Inf 2B: Ranking Queries on the WWW

Kyriakos Kalorkoti

School of Informatics
University of Edinburgh
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

God - Wikipedia, the free encyclopedia
This article is about the term "God" in the context of monotheism and henotheism. See deity or god (male deity) for details on polytheistic usages. ...
en.wikipedia.org/wiki/God - 154k - Cached - Similar pages

God.com
There are over six billion people in this world and each person has his or her own thoughts about God. How can a person know for sure what He is really like ...
www.god.com/ - 6k - Cached - Similar pages

View The interview With God Movie
Inspirational screensaver with landscape images, peaceful music and uplifting messages.
www.theinterviewwithgod.com/ - 6k - Cached - Similar pages

Video results for God

The Root of All Evil? - The God Delusion (1 of 2)
48 min
video.google.com

Martina McBride - God's Will
4 min 48 sec
www.youtube.com

CATHOLIC ENCYCLOPEDIA: God
How can you know that God exists? What is He really like? What is His relationship to the universe? What is the Blessed Trinity?
www.newadvent.org/cathen/08608a.htm - 9k - Cached - Similar pages

Does God Exist?
Bimonthly online journal that provides scientific evidence for God's existence. Site also includes lecture schedules, children's story, ...
www.docsgodexist.org/ - 34k - Cached - Similar pages

GOD. Global Online Directory.
The entry page to the Global online directory. ... G.O.D is coming... click here to advertise.
www.god.co.uk/ - 2k - Cached - Similar pages

Welcome to GOD.TV
With a vision to win one billion souls for the Kingdom of God, GOD TV is constantly exploring increased distribution opportunities to broadcast the Gospel ...
www.god.tv/ - 5k - Cached - Similar pages

god - process and task monitoring done right
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™).
Webgraph for a particular query:

- vertices $V = [N]$ where $[N] = \{1, 2, \ldots, N\}$ 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.

- $In(u)$ is the set of in-edges to $u$. The in-degree $in(u)$ is $in(u) = |In(u)|$.
- $Out(u)$ is the set of out-edges from $u$. The out-degree $out(u)$ is $out(u) = |Out(u)|$. 

---

**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.
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:

$$R(u) = \sum_{v \in \text{ln}(u)} \frac{R(v)}{\text{out}(v)}$$

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

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

Shorthand version:

\[ R^T = R^T P, \]  \hspace{1cm} (1)

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

Equivalent to asking for

\[ R = P^T R, \]  \hspace{1cm} (2)

Looks like condition for \( R \) to be an eigenvector of \( P^T \) with eigenvalue \( \lambda = 1 \).
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?
  
  - i.e., any two vectors that satisfy $P^T R = R$ are the same up to a non-zero constant multiple (linearly dependent).
  
  **Answer:** Under our assumptions, there is just one linearly independent eigenvector for 1.
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).
Example

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

\[(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} \).

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

\[(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} \\ 0 & 0 & 0 & 0 \\ \frac{1}{3} + \frac{1}{6} & \frac{1}{3} + \frac{1}{6} & \frac{1}{3} & 0 \end{pmatrix} \).
Example (continued)

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

$$(R_u, R_v, R_z) = (R_u, R_v, R_z) \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$
Example (continued)

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

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

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

Expand vector-matrix product:

\[ R_u = \frac{1}{2} R_v + \frac{1}{2} R_w + \frac{1}{3} R_z \]
\[ R_v = \frac{1}{2} R_u + \frac{1}{2} R_w + \frac{1}{3} R_z \]
\[ R_w = \frac{1}{3} R_z \]
\[ R_z = \frac{1}{2} R_u + \frac{1}{2} R_v. \]

- Subtract the second equation from the first:
  \[ R_u - R_v = \frac{1}{2} R_v - \frac{1}{2} R_u \]
- It follows that \( R_v = R_u \).
- Substituting into the fourth equation: \( R_z = R_u \).
- This method is probably preferable for such small examples.
Solutions are $R_u = R_v = R_z$, $R_w = R_z/3$, i.e.,

$$(R_u, R_v, R_w, R_z) = c(1, 1, 1/3, 1)$$

where $c$ is a constant.

Not the same as counting in-degree (for this example).
General PageMark™ model

- Remove all our assumptions (dead-end pages, connectivity).
- $\lambda$ cannot be assumed to be 1.
- Need to tinker the model. See Lecture Notes.
Further Reading

Nothing in [GT] or [CLRS].

Papers on the web:


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