in the principal eigenvector of this matrix.
- That importance is also the probability that a Web surfer, starting at a random page, and following random links from each page will be at the page in question after a long series of links.

Example 5.1: In 1839, the Web consisted on only three pages — Netscape, Microsoft, and Amazon. The links among these pages were as shown in Fig. 9.

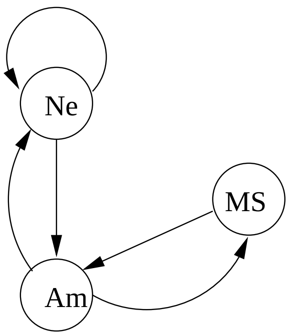

Figure 9: The Web in 1839

Let $[n, m, a]$ be the vector of importances for the three pages: Netscape, Microsoft, Amazon, in that order. Then the equation describing the asymptotic values of these three variables is:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

For example, the first column of the matrix reflects the fact that Netscape divides its importance between itself and Amazon. The second column indicates that Microsoft gives all its importance to Amazon.

We can solve equations like this one by starting with the assumption $n = m = a = 1$, and applying the matrix to the current estimate of these values repeatedly. The first four iterations give the following estimates:

$$\begin{array}{lclllll} n & = & 1 & 1 & 5/4 & 9/8 & 5/4 \\ m & = & 1 & 1/2 & 3/4 & 1/2 & 11/16 \\ a & = & 1 & 3/2 & 1 & 11/8 & 17/16 \end{array}$$

In the limit, the solution is $n = a = 6/5$; $m = 3/5$. That is, Netscape and Amazon each have the same importance, and twice the importance of Microsoft (well this was 1839). □

- Note that we can never get absolute values of n, m, and a, just their ratios, since the initial assumption that they were each 1 was arbitrary.
- Since the matrix is *stochastic* (sum of each column is 1), the above *relaxation* process converges to the principal eigenvector.

5.2 Problems With Real Web Graphs

1. *Dead ends*: a page that has no successors has nowhere to send its importance. Eventually, all importance will "leak out of" the Web.
2. *Spider traps*: a group of one or more pages that have no links out of the group will eventually accumulate all the importance of the Web.

Example 5.2: Suppose Microsoft tries to duck charges that it is a monopoly by removing all links from its site. The new Web is as shown in Fig. 10, and the matrix describing transitions is:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

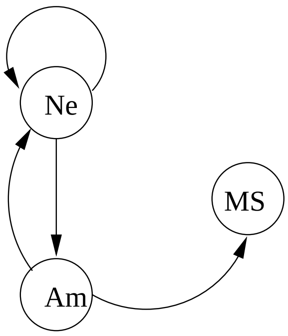

Figure 10: Microsoft becomes a dead end

The first four steps of the iterative solution are:

$$\begin{array}{lclllll} n & = & 1 & 1 & 3/4 & 5/8 & 1/2 \\ m & = & 1 & 1/2 & 1/4 & 1/4 & 3/16 \\ a & = & 1 & 1/2 & 1/2 & 3/8 & 5/16 \end{array}$$

Eventually, each of n, m, and a become 0; i.e., all the importance leaked out. □

Example 5.3: Angered by the decision, Microsoft decides it will link only to itself from now on. Now, Microsoft has become a spider trap. The new Web is in Fig. 11, and the equation to solve is:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix}$$

The first steps of the solution are:

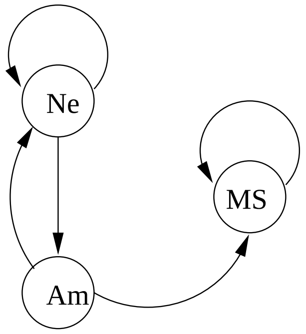

Figure 11: Microsoft becomes a spider trap

$$\begin{array}{lcllllll} n & = & 1 & 1 & 3/4 & 5/8 & 1/2 \\ m & = & 1 & 3/2 & 7/4 & 2 & 35/16 \\ a & = & 1 & 1/2 & 1/2 & 3/8 & 5/16 \end{array}$$

Now, m converges to 3, and $n = a = 0$. □

5.3 Google Solution to Dead Ends and Spider Traps

Instead of applying the matrix directly, "tax" each page some fraction of its current importance, and distribute the taxed importance equally among all pages.

Example 5.4: If we use a 20% tax, the equation of Example 5.3 becomes:

$$\begin{bmatrix} n \\ m \\ a \end{bmatrix} = 0.8 \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} n \\ m \\ a \end{bmatrix} + \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$$

The solution to this equation is $n = 7/11$; $m = 21/11$; $a = 5/11$.

- Note that the sum of the three values is not 3, but there is a more reasonable distribution of importance than in Example 5.3.

□

5.4 Google Anti-Spam Devices

"Spamming" is the attempt by many Web sites to appear to be about a subject that will attract surfers, without truly being about that subject.

- Google, like other search engines, tries to match the words in your query to the words on the Web pages. However, Google, unlike other engines tends to believe what others say about you in their anchor text, making it harder for you to *appear* to be about something you are not.

- The use of Page rank to measure importance, rather than the more naive "number of links into the page" also protects against spammers. The naive measure can be fooled by the spammer who creates 1000 pages that mutually link to one another, while Page rank recognizes that none of the pages have any real importance.

5.5 Hubs and Authorities

Intuitively, we define "hub" and "authority" in a mutually recursive way: a hub links to many authorities, and an authority is linked to by many hubs.

- Authorities turn out to be pages that offer information about a topic, e.g., the Quest home page about the IBM data-mining project.

- Hubs are pages that don't provide the information, but tell you where to find the information, e.g., the CS345 home page.

- Uses a matrix formulation similar to that of Page rank, but without the stochastic restriction. We count each link as 1, regardless of how many successors or predecessors a page has.

- Repeated application of the matrix leads to divergence, but we can introduce scaling factors and keep the computed values of "authority" and "hubbiness" for each page within finite bounds.

Define a matrix A whose rows and columns correspond to Web pages, with entry $A_{ij} = 1$ if page i links to page j, and 0 if not.

- Notice that A^T, the transpose of A, looks like the matrix used for computing Page rank, but A^T has 1's where the Page-rank matrix has fractions.

Let $\vec{a}$ and $\vec{h}$ be vectors, whose ith component corresponds to the degrees of authority and hubbiness of the ith page. let λ and μ be suitable scaling factors to be determined later. Then we can state:

1. $\vec{h} = \lambda A\vec{a}$. That is, the hubbiness of each page is the sum of the authorities of all the pages it links to, scaled by λ.

2. $\vec{a} = \mu A^T\vec{h}$. That is, the authority of each page is the sum of the hubbiness of all the pages that link to it, scaled by μ.

We can derive from (1) and (2), using simple substitution, two equations that relate vectors $\vec{a}$ and $\vec{h}$ only to themselves:

$$\vec{a} = \lambda\mu A^T A\vec{a}; \; \vec{h} = \lambda\mu AA^T\vec{h}$$

As a result, we can compute $\vec{h}$ and $\vec{a}$ by relaxation, giving us the principal eigenvectors of the matrices AA^T and A^TA, respectively.

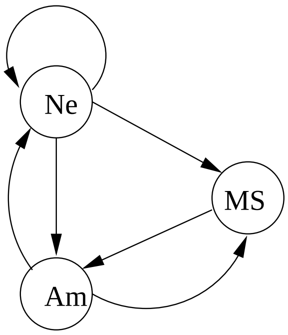

Figure 12: Web for Example 5.5

Example 5.5: Consider the Web of Fig. 12. The relevant matrices are:

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad A^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad AA^T = \begin{bmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{bmatrix} \quad A^TA = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

If we use $\lambda = \mu = 1$ and assume that the vectors $\vec{h} = [h_n, h_m, h_a]$ and $\vec{a} = [a_n, a_m, a_a]$ are each initially $[1, 1, 1]$, the first three iterations of the equations for $\vec{a}$ and $\vec{h}$ are:

$$
\begin{array}{lcllll}
a_n & = & 1 & 5 & 24 & 114 \\
a_m & = & 1 & 5 & 24 & 114 \\
a_a & = & 1 & 4 & 18 & 84 \\
\\
h_n & = & 1 & 6 & 28 & 132 \\
h_m & = & 1 & 2 & 8 & 36 \\
h_a & = & 1 & 4 & 20 & 96
\end{array}
$$

For instance, the vector $\vec{a}$, properly scaled, will converge to a vector where $a_n = a_m$, and each of these is greater than a_a in the ratio $1 + \sqrt{3}\ :\ 2$, or about 1.36. $\square$