CS 493: Algorithms for Massive Data Sets	Hubs and Authorities
DATE Tuesday, 4/23/2002	Scribe: Keith Vallerio

1 Page Rank

From Larry "Page". The basic idea is that you start with a value for R_0 and apply iteratively. Hopefully, it will converge. Those ranks will be the pagerank The result is an independent quality assessment of the page.

$$R(n) = c(\sum_{v_i,(v,u)\in E} \frac{R(v)}{N_v} + E(u))$$

2 Hubs and Aurhorities

Based on work by Kleinberg. The basic idea is to take a query and get a basic starting set. From there, expand the set using "points to it" and "it points to" link info from the pages. The idea is that authoritative pages are pointed to by many good hubs. This is accomplished by starting with the basic pages, performing I and O operations, and iterating:

a(p): authority score ("I" operation)

$$a(p) = \sum_{q_i,(q,p)\in E} h(q)$$

h(p): hub score ("O" operation)

$$h(p) = \sum_{q_i,(q,p)\in E} a(q)$$

3 Adjacency Matrix

The adjacnecy matrix for determining hubs and authorities is based on the following:

$$A_{ij} = \begin{cases} 1 & i \rightarrow j \\ 0 & otherwise \end{cases}$$

- a_i: a in the i^{th} iteration
- h_i: h in the i^{th} iteration
- $a_{i+1} = A^T h_o$
- $a_i = (A^T A)^{i-1} A^T h_o$
- $a_i = (A^T A)^{i-1} A^T h_o$

4 Page Rank

$$R = C\left(\overbrace{\left[\begin{array}{cc|c} & & v \\ & & \\ u & & \\ \hline & & \end{array}\right]}^{A}\overbrace{\left[\begin{array}{c} \\ \\ \\ \\ \end{array}\right]}^{R} + \overbrace{\left[\begin{array}{c} \\ \\ \\ \\ \end{array}\right]}^{E}\right)$$

- $R = C(AR + E)$
- $A_{u,v} = \frac{1}{N}$, if $(v, u) \in E$
- Assume $\|R\| = 1$
- $R = C(A + E \times 1_n)R$
  (Since $E \times 1_n \times R = E$)
- $R = C(A + E \times 1_n)R$
- Page values = eigenvalues of A

5 Linear Algebra

- $A_{n \times n}x = \lambda x$
- n eigenvalues (λ_i) and n corresponding eigenvectors (x_i) such that $Ax_i = \lambda x_i$
- Assume $|\lambda_1|$ has the greatest eigenvalue, then this method will return x_1
- Look at x_1 to x_n (All orthonormal)
- Any x can be expressed $\sum \alpha_i x_i$
- $Ax = A \sum \alpha_i x_i = \sum \alpha_i A x_i = \sum \alpha_i \lambda_i x_i$
- So, $A^j x = \sum \alpha_i \lambda_i^j x_i$
- a_i converges to a^* (a^*: dominant eigenvector of $A^T A$)
- h_i converges to h^* (h^*: dominant eigenvector of AA^T)
- $A^T A_i j$: nodes that point to both i and j
- $AA_i^T j$: nodes pointed to by both i and j

The problem with this technique is that a query for "jaguar" could be the car, the animal, or the team and leads to only finding one of the "communities".

6 Singular Value Decomposition (SVD)

$$A = U\Sigma V^T = \left[\begin{array}{c|c} & u \\ & \\ & \\ & \end{array}\right] \left[\begin{array}{cccc} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_d \end{array}\right] \left[\begin{array}{c} \\ v \\ \hline \\ \end{array}\right]$$

- $A_{m\times n}, U_{m\times d}, \Sigma_{d\times d}, V_{n\times d}$
- A is a real, positive symetric matrix so all λ are positive
- The columns of U and the rows of V are orthogonal (i.e. $U^TU = I$ and $VV^T = I$)
- $A = \sum \sigma_i u_i v_i^T$
- $AA^T = (U\Sigma V^T)(V\Sigma U^T) = U\Sigma^2 U^T = \sum \sigma_i^2 u_i u_i^T$
- σ_i are the eigenvalues of AA^T
- u_i^T are the eigenvectors of AA^T
- $AA^T u_i = \sigma_i^2 u_i$
- $A^TA = (V\Sigma U^T)(U\Sigma V^T) = V\Sigma^2 V^T = \sum \sigma_i^2 v_i v_i^T$
- σ_i are the eigenvalues of A^TA
- u_i^T are the eigenvectors of A^TA

7 Problems with Link Analysis Methods

- Hubs and Authorities can suffer from "topic drift." This is caused when in and out links are more closely related to each other than the original results.
- Lempel and Moran found that a search for "jaguar" revealed a lot of pages about Cincinatti since it drew may pages from the paper by that name. As a result, the suggested a different methodology "SALSA."

7.1 SALSA

Start with a root set. From a page, perform a combination of the following operations:

1. Goto a page it points to
2. Goto a page that points to it

A hub is computed by sequences of (1)(2), (1)(2), ... and an authority is computed by (2)(1), (2)(1) ... looking at stationary distribution.

7.1.1 Random Walk Probabilities

$$hub:\begin{array}{ccc} B(i) & & N(i) \\ \searrow & & \nearrow \\ \rightarrow & i & \rightarrow \\ \nearrow & & \searrow \end{array}$$

$$(i,j) = \Sigma_K \frac{1}{|N(i)|}\frac{1}{B(K)}$$

$$hub = \frac{|N(i)|}{\Sigma|N(i)}$$

$$authority = \frac{|B(i)|}{\Sigma|B(i)|}$$