Data Mining and Information Retrieval

PageRank and Web Spam
Ranking Web Pages

- Web pages are not equally “important”
  - www.joe-schmoe.com vs www.stanford.edu

- Inlinks as votes
  - www.stanford.edu has 23,400 inlinks
  - www.joe-schmoe.com has 1 inlink

- Are all inlinks equal?
  - Recursive question!
Simple Recursive Formulation

- Each link’s vote is proportional to the importance of its source page.
- If page $P$ with importance $x$ has $n$ outlinks, each link gets $x/n$ votes.
- Page $P$’s own importance is the sum of the votes on its inlinks.
Simple “Flow” Model

The web in 1839

\[ y = y/2 + a/2 \]
\[ a = y/2 + m \]
\[ m = a/2 \]
Solving the Flow Equations

- 3 equations, 3 unknowns, no constants
  - No unique solution
  - All solutions equivalent modulo scale factor
- Additional constraint forces uniqueness
  - $y + a + m = 1$
  - $y = 2/5$, $a = 2/5$, $m = 1/5$
- Gaussian elimination method works for small examples, but we need a better method for large graphs
Matrix formulation

- Matrix $M$ has one row and one column for each web page.
- Suppose page $j$ has $n$ outlinks:
  - If $j \rightarrow i$, then $M_{ij} = 1/n$
  - Else $M_{ij} = 0$
- $M$ is a column stochastic matrix:
  - Columns sum to 1
- Suppose $r$ is a vector with one entry per web page:
  - $r_i$ is the importance score of page $i$
  - Call it the rank vector
Example

Suppose page $j$ links to 3 pages, including $i$.

\[
M = \begin{bmatrix}
    & j \\
  i & & & \\
    & & M & \\
  r & & & r
\end{bmatrix}
\]

\[
1/3 = i
\]
The flow equations can be written

\[ r = Mr \]

So the rank vector is an eigenvector of the stochastic web matrix

In fact, its first or principal eigenvector, with corresponding eigenvalue 1
Example

\[
y = \frac{y}{2} + \frac{a}{2}
\]

\[
a = \frac{y}{2} + m
\]

\[
m = \frac{a}{2}
\]

\[
\begin{bmatrix}
y \\
a \\
m
\end{bmatrix} = \frac{1}{2} \begin{bmatrix}
y \\
a \\
m
\end{bmatrix} + \begin{bmatrix}
1/2 & 1/2 & 0 \\
1/2 & 0 & 1 \\
0 & 1/2 & 0
\end{bmatrix} \begin{bmatrix}
y \\
a \\
m
\end{bmatrix}
\]

r = Mr
Power Iteration method

- Simple iterative scheme
- Suppose there are N web pages
- Initialize: \( r^0 = [1/N, \ldots, 1/N]^T \)
- Iterate: \( r^{k+1} = Mr^k \)
- Stop when \( |r^{k+1} - r^k|_1 < \varepsilon \)

\(|x|_1 = \sum_{1 \leq i \leq N} |x_i|\) is the L1 norm

Can use any other vector norm
Power Iteration Example

Yahoo

Amazon

M’soft

\[
\begin{bmatrix}
y & a & m \\
1/2 & 1/2 & 0 \\
1/2 & 0 & 1 \\
0 & 1/2 & 0 \\
\end{bmatrix}
\]

\[
\begin{align*}
y &= 1/3 & 1/3 & 5/12 & 3/8 & 2/5 \\
a &= 1/3 & 1/2 & 1/3 & 11/24 & \ldots & 2/5 \\
m &= 1/3 & 1/6 & 1/4 & 1/6 & 1/5 \\
\end{align*}
\]
Random Walk Interpretation

- Imagine a **random web surfer**
  - At any time $t$, surfer is on some page $P$
  - At time $t+1$, the surfer follows an outlink from $P$ uniformly at random
  - Ends up on some page $Q$ linked from $P$
  - Process repeats indefinitely

- Related to *Markov Chain* model
PageRank

A page is important if many other important pages point to it.

\[ P_0 = d \sum_i P_i / \text{out}(i) + (1-d) \]

- **PageRank Score of page** \( p_0 \)
- **PageRank Score of page** \( p_i \) that points to page \( p_0 \)
- **Damping factor** \( d \approx 0.85 \)
- **Out degree of page** \( p_i \)
- **Random jump probability** \( 0.15 \approx 0.15 \)
What is Web Spam?
Increasing exposure on the World Wide Web may yield significant financial gains for the Web site owners!

The increasing importance of search engines to commercial Web sites has given rise to a phenomenon called “Web Spam”!
Why Web Spam

- E-commerce is rapidly growing
  - Projected to $329 billion by 2010
- More traffic $\rightarrow$ more money
- Large fraction of traffic from Search Engines
- Increase Search Engine referrals:
  - Place ads $\smiley$
  - Provide genuinely better content $\smiley$
  - Create Web spam… $\frown$
Web Spam Examples
(you know it when you see it)
Defining Web Spam

- Spam Web page
  - A page created for the sole purpose of attracting search engine referrals (to this page or some other "target" page)

- Ultimately a judgment call
  - Some Web pages are borderline cases
Why Web Spam is Bad

- **Bad for users**
  - Makes it harder to satisfy information need
  - Leads to frustrating search experience

- **Bad for search engines**
  - Wastes bandwidth, CPU cycles, storage space
  - Pollutes corpus (infinite number of spam pages!)
  - Distorts ranking of results
Detecting Web Spam

- Spam detection: A classification problem
  - Given salient features of a Web page, decide whether the page is spam

- Which “salient features”?  
  - Need to understand spamming techniques to decide on features  
  - Finding right features is “alchemy”, not science
Ask yourself a question:
- Why Web spam exists?
- Spammers did, because they are trying to mislead Web search engines

Thus, in order to detect Web spam
- Thinking in the spammers’ way
- If I am a spammer, what shall I do to mislead the search engines as much as possible?

So, before going to detect Web spam
- Try to understand how a search engine ranks Web pages...
Web Spam Taxonomy

Web Spam = misleading search engines to obtain higher-than-deserved ranking

Given a URL, spam Web servers return different HTML document to different Web crawler

Automatically redirecting the browser to another URL as soon as the page is loaded

Information retrieval and text query

Web Spam Techniques

Boosting

Term

Link

Hiding

Content Hiding

Cloaking

Redirection

HITS, PageRank
How to Detect Web Spam

Ask yourself following questions
- What kind of features can be useful to detect spam Web pages?
- Once we get those features, what kind of data mining methods can be used to detect spam Web pages?
- Once we have Web spam detection methods, what kind of evaluation metrics can be used to evaluate the results?
Reading References

Search Engine Webmaster Guidelines

- Google
  - http://www.google.com/support/webmasters/bin/answer.py?answer=35769

- Yahoo!

- Microsoft Live Search