

Inf 2B: Ranking Queries on the WWW

Kyriakos Kalorkoti

School of Informatics
University of Edinburgh

Queries

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 - Google Search

http://www.google.com/search?client=safari&rls=en-us&q=God&ie=UTF-8&oe=UTF-8 - Q- God

Apple Google Maps News (2718) YouTube Popular Yahoo Gáidhlig Greek University Goodies Photography Research

Web Images Maps News Video Mail more Sign in

Google God Search [Advanced Search](#) [Preferences](#)

Web Video Images Results 1 - 10 of about 470,000,000 for **God** [definition]. (0.24 seconds)

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/06608a.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.doesgodexist.org/ - 34k - [Cached](#) - [Similar pages](#)

GOD, Global Oline 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

Sponsored Links

Who and What is God?
Why is **God** a mystery to mankind?
Astonishing answers in free book.
www.TheTrumpet.com/mysteryofthegods

Looking for God?
Want to know where **God** is?
Need to find him? look here
www.lookingforgod.com

+pictures +of +god.
Get +pictures +of +**god**
Find +pictures +of +**god** with Ask.
www.Ask.com

God -total Union with God
Why Jesus is the Only Way to Union?
The Answer May Surprise You.
www.hallworthington.com

Ask God Into Your Heart
Jesus Christ Loves You. Here is a Prayer That Can Change Your Life
www.Jesus2020.com

Last Chaos Fantasy MMORPG
Medieval Fantasy, Endless Adventure
Join the Biggest Gaming Community!
lastchaos.aeriagames.com

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™

Webgraph for a particular query:

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

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:

$$R(u) = \sum_{v \in In(u)} R(v) / out(v)$$

Rank of u is the “total amount of Rank” given from the incoming links to u .

PageRank™ in matrix form

$$(R_1, R_2, \dots, R_N) = (R_1, R_2, \dots, R_N) \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1N} \\ p_{21} & p_{22} & \dots & p_{2N} \\ \dots & \dots & \dots & \dots \\ p_{N1} & p_{N2} & \dots & p_{NN} \end{pmatrix}$$

where

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

PageRank™ in matrix form

Shorthand version:

$$R^T = R^T P, \quad (1)$$

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

Equivalent to asking for

$$R = P^T R, \quad (2)$$

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

Questions and Answers

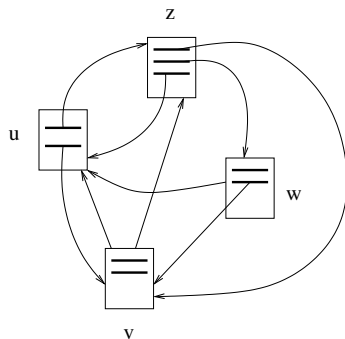
- ▶ How do we know that 1 is an eigenvalue of the matrix P^T ?

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*).

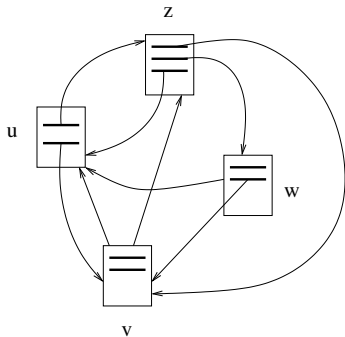
Answer: Under our assumptions, there is just one **linearly independent** eigenvector for 1.

Example



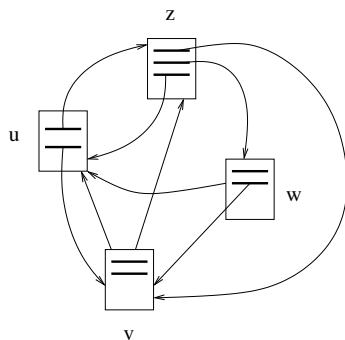
Example webgraph returned by a **rare query** in **ancient times**.

Example



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):

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

Example (continued)

$$(R_U, R_V, R_Z) = (R_U, R_V, R_Z) \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \iff$$

$$(R_U, R_V - R_Z, R_Z) = (R_U, R_V, R_Z) \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

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:

$$R_U = \frac{1}{2}R_V + \frac{1}{2}R_W + \frac{1}{3}R_Z$$

$$R_V = \frac{1}{2}R_U + \frac{1}{2}R_W + \frac{1}{3}R_Z$$

$$R_W = \frac{1}{3}R_Z$$

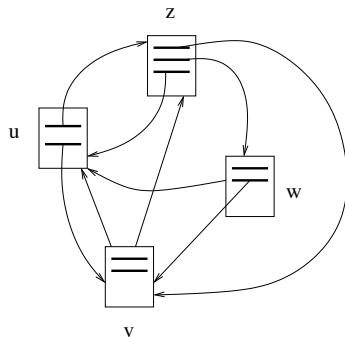
$$R_Z = \frac{1}{2}R_U + \frac{1}{2}R_V.$$

- ▶ 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.

Example (continued)



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).
- ▶ λ 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:

- ▶ An Anatomy of a Large-Scale Hypertextual Web Search Engine, by Sergey Brin and Lawrence Page, 1998. Online at:
<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:
<http://dbpubs.stanford.edu:8090/pub/1999-66>
- ▶ Authoritative Sources in a Hyperlinked Environment, by Jon Kleinberg. Available Online from Jon Kleinberg's webpage:
<http://www.cs.cornell.edu/home/kleinber/>