Exploiting Link Structure

Problem

How do I find on the WWW good sources of information?

eXtreme Theory

Background

1990: Alan Emtage at McGill University creates 'Archie' ftp search tool
1991: Mark McCahill at University of Minnesota introduces 'Gopher' as alternative to Archie
Early 90s: WWW born, grows very quickly
1993: Matthew Gray creates World Wide Web Wanderer at MIT
1993: Martijn Koster creates Archie-Like Indexing of the Web (ALIWEB)
1994: Brian Pinkerton introduces WebCrawler
1994: Carnegie Mellon launches Lycos search engine with directory of 54,000 documents
Mid 90s: Web crawlers-based search engines overwhelmed by bad quality results
1994: David Filo & Jerry Yang at Stanford University start Yahoo!
Late 90s: Web continues to grow. Human-based indexing cannot keep pace
Finding Authoritative Sources on the WWW

Jon Kleinberg

Queries

Specific: Will Blue Choice cover my nose ring?

Broad: What does my health insurance cover?

Similar: Find all pages similar to bluecross.com
Hubs

webpages that have a lot of outlinks

http://dir.yahoo.com/Arts/Performing_Arts/Circus_Arts/Circuses/

Authorities

webpages that have a lot of inlinks

http://www.yahoo.com
http://www.rit.edu
http://www.gnu.org

Reinforcement

Measuring “hubiness” (and “authoritativeness”)

http://www.yahoo.com
http://www.rit.edu
http://www.gnu.org
**Notation**

$S_\alpha$ is the set of all pages under consideration

For each $p \in S_\alpha$,

$x^p$ is the relative authority weight of $p$

$$\sum_{p \in S_\alpha} x^p = 1$$

$y^p$ is relative hub weight of $p$

$$\sum_{p \in S_\alpha} y^p = 1$$

**Reinforcement**

Authority weight:

$$x^p = \sum_{q \in S_\alpha : q \rightarrow p} y^q$$

Hub weight:

$$y^p = \sum_{q \in S_\alpha : p \rightarrow q} x^q$$
**Computation**

Assume $n$ pages.

Let $x_0$ estimate $(x^{<1>}, \ldots, x^{<n>})$ at time $0$

Let $y_0$ estimate $(y^{<1>}, \ldots, y^{<n>})$ at time $0$

Let $x_0 = y_0 = (1/\sqrt{n}, \ldots, 1/\sqrt{n}) \in \mathbb{R}^n$

**Refining hub estimate**

Let

$$y_i = (\mathcal{O}(i, 1), \ldots, \mathcal{O}(i, n)) / \sqrt{\mathcal{O}(i, 1) + \ldots + \mathcal{O}(i, n)}$$

where

$$\mathcal{O}(j, p) = \sum_{q \in S: p \rightarrow q} x_j^{<q>}$$

**Refining authority estimate**

Let

$$x_i = (\mathcal{A}(i-1, 1), \ldots, \mathcal{A}(i-1, n)) / \sqrt{\mathcal{A}(i-1, 1) + \ldots + \mathcal{A}(i-1, n)}$$

where

$$\mathcal{A}(j, p) = \sum_{q \in S: q \rightarrow p} y_j^{<q>}$$

**How many iterations?**

Do

$$y_i = (\mathcal{O}(i, 1), \ldots, \mathcal{O}(i, n)) / \sqrt{\mathcal{O}(i, 1) + \ldots + \mathcal{O}(i, n)}$$

and

$$x_i = (\mathcal{A}(i-1, 1), \ldots, \mathcal{A}(i-1, n)) / \sqrt{\mathcal{A}(i-1, 1) + \ldots + \mathcal{A}(i-1, n)}$$

even converge?
How many iterations?

**Theorem 1:**

As \( k \to \infty \),

\[
x_k \uparrow x^*
\]

\[
y_k \uparrow y^*
\]

Where \( x^* \) and \( y^* \) are very special vectors (as we shall see)

Proof. Start with adjacency matrix \( A \) of \( S_\sigma \)

\[
\begin{array}{cccccccccc}
\text{a} & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
\text{b} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{c} & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\text{d} & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
\text{e} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{f} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{g} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{h} & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
\text{i} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\end{array}
\]

Let \( A^T \) be the transpose of \( A \), i.e., if \( B = A^T \) then \( b_{ij} = a_{ji} \)

\[
A^T
\]

\[
\begin{array}{cccccccccc}
\text{a} & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\text{b} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\text{c} & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\text{d} & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
\text{e} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{f} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{g} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{h} & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
\text{i} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\end{array}
\]

\[\mathcal{I}(i-1, p) = " q# S_\sigma : q$ p y_{i-1} <q> = (A^T y_{i-1}) <p>\]

\[\mathcal{R}(i-1, p) = \sum_{q \in S_\sigma : q \rightarrow p} y_{i-1} q \phi = (A^T y_{i-1}) <\phi>\]

\[
\begin{array}{cccccccccc}
\text{a} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{b} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{c} & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\text{d} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{e} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{f} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\text{g} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{h} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{i} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]
How many iterations?

\[ \mathcal{R}(i-1, p) = \sum_{q \in S \sigma: q \rightarrow p} y_{i-1}^{\varphi q} = (A^T y_{i-1})^{\varphi q} \]

\[ A^T \]

\[ y_0 \]

\[ a \ b \ c \ d \ e \ f \ g \ h \]

\[ a \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ b \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ c \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ d \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ e \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ f \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1/n \]
\[ g \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1/n \]
\[ h \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ i \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]

How many iterations?

\[ \mathcal{O}(i, p) = \sum_{q \in S \sigma: q \rightarrow p} x_i^{\varphi q} = (Ax_i)^{\varphi q} \]

\[ A \]

\[ x_0 \]

\[ a \ b \ c \ d \ e \ f \ g \ h \ i \]

\[ a \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ b \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ c \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ d \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1/n \]
\[ e \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ f \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ g \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1/n \]
\[ h \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]
\[ i \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1/n \]

\[ A^T, AA^T \] are both symmetric, i.e., \( a_{ij} = a_{ji} \)
How many iterations?

$A^TA, AA^T$ are both symmetric, i.e., $a_{ij} = a_{ji}$

<table>
<thead>
<tr>
<th>$A$</th>
<th>$A^T$</th>
<th>$AA^T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 1 0 0 0 0 0</td>
<td>0 1 0 0 0 0 0 0 0</td>
<td>2 0 0 0 0 0 1 1</td>
</tr>
<tr>
<td>1 0 0 0 0 0 0 0 0</td>
<td>0 1 0 0 0 0 0 0 0</td>
<td>0 1 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>0 1 0 0 1 0 0 0 0</td>
<td>1 0 0 0 0 0 0 0 0</td>
<td>0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>0 0 0 0 1 0 0 0 0</td>
<td>0 0 1 0 0 0 0 0 0</td>
<td>0 0 1 0 0 0 1 0</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 0 0</td>
<td>0 0 0 1 0 0 0 0 0</td>
<td>0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 0 0</td>
<td>0 0 0 0 1 0 0 0 0</td>
<td>0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 0 0</td>
<td>0 0 0 0 0 0 1 0 0</td>
<td>0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>0 0 0 0 0 1 1 0 0</td>
<td>0 0 0 0 0 0 0 0 0</td>
<td>1 0 0 1 0 0 2 1</td>
</tr>
<tr>
<td>0 0 0 0 0 1 0 0 0</td>
<td>0 0 0 0 0 0 0 0 0</td>
<td>1 0 0 0 0 0 1 1</td>
</tr>
</tbody>
</table>

How many iterations?

If $X = BC$ then

$x_{ij} = \sum_{k \in \{1, \ldots, n\}} b_{ik} c_{kj}$

If $X = AA^T$, $B = A$, $C = A^T$ then $b_{ij} = a_{ij}$ but $c_{ij} = a_{ji}$

$x_{ij} = \sum_{k \in \{1, \ldots, n\}} b_{ik} c_{kj} = \sum_{k \in \{1, \ldots, n\}} a_{ik} a_{kj} = \sum_{k \in \{1, \ldots, n\}} c_{kj} b_{kj} = x_{ji}$

How many iterations?

If $M$ is an $n \times n$ matrix.

A number $\lambda$ is an eigenvalue if, for some vector $\omega$ (with $|\omega| = 1$),

$M \omega = \lambda \omega$

$M$ has, at most, $n$ distinct eigenvalues ($\lambda_1(M) > \ldots > \lambda_n(M)$) and corresponding eigenvectors ($\omega_1(M) > \ldots > \omega_n(M)$)

How many iterations?

Lemma 1:

If $M$ is a symmetric matrix and $v$ is a vector such that $v^T \omega_1(M) \neq 0$ then $M^k v \uparrow \omega_1(M)$ as $k \uparrow \infty$
Lemma 2:

If $M$ is a symmetric matrix and has only nonnegative entries then $\omega_1(M)$ has only nonnegative entries.

Since $AA^T$ is symmetric and has only nonnegative entries, $\omega_1(AA^T)$ has only nonnegative entries.

Lemma 1:

If $M$ is a symmetric matrix and $v$ is a vector such that $v^T \omega_j(M) \neq 0$ then $M^k v \uparrow \omega_j(M)$ as $k \uparrow \infty$.

Because $AA^T$ is symmetric and $y_0^T \omega_j(AA^T) \neq 0$, it follows that $(AA^T)^k y_0 \uparrow \omega_j(AA^T)$ as $k \uparrow \infty$.

Similarly, $(A^T A)^k y_0 \uparrow \omega_j(A^T A)$ as $k \uparrow \infty$. 

How many iterations?
How many iterations?

Theorem 1 (restated):

As $k \to \infty$,

$$x_k \to \lambda_1(M^TM)$$
$$y_k \to \lambda_1(MM^T)$$

Thus, $x_k$ (respectively, $y_k$) approaches an eigenvector of the first eigenvalue of $M^TM$ (respectively, $MM^T$)

Homework

Prove lemmas 1 and 2 from the notes.

Submit this homework by 4/22