Suppose we have an Inverted Index for a set of webpages. 

**Disclaimer**
- Not really the scenario of Lecture 11.
- Indexing for the web is massive-scale:
  - many distributed networks working in parallel.

We search with a term $t$.

Index has many hits for $t$ (say 36,000 for this $t$).

How should we rank them?

---

**A real search**

**Ranking Queries**

Inverted Index (probably) stores the frequency of the term $t$ in each document $d$ (e.g., in previous lecture, our index contains $f_{d,t}$ values).

**Idea**
- Rank answers to queries in order of frequency of $t$ in the various webpages.

**Problem**
- Some great websites will not even contain the term $t$.
- For example, there are not many occurrences of the term “University of Edinburgh" on http://www.ed.ac.uk

**New Idea**
- Use structure of web to rank queries.
Ranking Queries using web structure

Principle:
Link from one webpage to another confers authority on the target webpage.

This is the concept behind:
- The Hub-Authority model of Kleinberg.
- PageRank™ ranking system of Google™.
  In early 90s, while PhD students at Stanford, Sergey Brin and Larry Page invented PageRank™ (and founded Google™).

PageRank™

Could use in-degree to measure ranking directly.
But:
- Want pages of high rank to confer more authority on the pages they link to.
- A page with few links should transfer more of its authority to its linked pages than one with many links.

Assumptions: (for basic PageRank™)
- No “dead-end” pages.
- Every page can hop to every other page via links.
- Aperiodic.

PageRank™

Webgraph for a particular query:
- vertices $V = \{1, 2, \ldots, N\}$ corresponding to pages;
- links are the directed edges of the graph, so $E \subseteq [N] \times [N]$.

Let $G = (V, E)$. Recall:

Definition
Let $u$ denote some page $u \in [N]$ in the webgraph.
- $\text{in}(u)$ is the set of in-edges to $u$. The in-degree $\text{in}(u)$ is $\text{in}(u) = |\text{in}(u)|$.
- $\text{Out}(u)$ is the set of out-edges from $u$. The out-degree $\text{out}(u)$ is $\text{out}(u) = |\text{Out}(u)|$.

Principle of PageRank™

Let $R(v)$ denote the rank of $v$ for any webpage $v \in [N]$.
For every webpage $u$ in our collection, the following equality should hold:
$$ R(u) = \sum_{v \in \text{in}(u)} R(v) / \text{out}(v) $$

Rank of $u$ is the “total amount of Rank” given from the incoming links to $u$. 
PageRank™ in matrix form

\[
(R_1, R_2, \ldots, R_N) = (R_1, R_2, \ldots, R_N) \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1N} \\
p_{21} & p_{22} & \cdots & p_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
p_{N1} & p_{N2} & \cdots & p_{NN}
\end{pmatrix}
\]

where

\[
p_{uv} = \begin{cases} 
1/\text{out}(u), & \text{if } v \in \text{Out}(u); \\
0, & \text{otherwise.}
\end{cases}
\]

PageRank™

Questions and Answers

- How do we know that 1 is an eigenvalue of the matrix \( P^T \)?
  
  Answer: \( P^T \) is a stochastic matrix (each column adds to 1), so has eigenvalue 1.

- If 1 is an eigenvalue of \( P^T \), is it guaranteed to be a simple eigenvalue?
  
  Answer: Under our assumptions, there is just one linearly independent eigenvector for 1.

PageRank™ in matrix form

Shorthand version:

\[
R^T = R^T P,
\]

where \( P = [p_{uv}]_{u,v \in [N]} \) and \( R \) is the vector of ranks for \([N]\).

Equivalent to asking for

\[
R = P^T R,
\]

Looks like condition for \( R \) to be an eigenvector of \( P^T \) with eigenvalue \( \lambda = 1 \).

Example

Example webgraph returned by a rare query in ancient times.
Satisfies all the nice conditions for Basic PageRank™ model (no dead-end pages, can move from any vertex $x$ to any other vertex $y$, aperiodic).

\begin{align*}
(R_u, R_v, R_w, R_z) &= (R_u, R_v, R_w, R_z) \\
&= \begin{pmatrix}
0 & \frac{1}{2} & 0 & \frac{1}{2} \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0
\end{pmatrix}.
\end{align*}

Can “read-off” $R_w = R_z/3$, and propagate this into matrix:

\begin{align*}
(R_u, R_v, R_w, R_z) &= (R_u, R_v, R_w, R_z) \\
&= \begin{pmatrix}
0 & \frac{1}{2} & 0 & \frac{1}{2} \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0
\end{pmatrix}.
\end{align*}

Now remove $R_w$ (keeping $R_w = R_z/3$ to side):

\begin{align*}
(R_u, R_v) &= (R_u, R_v) \\
&= \begin{pmatrix}
0 & \frac{1}{2} & 0 & \frac{1}{2} \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0
\end{pmatrix}.
\end{align*}
Example (continued)

\[
(R_u, R_v, R_z) = (R_u, R_v, R_z) \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \iff \\
(R_u, R_v - R_z, R_z) = (R_u, R_v, R_z) \begin{pmatrix} 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & -\frac{1}{2} & 0 \end{pmatrix}.
\]

Middle equation reads \(R_v - R_z = \frac{1}{2}(R_z - R_u)\), so \(R_v = R_z\).

Final equation says \(R_z = 1/2(R_u + R_v)\), so \(R_z = R_u\) too.

Solution: \(R_u = R_v = R_z\), \(R_w = R_z/3\).

Example (continued)

Alternative (Equivalent) Approach)

Expand vector-matrix product:

\[
\begin{align*}
R_u &= \frac{1}{2} R_v + \frac{1}{2} R_w + \frac{1}{3} R_z \\
R_v &= \frac{1}{2} R_u + \frac{1}{2} R_w + \frac{1}{3} R_z \\
R_w &= \frac{1}{3} R_z \\
R_z &= \frac{1}{2} R_u + \frac{1}{2} R_v.
\end{align*}
\]

\(\triangleright\) Subtract the second equation from the first: \(R_u - R_v = \frac{1}{2} R_v - \frac{1}{2} R_u\)

\(\triangleright\) It follows that \(R_v = R_u\).

\(\triangleright\) Substituting into the fourth equation: \(R_z = R_u\).

\(\triangleright\) This method is probably preferable for such small examples.

Example (continued)

\[
\begin{align*}
\text{Solutions are } & R_u = R_v = R_z, \ R_w = R_z/3, \text{i.e.,} \\
(R_u, R_v, R_w, R_z) &= c(1,1,1/3,1)
\end{align*}
\]

where \(c\) is a constant.

Not the same as counting in-degree (for this example).

General PageMark™ model

\(\triangleright\) Remove all our assumptions (dead-end pages, connectivity).

\(\triangleright\) \(\lambda\) cannot be assumed to be 1.

\(\triangleright\) Need to tinker the model. See Lecture Notes.
Further Reading

Nothing in [GT] or [CLRS].

Papers on the web:

  http://www-db.stanford.edu/backrub/google.html

- The PageRank Citation Ranking: Bringing Order to the Web, by Page, Brin, Motwani and Winograd, 1998. Available online from:

- Authoritative Sources in a Hyperlinked Environment, by Jon Kleinberg. Available Online from Jon Kleinberg's webpage:
  http://www.cs.cornell.edu/home/kleinber/