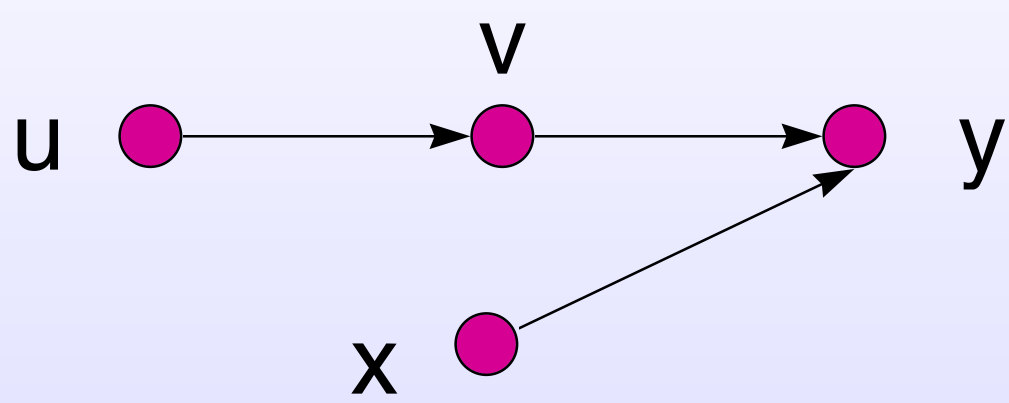
What is this Page Known for? Computing Web Page Reputations

Davood Rafiei, Alberto Mendelzon
University of Toronto

Introduction

- ***Ranking*** plays an important role in searching the Web.
- But the ***importance*** is a subjective measure.
- A high quality page in *computer graphics* is not necessarily a high quality page in *databases*.
- How do search engines address this problem?

Simple Importance Ranking

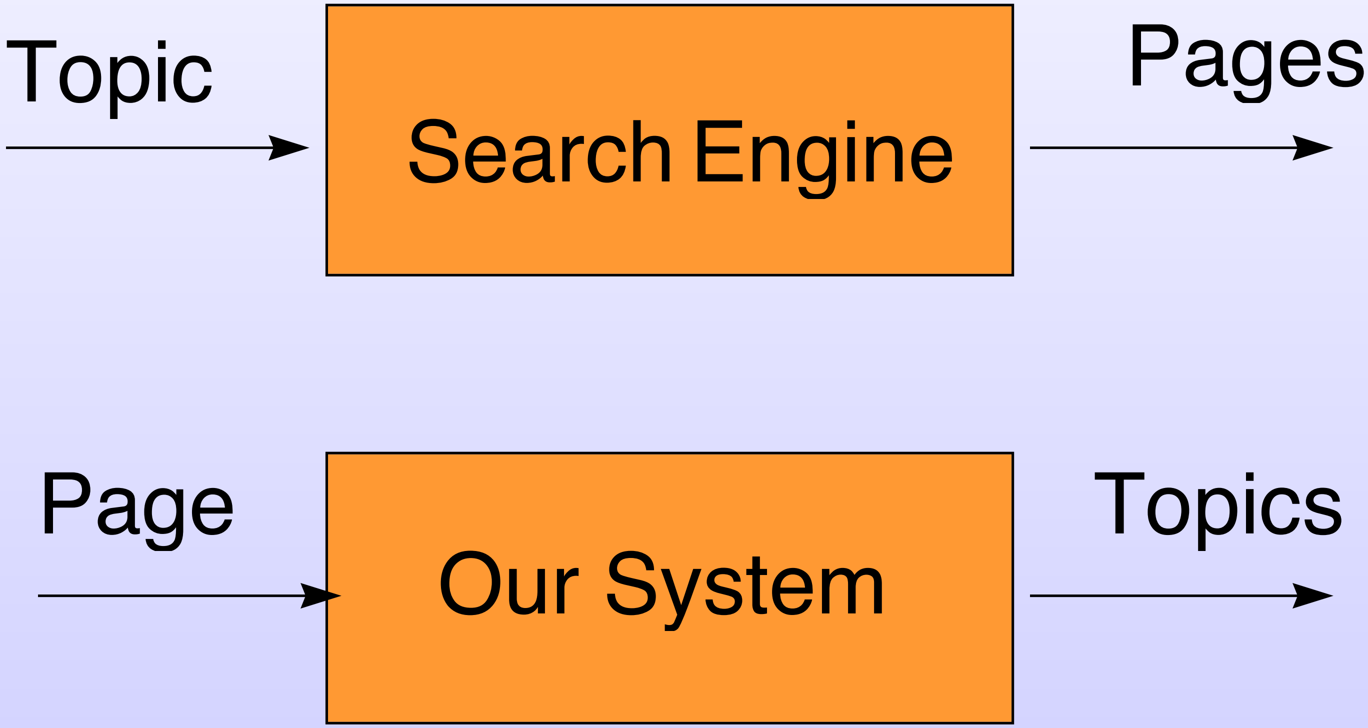


- Rank by in-degree:
  - used in citation analysis (1970s).
  - idea: important journals are frequently cited by other journals.

Importance Ranking: PageRank

- The rank of a page depends on
  - not only the number of its incoming links,
  - but also the ranks of those pages.
- Adopted by Google search engine.
  - high-ranked pages are returned first.
- Limitation: each page is assigned a universal rank, independent of its topic.

Our Goal



Example

What is the page
sunsite.unc.edu/javafaq/javafaq.html
good for?

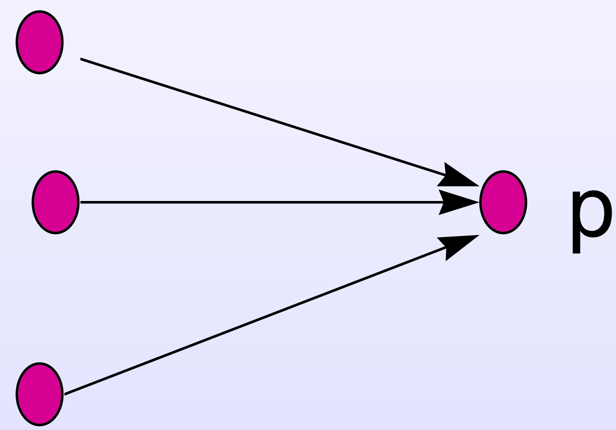
- **Java FAQ**
- **comp.lang.java FAQ**
- **Java Tutorials**
- **Java Stuff**

The Idea

search engines compared

my favorite search engines

a review of search engines



What can we say about the content of Page *p*?

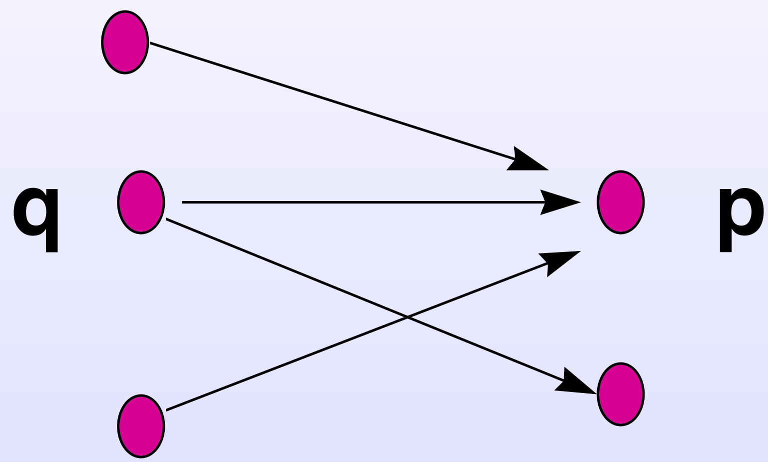
Random Walk Model 1

- Imagine a user searching for pages on topic ***t***.
- The user at each step
  - either jumps into a page on topic ***t*** chosen uniformly at random or
  - follows an outgoing link of the current page.
- The one-level rank of a page on topic ***t*** is the number of visits the user makes into the page if the walk goes forever.

Random Walk Model 1

- **d :** the fraction of times the user makes a random jump.
- **(1-d) :** the fraction of times the user follows a link.
- N_t **:** number of pages on topic t
- $R^n(p,t)$ **:** Prob. of visiting page ***p*** for topic ***t*** at step ***n.***

Probability of Visiting a Page



$$R^{n}(p,t) = (1-d)\sum_{q \rightarrow p}\frac{R^{n-1}(q,t)}{O(q)} + \begin{cases} \dfrac{d}{N_t} & \text{if page } p \text{ is on topic } t \\ 0 & \text{otherwise} \end{cases}$$

Second Scenario

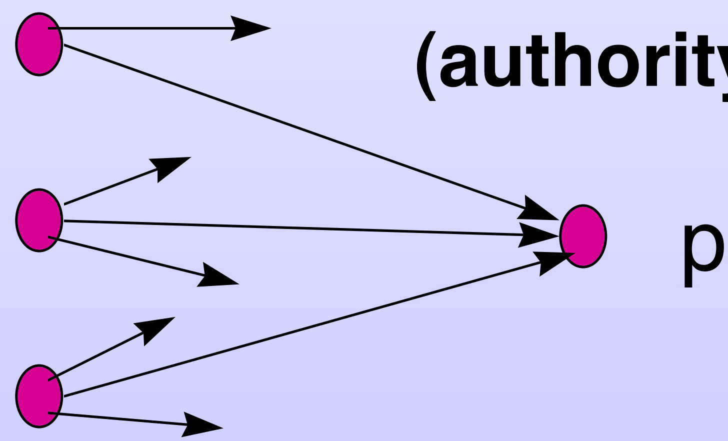
Good source of links (hub)

Good content (authority)

search engines compared

my favorite search engines

a review of search engines



Random Walk Model 2

- Imagine the user at each step
  - ❶ either jumps into a page on topic ***t*** chosen uniformly at random,
  - ❷ follows an outgoing link of the current page (*forward visit*),
  - ❸ or jumps into a page that points to the current page (*backward visit*).
  - ✻ The walk strictly alternates between steps 2, 3.
- The number of forward (backward) visits the user makes into a page is its authority (hub) rank on topic ***t*** if the walk goes forever.

Random Walk Model 2

- **d, (1-d),** N_t : defined similarly.
- $A^n(p,t)$ **:** Prob. of a forward visit into page p at step n.
- $H^n(p,t)$ **:** Prob. of a backward visit into page p at step n.

Probability of Visiting a Page

$$A^n(p,t) = (1-d)\sum_{q\to p}\frac{H^{n-1}(q,t)}{O(q)} + \begin{cases} \dfrac{d}{2N_t} & \text{if page } p \text{ is on topic } t \\ 0 & \text{otherwise} \end{cases}$$

$$H^n(p,t) = (1-d)\sum_{p\to q}\frac{A^{n-1}(q,t)}{I(q)} + \begin{cases} \dfrac{d}{2N_t} & \text{if page } p \text{ is on topic } t \\ 0 & \text{otherwise} \end{cases}$$

Rank Computation

- Done using iterative methods.
- First iteration:
  - Topics are extracted from the content of pages,
  - Ranks are initialized.
- Next iterations:
  - Ranks are propagated through hyperlinks.

Rank Approximation

- A given page ***p*** can acquire a high rank on an arbitrarily chosen topic ***t*** if
  - page ***p*** is on topic ***t***,
  - ***p*** can be reached within a few steps from a large fraction of pages on topic ***t***,
  - or ***p*** can be reached within a few steps from pages with high reputations on topic ***t***.
- An approximate algorithm will examine page ***p*** and only those pages not far away from page ***p***.

Computing One-Level Reputation

For every page **p** and term **t**

R(p,t) = 1/ N_t if term **t** appears in page **p**,

R(p,t) = 0 otherwise

While **R** has not converged

R1(p,t) = 0 for every page **p** and term **t**

For every link $q \rightarrow p$

R1(p,t) += R(q,t) / O(q)

R(p,t) = (1-d) R1(p,t) for every page **p** and term **t**

R(p,t) += d/ N_t if term **t** appears in page **p**.

Computing Two-level Reputation

For every page **p** and term **t**

A(p,t) = H(p,t) = 1/2N_t if term **t** appears in page **p**,

A(p,t) = H(p,t) = 0 otherwise

While both **H** and **A** have not converged

A1(p,t) = H1(p,t) = 0 for every page **p** and term **t**

For every link $q \rightarrow p$

A1(p,t) += H(q,t) / O(q)

H1(q,t) += A(p,t) / I(p)

A(p,t) = (1-d) A1(p,t) and **H(p,t) = (1-d) H1(p,t)**

for every page **p** and term **t**

A(p,t) += d/2N_t and **H(p,t) += d/2**N_t

if term **t** appears in page **p**.

Current Implementation

- Given a page, request its incoming links from Alta Vista.
- Collect the “snippets” returned by the engine and extract candidate terms and phrases.
- Remove stop words.
- Set O(p) = 7.2 for every page p.
- Initialize the weights and propagate them within one iteration.
- Return highly-weighted terms/phrases.

Example

Reputation of
www.macleans.ca :

1 - Maclean's Magazine
2 - macleans
3 - Canadian Universities

Example: Authorities on (+censorship +net)

- *www.eff.org*
  - Anti-Censorship, Join the Blue Ribbon, Blue Ribbon Campaign, Electronic Frontier Foundation
- www.cdt.org
  - Center for Democracy and Technology, Communications Decency Act, Censorship, Free Speech, Blue Ribbon
- www.aclu.org
  - ACLU, American Civil Liberties Union, Communications Decency Act

Example: Personal Home Pages

- www.w3.org/People/Berners-Lee
  - History Of The Internet, Tim Berners-Lee, Internet History, W3C
- www-db.stanford.edu/~ullman
  - Jeffrey D Ullman, Database Systems, Data Mining, Programming Languages
- www.cs.toronto.edu/~mendel
  - Alberto Mendelzon, Data Warehousing and OLAP, SIGMOD, DBMS

Example: Site Reputation

What is this site known for?

- Russia
- Computer Vision
- Images
- Hockey

Example: Site Reputation

Reputation of the **Faculty of Mathematics, Computer Science, Physics and Astronomy** at the **University of Amsterdam** (www.wins.uva.nl):

- Solaris 2 FAQ
- Wiskunde
- Frank Zappa

Limitations

- Our computations are affected by the following two factors:
  - how well is a topic represented on the Web?
  - how well is a page connected?
    - a few pages such as www.microsoft.com have links from a large fraction of all pages on the Web.
    - a large number of pages only have a few incoming links.

Conclusions

- Introduced a notion of reputation
  - combining the textual content and the linkage structure.
- Duality of Topics and Pages
  - Given a page, we currently find a ranked list of topics for the page.
  - However, given a topic, we can also find a ranked list of pages on that topic.

Conclusions

- Our proposed methods generalize earlier ranking methods
  - One-level reputation ranking generalizes PageRank,
  - Two-level reputation ranking generalizes the hubs-and-authorities model.
- Ongoing Work:
  - large-scale implementation of the proposed methods.