Queries

- Specific queries

  Does Netscape support the JDK 1.1 code signing API?

  Usually there are few pages

- Broad-topic queries

  Find information about the Java programming language

  Usually there are too many pages

Slide 1

- Similar page queries

  Find pages similar to `java.sun.com`

  Depends on page

References

- Authoritative Sources in a Hyperlinked environment.
  J.M Kleinberg
  Journl of tha ACM, 46, 1999

- The stochastic approach for link-structure analysis (SALSA) and the TKC effect.
  R. Lempel and S. Moran.
  WWW9, May 2000, Amsterdam

Slide 2

- Finding Authorities and Hubs from link structures on the www.
  A. Borodin, G.O. Roberts, J.S. Rosenthal and P. Tsaparas
  WWW10, May 1-5, 2001, Hong Kong.

- Finding Related pages in the WWW
  J. Dean and M.R. Henzinger

Page type

PageRank is not enough to classify pages. There are other proposed alternatives.

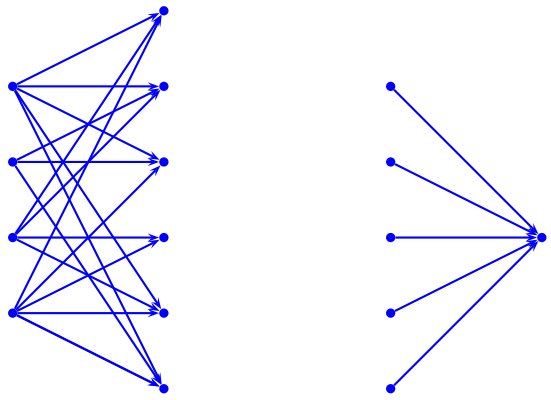
All them look at the web as a linked structure and try to answer broad search queries by finding authoritative information sources

- authoritative pages
- hub pages
- popular pages

Slide 3

from a link point of view?

Slide 4



hubs authorities popular

Why it is difficult to find them

First problem: authoritative sources often do not contain the searched term.

Query: Harvard

`www.harvard.edu` should be one of the authoritative pages, but it does not use the term Harvard very often.

Slide 5

Second problem: good balance between authority and popularity.

popular pages will have high in-degree.

Construct a focused subgraph of the web

We woud like to focus on a collection of pages S_σ such that

- is relatively small
- contains relevant pages
- contains most of the strongests authorities

Slide 6

The construction

R_σ is formed by the t-highest ranked pages from a text-based query.
Set $S_\sigma = R_\sigma$
For each page $p \in R_\sigma$
  Let $\Gamma^+(p)$ be the set of all pages p points to.
  Let $\Gamma^-(p)$ be the set of all pages pointing to p.
  Add al pages in $\Gamma^+(p)$ to S_σ.
  If $|\Gamma^-(p)| \leq d$ then
Slide 7
    Add all pages in $\Gamma^-(p)$ to S_σ
  Else
    Add an arbitrary set of d pages in $\Gamma^-(p)$ to S_σ
  endIf
endfor

The HITS algorithm [Kleinberg]

To each page p in S_σ associate an authority weight $x(p)$ and a hub weight $y(p)$, normalized so that the total weight is one

Starting from some initial assignment apply iteratively

Operation $\mathcal{I}$

$$x(p) := \sum_{(q,p)\in E} y(q)$$

Slide 8

Operation $\mathcal{O}$

$$y(p) := \sum_{(p,q)\in E} x(q)$$

until equilibrium is achieved

The pages with the c largest x value are the authorities

The pages with the c largest y value are the hubs

The SALSA algorithm [Lempel,Moran]

Start by constructing a Base Set.

The algorithm performs a random walk by alternatively

1. going uniformly to one of the pages which links to the current page
2. going uniformly to one of the pages linked to by the current page

The authority wheights are given by the stationary distribution of the two-step Markov chain doing 1 and then 2.

Slide 9

The hub wheights are given by the stationary distribution of the two-step Markov chain doing 2 and then 1.

PageRank provides a ranking that helps to determine how to order the pages returned by a web search query.

HITS considers that web page importance can be measured by two separated measures the authority rating and the hub rating.

SALSA looses the mutually reinforcing structure of HITS, and produces the relative authority/hub scorings from local links.

Slide 10

A normal misbehaviour

Search term jaguar

HITS converges to a collection of sites about the city of Cincinnati!

The cause is a large number of on-line newspaper articles about the Jacksonville Jaguars football team in the Cincinnati Enquirer. All of them link to the same web page of the newspaper.

Therefore not all hubs have the same quality. This is the basis for some variants of the HITS algorithm.

Slide 11

The Hub-Averaging-Kleinberg Algorithm

An hybrid of HITS and SALSA

Starting from some initial assignment apply iteratively

Operation $\mathcal{I}$

$$x(p) := \sum_{(q,p)\in E} y(q)$$

Operation $\mathcal{O}$

$$y(p) := \frac{\sum_{(p,q)\in E} x(q)}{|\Gamma^+(p)|}$$

Slide 12

until equilibrium is achieved

The Threshold-Kleinberg Algorithm

Starting from some initial assignment apply iteratively

Operation $\mathcal{I}$

$$x(p) := \sum_{\{(q,p)\in E | y(p) > H\}} y(q)$$

Operation $\mathcal{O}$

$$y(p) := \sum_{\{(p,q)\in E | x(p) > A\}} x(q)$$

Slide 13

until equilibrium is achieved.

Finding Related pages in the www

Normal feature in a searcher.

The algorithm is an adaptation of the HITS algorithm.

Slide 14

The Companion algorithm

Takes as input a starting URL u and consists of 4 steps.

- Build a vicinity graph for u.
- Contract duplicates and near-duplicates in this graph.
- Compute edge weights based on host to host connections.
- Compute a hub score and an authority score for each node in the graph and return the top ranked authority nodes.

Slide 15

Takes into account the relative order of links inside a page.

The vicinity graph

Uses four parameters B, BF, F, and FB.

The graph consists of

1. u
2. up to B parents of u, and for each parent, up to BF of its children different form u, and
3. up to F children of u, and for each child, up to FB of its parents different form u.

Slide 16

Normal values with good results $F, B = 2000$ and $BF = 8$.

Maintain a list STOP of very popular URLs that are unrelated to many queries, generated through queries history.

If u is in STOP set STOP to empty.

Slide 17

STOP contains most popular search engines and portals

Go Back (B) and Back Forward (BF)

If u has more than B parents
    add B random parents not in STOP
Else
    add all u's parents
If a parent x of u has more than $BF + 1$ children
    add up to $BF/2$ children pointed to by the $BF/2$ links just before u
    add up to $BF/2$ children pointed to by the $BF/2$ links immediatly after u
Else
    Add all its children to the graph

Slide 18

Go Forward (F) and Forward Back (FB)

If u has more than B children
    add the children pointed to by the first F links of u
Else
    add all u's children
If a child x of u has more than BF parents not in STOP
    add the BF parents with highest in-degree
Else
    Add all the parents not in STOP.

Slide 19

Duplicate elimination

Two nodes are near duplicates if

- both have more than 10 links
- they have at least 95% of their links in common

Slide 20

Combine two near duplicates replacing the two nodes by one node whose links are the union of the links of the two near duplicates.

Assign edge weights

- An edge between two nodes in the same host has weight 0
- If there are k edges from documents in a host to a single document in a second host each edge gets an authority weight of $1/k$.
- If there are l edges from a single document in a host to a set of documents in a second host each edge gets a hub weight of $1/k$.

Slide 21

Scoring hubs and authorities

An extension of HITS to weighted graphs considering URLs in the vicinity graph.

Initialize hub score $y[u] = 1.0$ for all u.
Initialize authority score $x[u] = 1.0$ for all u.
While the vectors x and y do no converge
  For all u
    $x[u] = \sum_{(u',u)\in E} y[u'] \times \text{authority-weight}(u', u)$
  For all u
    $y[u] = \sum_{(u,u')\in E} x[u'] \times \text{hub-weight}(u, u')$
  Normalize the x and y vectors

Slide 22