707.000
Web Science and Web Technology
„Link Analysis and Search“

Markus Strohmaier

Univ. Ass. / Assistant Professor
Knowledge Management Institute
Graz University of Technology, Austria

e-mail: markus.strohmaier@tugraz.at
web: http://www.kmi.tugraz.at/staff/markus

Web-Science Summer Academy

7 courses

3 invited talks

Students from all over Europe

Earn credits

Have fun!

Koblenz, Germany

4 Weeks in June/July 2010

WEB SCIENCE SUMMER ACADEMY

Nothing like the Web has ever happened in all of human history. The Web is the largest human information construct in history. The scale of its impact and the rate of its adoption are unparalleled. If we are to ensure that the Web benefits the human race we must do our best to understand it, engineer its future and ensure its social benefit. Web Science is the new interdisciplinary field targeting these objectives.

The Web Science Summer Academy offers a unique combination of 7 courses and 3 invited talks in Web Science. The courses cover socio-economic as well as computer science subjects. All courses award credits for transfers to home institutions. Courses are held in English in a 4 weeks period in June/July 2010.

Join Web Science Summer Academy and:

- explore interdisciplinary facets of web science.
- earn credit points for your studies at home institutions.
- meet a lot of interesting people.
- have fun enjoying a variety of social events.
- if you want: aquire basic skills in the German language.

Courses 2010

Socio-Economic Track

- Virtual Goods
- Social Web and Bibliometry
- System Analysis and Mathematical Modelling
- Agent Based Simulation

Computer Science Track

- Web Retrieval
- Semantic Web
- Web Engineering

Additionally three invited guest lectures will be offered as "Highlights in Web Science."

* Courses and timetable are subject to change.

Previously

Past lectures:

- The Small World Problem
- Network Theory and Terminology
- Social Network Analysis
- Affiliation Networks

Today:

- Link Analysis and Search

Overview

Today‘s agenda

Architecture of search on the web including an overview of

- Crawling, indexing
- Link analysis
- Search Evaluation

Slides based on

- M. Lux, Information Retrieval I&II, Web-based Retrieval, http://www.itec.uni-klu.ac.at/~mlux/
- C. Gütl, Information Search and Retrieval, http://www.iicm.tugraz.at/isr/

Google Wacking: The Belle de Jour Case

http://www.timemachinego.com/linkmachinego/2009/11/16/me-and-belle-de-jour-could-it-be-brooke/

flickr/lmg

[subscribe]
LMG Web Feed

SUBSCRIBE

BOOKMARK

[projects]
TimeMachineGo

Evening Standard

Headline Pictures

Virtual Worlds London

Search LMG

[UK Bloggers]
sashinka
plasticbag.org
Diamond Geezer
Pete's Weblog
As Above
meish dot org
Feeling Listless
Blogadoon
Grayblog
Back in a Bit
Troubled Diva
mondo a-go-go
Random Acts Of Reality
Sore Eyes
plep
Orbyn
info overloaded
The Copydesk
Coffee Grounds
Girl With A...
blackbeltjones
Pandemian
Blah Blah Flowers
Technovia
The View from Here
Belle de Jour
Venusberg
Blogjam
Parallax View
methylsalicylate
Scary Duck

Me and Belle de Jour – 'Could it be Brooke?'

Let's break out of this self-imposed link blogging format for just one post... it's not everyday the biggest secret you've ever kept gets revealed on the front pages of the national press.

I have an admission to make about Belle de Jour.

It's time for me to admit that I solved the puzzle of her identity almost at the very start of her blog after she (as Belle) sent me the link to BdJ to add to the list of Updated UK Blogs. Sending the link to me implied somebody who knew quite a lot about how UK blogging worked at the time. I also found it hard to believe that an escort that had starting blogging would use me to announce the blog to the world. And after BdJ proceeded to knock the ball out of the park in the blog writing department, I started to seriously consider if it was somebody I knew.

I was in the very lucky position of having met many London bloggers at the time and probably briefly read a large percentage of the UK written blogs during 2000 and 2001. UK blogging was (and still is) full of young, smart people and anyone of them might have written Belle's blog. I never believed that a professional writer could be BdJ – apparently effortless blog writing takes practice, and required an understanding of a new medium which not many people had at the time. So I asked myself: which blogger is it?

A couple of months went past, and after Belle de Jour won the award for Best Written Blog from the Guardian and the whole BdJ phenomenon kicked off, I had my eureka moment – I was sitting on the tube one morning and suddenly thought: 'Could it be Brooke?'

Brooke at the time ran a couple of blogs – A link blog called Methylsalicylate and another science blog called Cosmas. She also had done a few short, smart pieces of writing online – The Autopsy, What The Dead Remember and one called Malted. Malted was about whiskey and was the bit of writing that gave it all away. I remembered reading Malted a few months previously and realised the style and content was reminiscent of Belle's and was suddenly convinced I had the answer.

I then spent the first three months of 2004 "internet stalking" Brooke Magnanti, collecting together a whole bunch of circumstantial evidence that Brooke was indeed Belle. I also slowly became aware of the heightened stakes, as Belle became increasingly famous and obviously wanted to maintain her pseudonymity.

For a while I believed that Brooke would get outed immediately – but it turns out the British press could not investigate anything not handed to them on a plate, and were never looking in the right place – the small clique of people who starting blogging in the UK in 2000/2001. Belle de Jour remained pseudonymous and the mystery remained intact even after two TV series based on her books.

During this time I published a googlewack hidden in my blog – the words "Belle de Jour" "Brooke Magnanti" and "Methylsalicylate" were published and available in Google's index on a single page on the internet – my weblog. This "coincidental" collection of links could in no way reveal Belle's identity. But I wondered if anybody else knew the secret and felt that analysing my web traffic might confirm my strongly-held belief. If someone googled "Belle de Jour" "Brooke Magnanti", I would see it in the search referrers for LinkMachineGo.

I waited five years for somebody to hit that page (I'm patient). Two weeks ago I started getting a couple of search requests a day from an IP address at Associated Newspapers (who publish the Daily Mail) searching for "brooke magnanti" and realised that Belle's pseudonymity might be coming to an end. I contacted Belle via Twitter and let her know what was happening. I didn't expect to hear anything back.

And then early last weekend I received an email signed by Brooke that confirmed that she was outing herself in the Sunday Times because the Daily Mail had discovered her identity via an ex-boyfriend.

It was finally over, the secret was out. I no longer have to worry about inadvertently revealing her identity. If I'm honest, solving the puzzle of the biggest literary and blogging mystery of the last six years has been fun and exciting. I'm just really disappointed I don't get to dig up a gold hare as a prize!

One last thing: Good Luck Brooke, I'm very glad you've managed to maintain some control over how and when your real identity was revealed to the public. I think I probably owe you a bottle of your favourite whiskey. Let me know what you like and I'll see what I can do. — Darren/LMG.

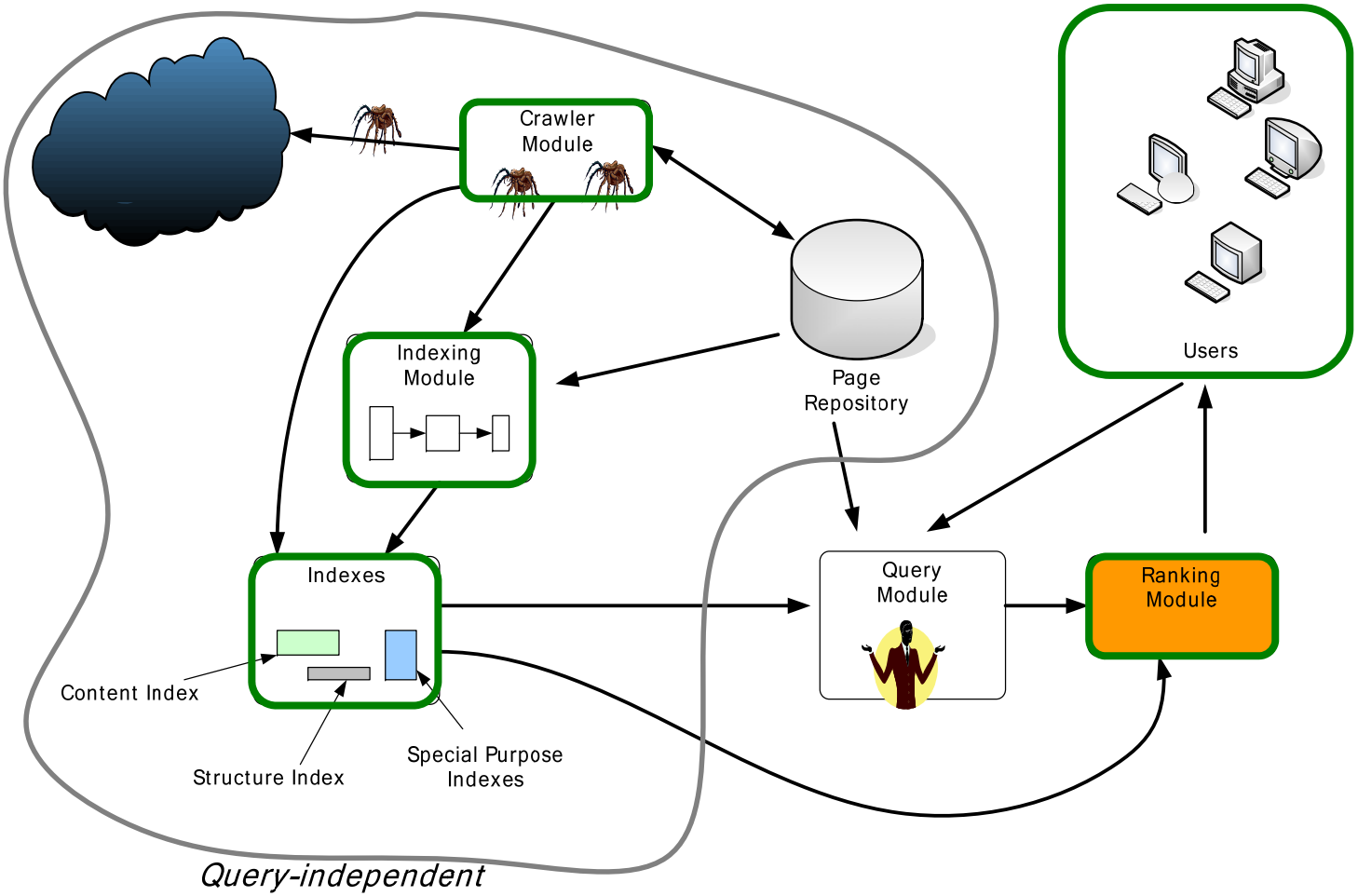
Update #1: Belle has confirmed the story in the comments to this post.

Update #2: Googlewack Screengrab Published

Update #3: How Belle de Jour's Secret Ally Googlewacked The Press — the Guardian cover the story.

Me and Belle de Jour – 'Could it be Brooke?'

Common Architecture



History of Crawlers [Witten 2007]

- World Wide Web Wanderer (1993)
  - Purpose not to index, but to measure its growth
- WebCrawler (1994)
  - First full-text index for entire web pages
- Lycos, Infoseek, Hotbot (1996)
- AskJeeves, Northern Light (1997)
- Others: OpenText, AltaVista
- Yahoo (What's that acronym?)
  - Two Stanford PhD students

officious
adj. eager to offer unwanted services; meddlesome; interfering; offering much unwanted advice

„Yet Another Hierarchical Officious Oracle"

And then came Google (1998)

- Another two Stanford PhD students (T. Winograd)
- Who are now allowed to land their private air planes on a NASA airfield close to Mountain View
  http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2007/09/13/BUPRS4MHA.DTL

Crawler

Crawlers, robots & spiders harvest sites

Starting with a **root set** of URLs

Following links, that are found on the pages

Applying **filters** to the links

- e.g. only .at domains -> Austrian web pages
- e.g. based on link title & position (focused crawling)

Crawlers: Index Update

- Which sites should be updated and when?
- A page content might have changed since last visit
  - last modified dates are possibly inaccurate
- Different strategies are possible:
  - Refresh only portions ...
  - Prefer most popular sites ...

Ethical Questions:

- How much bandwidth is used?
  - Hit counts ...
- What does that mean for the server load?
- Let loose several spiders at once
  - Decrease of crawling time
  - Increase of load

Crawling: Robots.txt

Robots.txt is an option for webmasters to

- restrict crawler access
- point crawlers to interesting URLs
- identify crawlers (via hits on the robots.txt)
- see http://www.robotstxt.org/wc/robots.html

Example

```
User-agent: *
Disallow: /wp-admin/
Disallow: /netadmin/
```

Crawler: Google sitemaps

XML schema to identify interesting portions & updates of a web page

Integration into CMS possible

Example:

```
<url>
    <loc>http://www.semanticmetadata.net/</loc>
    <lastmod>2007-02-06T11:26:06+00:00</lastmod>
    <changefreq>daily</changefreq>
    <priority>1</priority>
</url>
```

Crawler: Coverage, Freshness and Coherence [Witten 2007]

Coverage:

- The percentage of pages that a crawler indexes

Freshness:

- The reciprocal of the time that elapses between successive visits to websites

Coherence:

- The overall extent to which the index corresponds to the web itself

Indexing Module

Takes each new uncompressed page

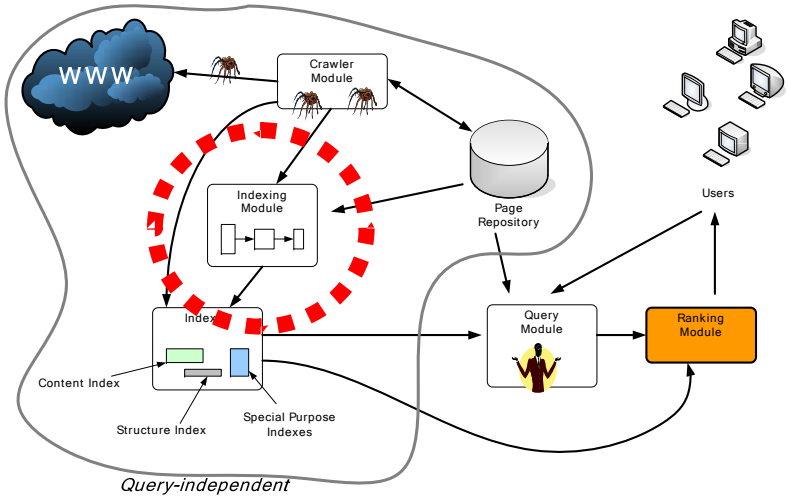
Extracts vital descriptors

- terms, positions, links

Creates compressed version of page

Stores

- Page in cache
- Descriptors in index



Constructing a Full-text Index [Witten 2007]

word	position in text
be	2 6 ...
is	8 ...
not	4 ...
or	3 ...
question	10 ...
that	7 ...
the	9 ...
to	1 5 ...

(a) The beginning of the index.

1	2	3	4	5	6	7	8	9	10
to	be	or	not	to	be	that	is	the	question . . .

(b) The text.

Figure 4.3 Making a full-text index.

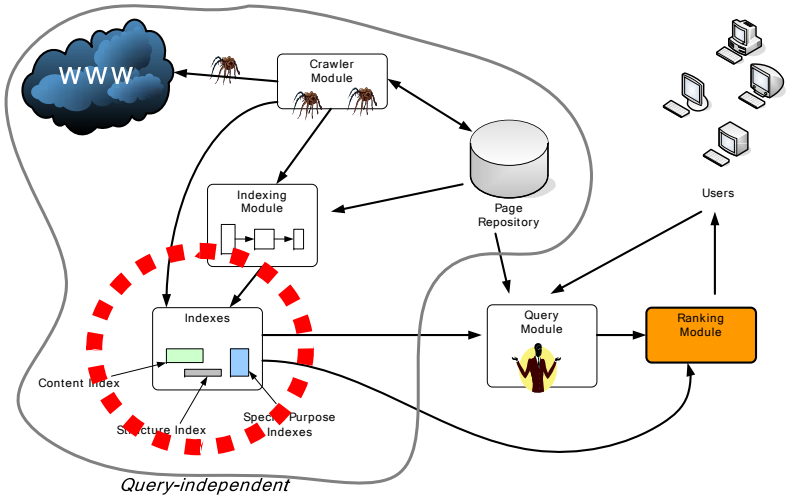
Indexes

Content Index

Structure Index

Special Purpose Index

- Document Formats (PDF, Doc, ...)
- Media (Images, Video, ...)



Indexes

Content Index

- Inverted Document Index
  - term x -> <d11>, <d28>, <d31>, ...
  - term y -> <d10>, <d35>, <d36>, ...
- Index is a
  - quick lookup table
  - smaller than documents

Structure Index

- Hyperlink Information
- In-links, out-links & self-links
- Stored for ...
  - Later analysis
  - Later queries (who links to whom)

Ranking Module

- Orders set of relevant pages
  - Input from query module
- Employs **ranking algorithm**
  - Based on several aspects (terms, links, etc.)
  - Overall score is combination of
    - Content score (TF*IDF)
    - Popularity score (PageRank, HITS, etc.)

Popularity Ranking

- 2 Algorithms developed independently
  - PageRank, Brin & Page
  - Hypertext Induced Topic Search (HITS), Kleinberg
- Basic idea of popularity
  - Someone likes a page
  - Gives a recommendation (on another page)
  - Using a hyperlink

Popularity Ranking: Basic Idea

There are different types of people:

- Regarding their idea of recommendation
  - People giving a lot of recommendations (links)
  - People giving few recommendations (links)
- Regarding their state of recommendation
  - Recommended by a lot of people
  - Recommended by few people

Combinations are possible:

- Having no recommendation, but recommending a lot, ...

Popularity Ranking: Basic Idea

Think of ....

people as pages

recommendations as links

PageRank (Google)

Therefore:

"Pages are popular, if popular pages link them"

"PageRank is a global ranking of all web pages, regardless of their content, based solely on their location in the Web's graph structure." [Page et al 1998]

A Tangled Web [Witten 2007]

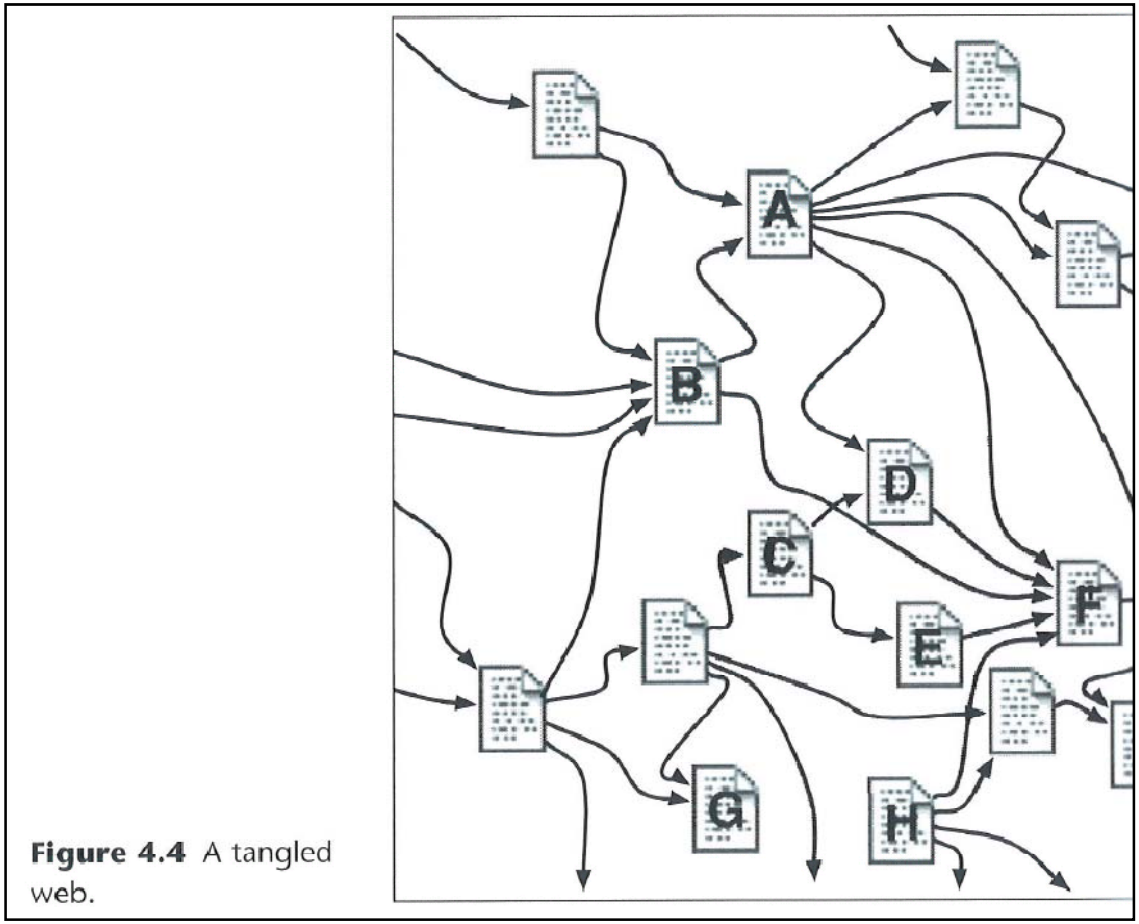
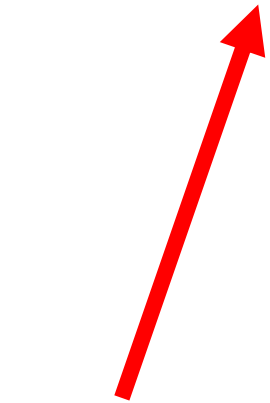


Figure 4.4 A tangled web.

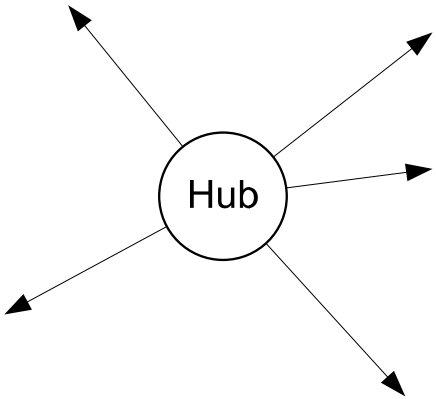
Popularity Ranking: Basic Idea

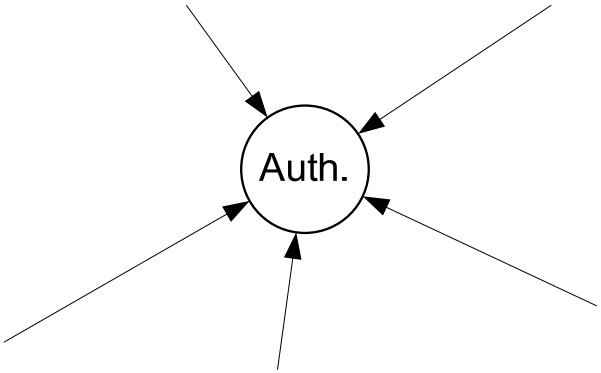
Additional assumptions:

- **Hubs** are pages that point to highly ranked vertices
- **Authorities** are pages, which are pointed to by highly ranked vertices



HITS





PageRank: Original Summation Formula

Original summation formula

– PageRank of page P_i is given by the summation of: all pages P_j that link to P_i (given by the set B_{Pi}) *divided by the set of outbound links of* P_j: $|P_j|$

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|},$$

Iterative formula, starting with rank 1/n for all n pages:

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

PageRank: Original Summation Formula [Page et al 1998]

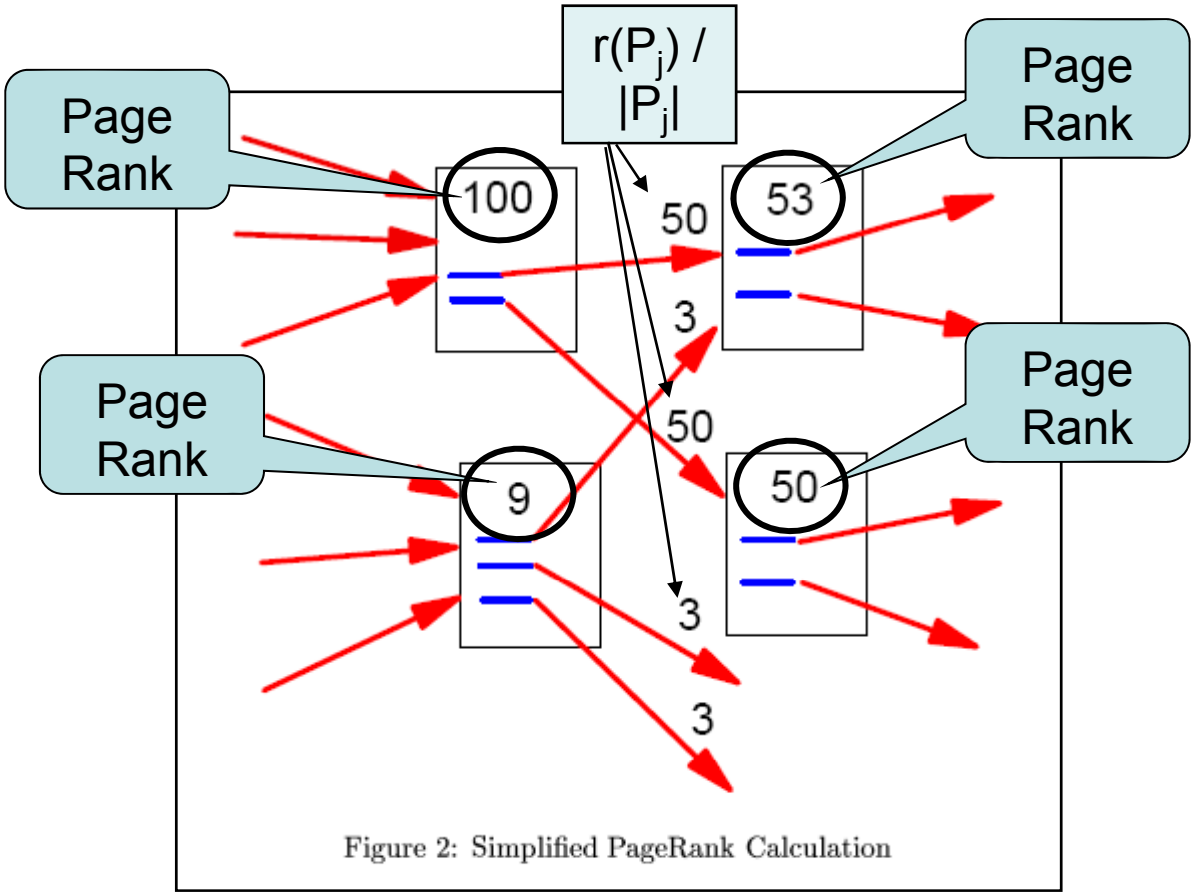
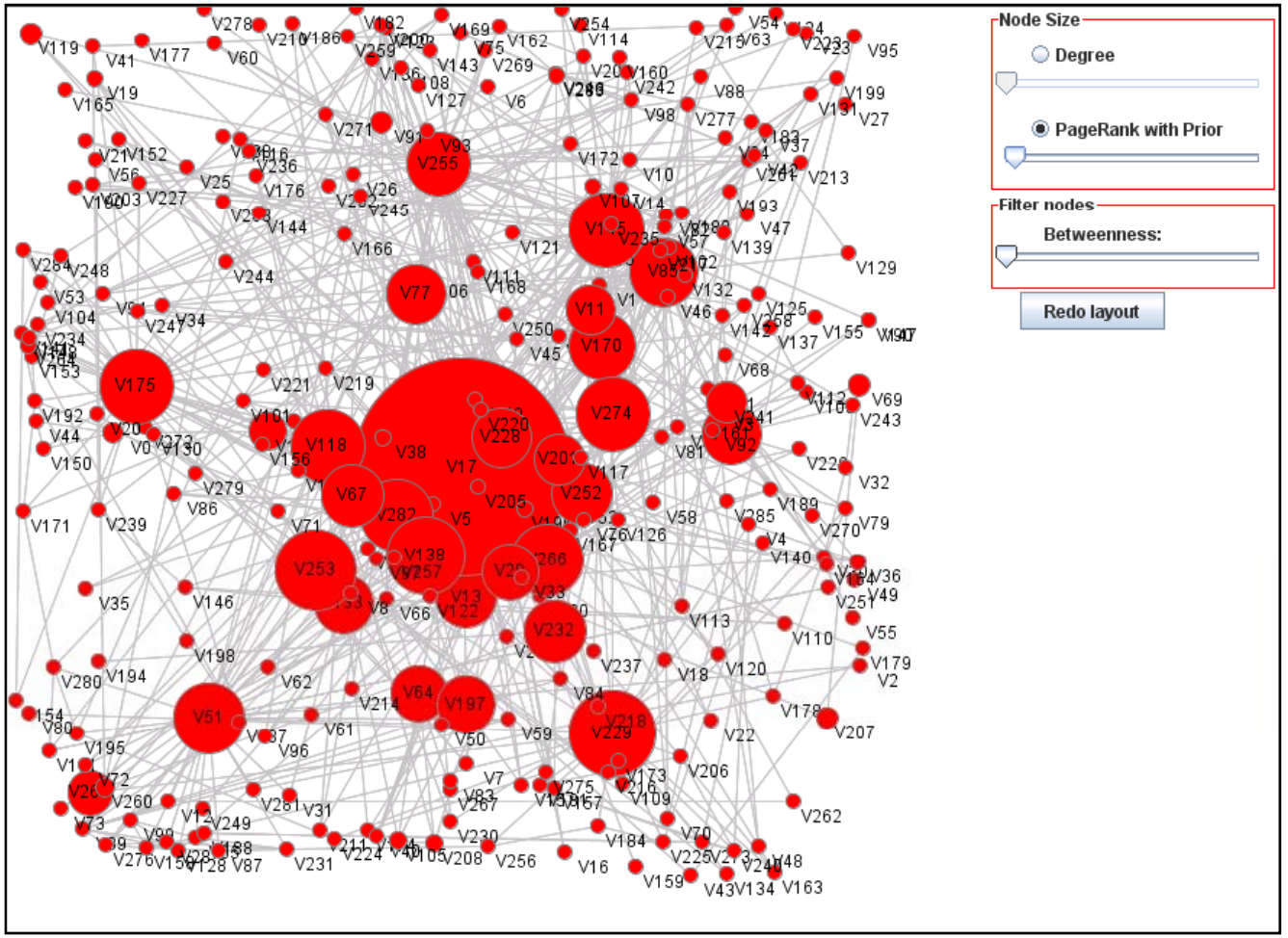


Figure 2: Simplified PageRank Calculation

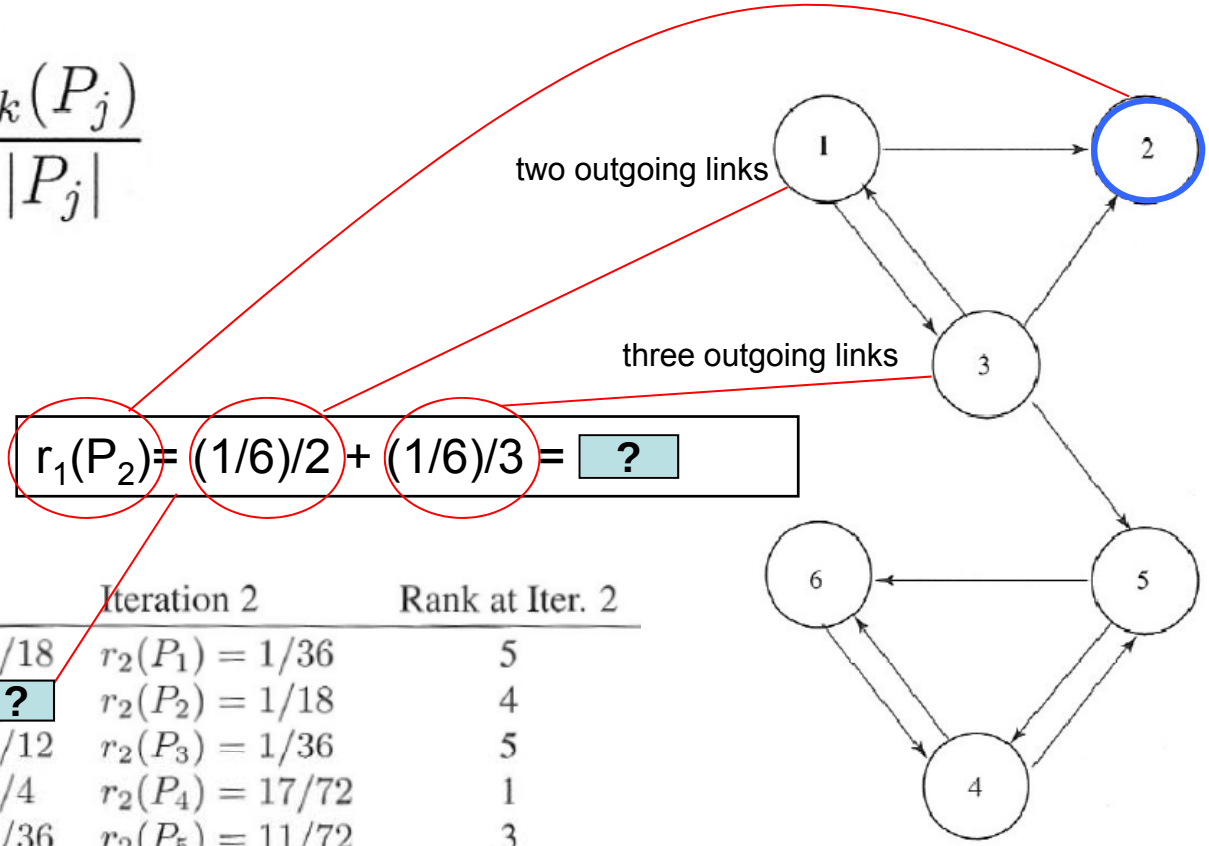
PageRank

http://jung.sourceforge.net/applet/rankingdemo.html



PageRank: Original Summation Formula [Amy N. Langville and Carl D. Meyer 2004]

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$



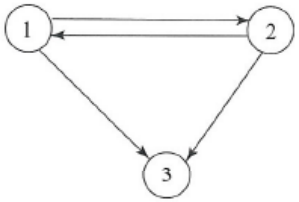
Iteration 0	Iteration 1	Iteration 2	Rank at Iter. 2
$r_0(P_1) = 1/6$	$r_1(P_1) = 1/18$	$r_2(P_1) = 1/36$	5
$r_0(P_2) = 1/6$	$r_1(P_2) =$?	$r_2(P_2) = 1/18$	4
$r_0(P_3) = 1/6$	$r_1(P_3) = 1/12$	$r_2(P_3) = 1/36$	5
$r_0(P_4) = 1/6$	$r_1(P_4) = 1/4$	$r_2(P_4) = 17/72$	1
$r_0(P_5) = 1/6$	$r_1(P_5) = 5/36$	$r_2(P_5) = 11/72$	3
$r_0(P_6) = 1/6$	$r_1(P_6) = 1/6$	$r_2(P_6) = 14/72$	2
$\sum = 1$	$\sum < 1$	$\sum < 1$	

sinks!

Initial Problems

Rank sinks & cycles:

- Some pages get all of the score, other pages none
- Cycles just flip the rank
- Some nodes do not have outlinks: Dangling nodes



How many iterations?

- Will the process converge?
- Will it converge to one single vector?

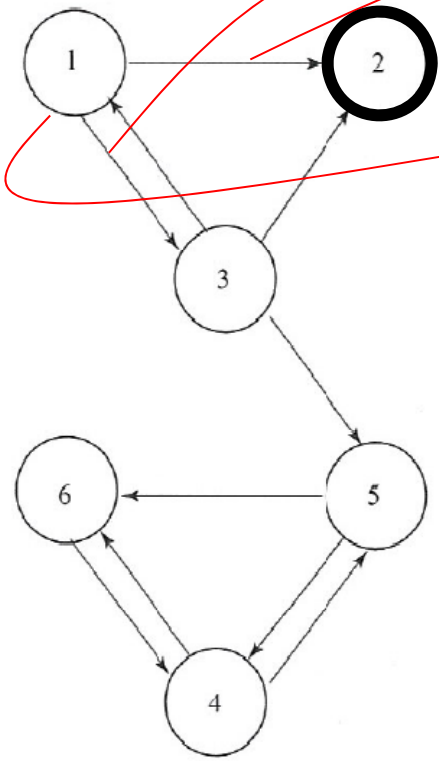
Approach of Brin & Page

Notion of the random surfer

- Someone navigates through the web using hyperlinks
- If there are 6 links, there is a probability of 1/6 that s/he takes a specific link
- On dangling nodes (without out links) s/he can jump everywhere with equal chance
- Furthermore s/he can leave the link path with a given probability every time

- What would happen **without the random surfer** model?
- http://projects.si.umich.edu/netlearn/GUESS/pagerank.html (Allow / Disallow sinks)

Approach of Brin & Page: Example taken from [Amy N. Langville and Carl D. Meyer 2004]



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

replace all zero rows, 0^T, with $1/n\ e^T$, where e^T is the row vector of all ones and n is the order of the matrix.

Dealing with dangling nodes

$$\mathbf{S} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Leaving the link structure: [Amy N. Langville and Carl D. Meyer 2004]

Introduction of the Google Matrix: $\mathbf{G} = \alpha \mathbf{S} + (1-\alpha) 1/n \, \mathbf{e}\mathbf{e}^T$

$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Considering dangling nodes

Random surfer

$$\mathbf{G} = .9\mathbf{H} + (.9 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + .1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}) \, 1/6 \, (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1)$$

S

$$= \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix}.$$

Brin and Page suggested a damping factor α = 0.85

„That means, roughly five-sixths of the time a web surfer randomly clicks on hyperlinks (i.e. following the structure of the web) while one-sixth of the time this web surfer will go to the URL line and type the address of a new page to „teleport“ to. “

Every node is now directly connected to every other node

The Google Matrix Step by Step

$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

$$\mathbf{G} = .9\mathbf{H} + (.9 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + .1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}) \, 1/6 \, (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1)$$

$$0.9 * \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 9/20 & 9/20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 3/10 & 3/10 & 0 & 0 & 3/10 & 0 \\ 0 & 0 & 0 & 0 & 9/20 & 9/20 \\ 0 & 0 & 0 & 9/20 & 0 & 9/20 \\ 0 & 0 & 0 & 9/10 & 0 & 0 \end{pmatrix}$$

The Google Matrix Step by Step

$$\mathbf{G} = .9\mathbf{H} + \left(.9 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + .1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}\right) 1/6 \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 0 \\ 9/10 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1/10 \\ 1/10 \\ 1/10 \\ 1/10 \\ 1/10 \\ 1/10 \end{pmatrix} = \begin{pmatrix} 1/10 \\ 10/10 \\ 1/10 \\ 1/10 \\ 1/10 \\ 1/10 \end{pmatrix}$$

The Google Matrix Step by Step

$$\mathbf{G} = .9\mathbf{H} + (.9 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + .1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}) \, 1/6 \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1/10 \\ 10/10 \\ 1/10 \\ 1/10 \\ 1/10 \\ 1/10 \end{pmatrix} * 1/6 \, (1\ 1\ 1\ 1\ 1\ 1) = \begin{pmatrix} 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \end{pmatrix}$$

The Google Matrix Step by Step

$$\mathbf{G} = .9\mathbf{H} + \left(.9 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + .1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}\right) 1/6 \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$G = \begin{pmatrix} 0 & 9/20 & 9/20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 3/10 & 3/10 & 0 & 0 & 3/10 & 0 \\ 0 & 0 & 0 & 0 & 9/20 & 9/20 \\ 0 & 0 & 0 & 9/20 & 0 & 9/20 \\ 0 & 0 & 0 & 9/10 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/60 & 1/60 \end{pmatrix} = \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix}$$

Result of the adaptations [Amy N. Langville and Carl D. Meyer 2004]

Iterative Formula

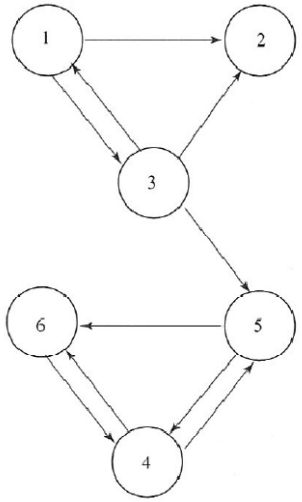
$$\pi^{(k+1)T} = \pi^{(k)T}\mathbf{G},$$

– Converges to a single PageRank vector

In our example:

$$\pi^T = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix}$$
$$\begin{pmatrix} .03721 & .05396 & .04151 & .3751 & .206 & .2862 \end{pmatrix}$$

taken from “Google’s PageRank & Beyond”, Langville & Meyer



Retrieval Evaluation: Motivation

Objectively compare different

- Search engines
- Models & Weighting Schemes
- Methods & Techniques

Scope

- Academic
- Commercial & Industrial

Axis

- Runtime, Retrieval performance

Retrieval Evaluation

Approaches since first prototypes differ in:

- Test collections
- Experts assessing retrieval performance
- Metrics
  - What's good? / What's bad?

Overall problem:

- What is relevant?

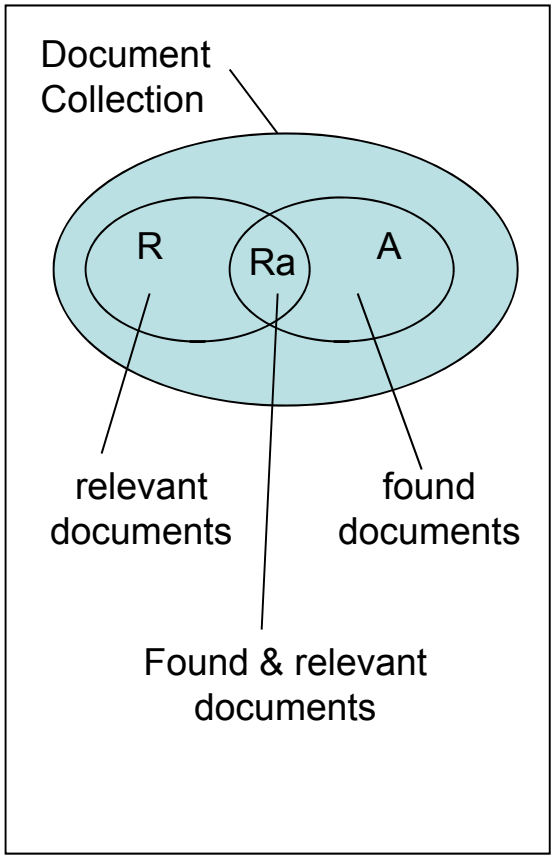
Metrics: Precision & Recall

Within a document collection *D* with a given query *q*

|R| .. num. of relevant docs

|A| .. num. of found docs

|Ra| .. num. found & relevant



Metrics: Precision

$$\text{Precision} = \frac{|Ra|}{|A|} = \frac{\text{found relevant docs}}{\text{found docs}}$$

Gives % how many of the actual found documents have been relevant

Between 0 and 1

- Optimum: 1 ... all found docs are relevant

Metrics: Recall

$$\text{Recall} = \frac{|Ra|}{|R|} = \frac{\text{found relevant docs}}{\text{relevant docs}}$$

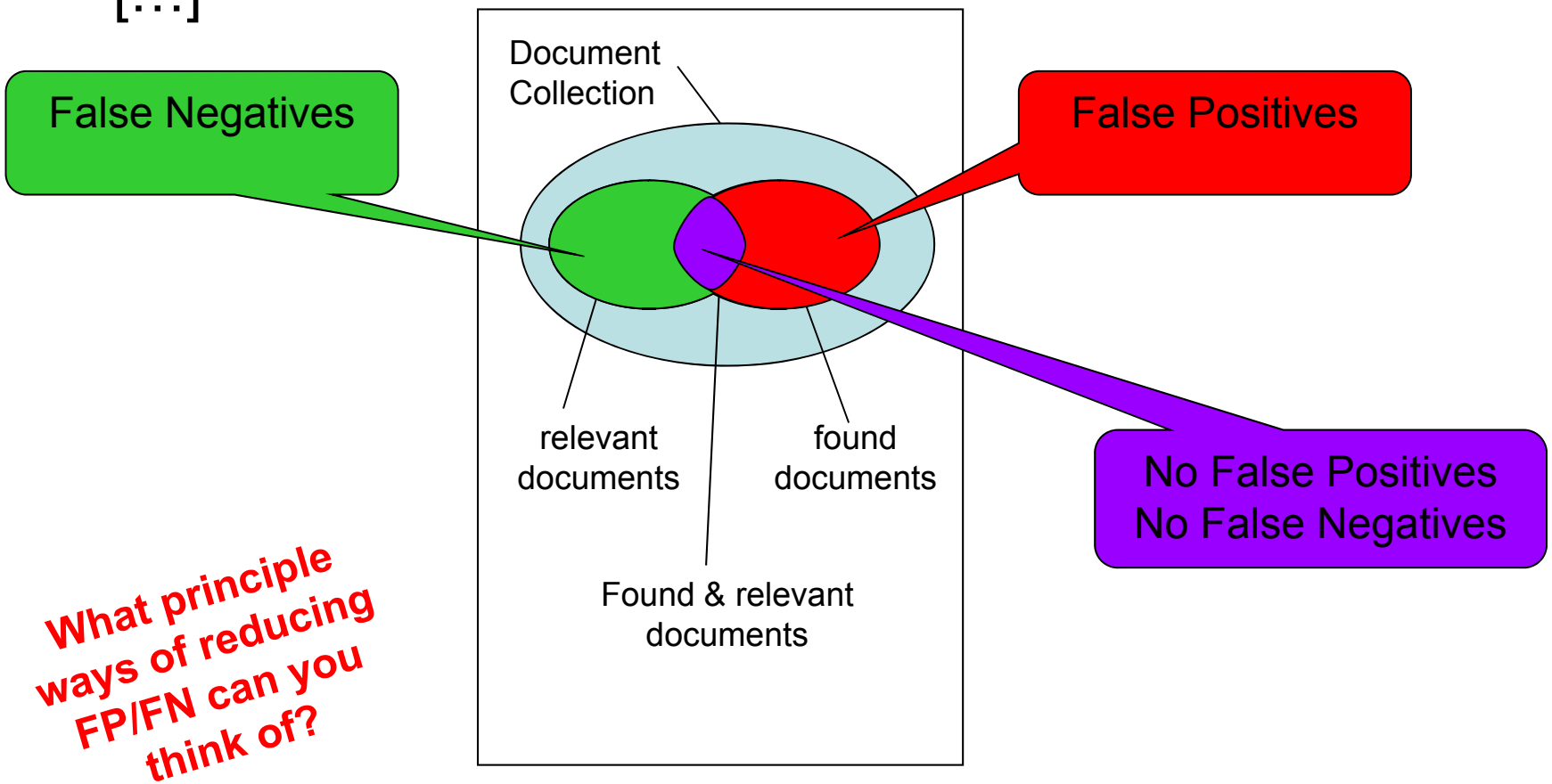
Gives % how many of the actual relevant documents have been found

Between 0 and 1

- Optimum: 1 ... all relevant docs are found

False Positives and False Negatives

[…]



What principle ways of reducing FP/FN can you think of?

Examples: Precision & Recall

With a query only 1 document has been found, but this one is relevant (100 would be relevant):

- Precision & Recall?

- **Precision = 1**
- **Recall = 0,01**

With a query all documents of D have been found (5% of D would be relevant)

- Precision & Recall?

- **Precision = 0,05**
- **Recall = 1**

Recall vs. Precision Plot

Assumption:

- Result list is sorted by descending relevance
- User investigates result list linearly
  - Precision and Recall change

Approach:

- Map different states to graph

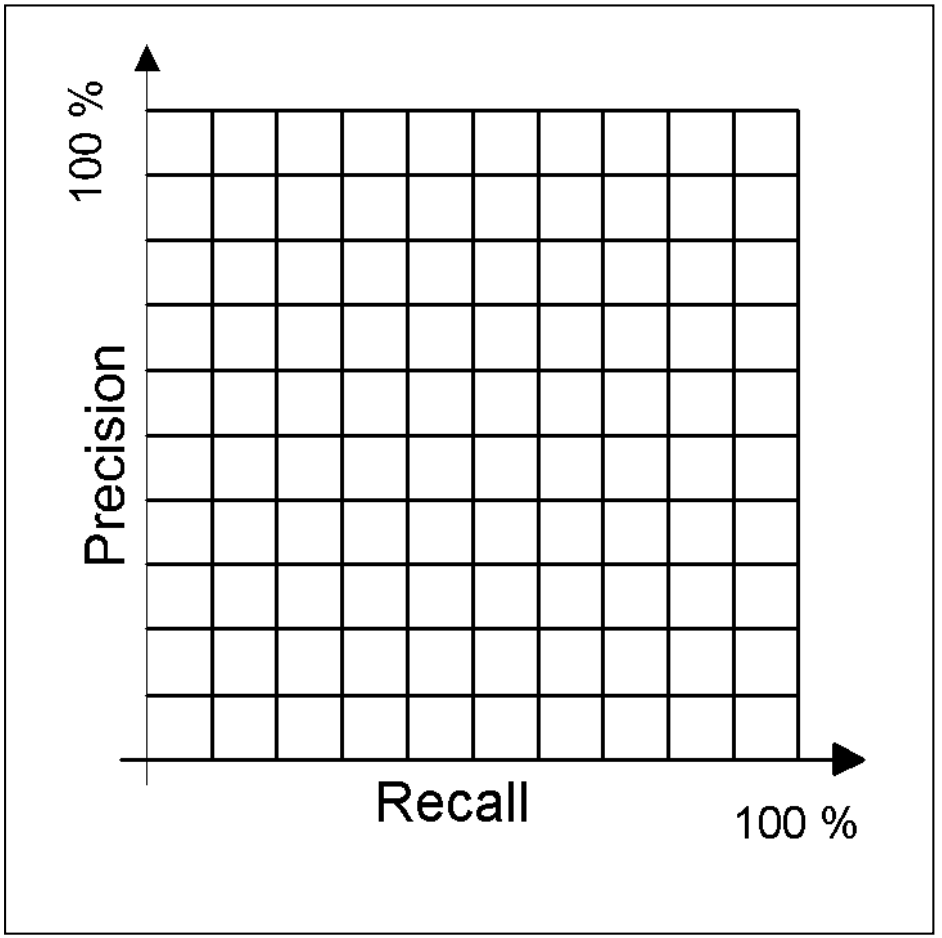
Recall vs. Precision Plot

Result Set:

01. d123 *	06. d9 *	11. d38
02. d84	07. d511	12. d48
03. d56 *	08. d129	13. d250
04. d6	09. d187	14. d113
05. d8	10. d25 *	15. d3 *

Relevant Results:

Rq={d3, d5, d9, d25, d39, d44, d56, d71, d89, d123} → ∑ 10



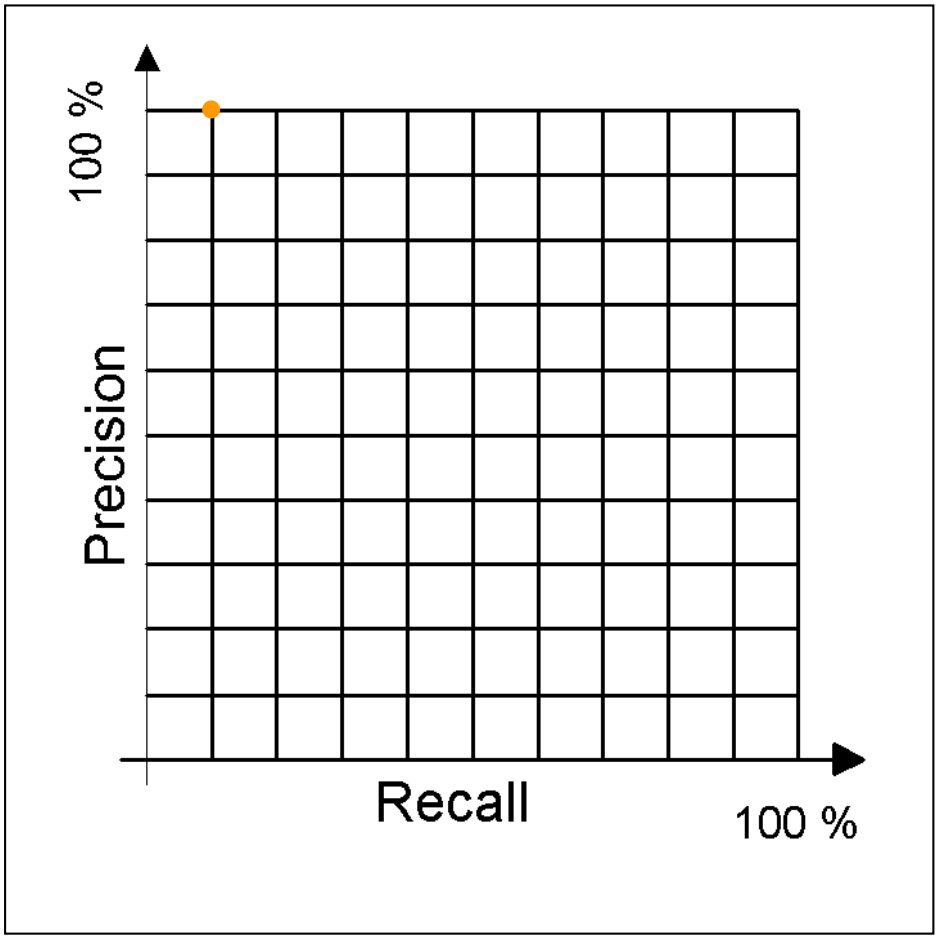
Recall vs. Precision Plot

01. d123 *	06. d9 *	11. d38
02. d84	07. d511	12. d48
03. d56 *	08. d129	13. d250
04. d6	09. d187	14. d113
05. d8	10. d25 *	15. d3 *

11 Standard Recall Levels
{0%, 10%, 20%, ... , 90%, 100%}

$$\text{Recall} = \frac{|Ra|}{R} = \frac{1}{10}$$

$$\text{Precision} = \frac{|Ra|}{A} = \frac{1}{1}$$

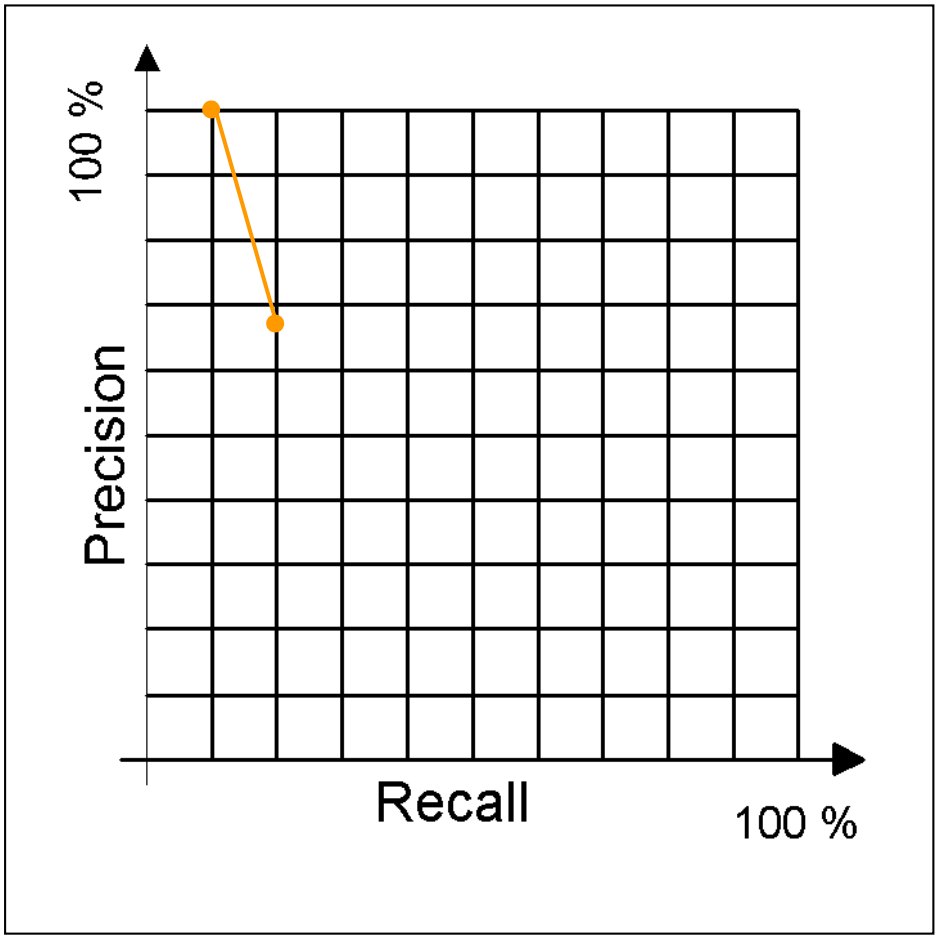


Recall and Precision

01. d123 *	06. d9 *	11. d38
02. d84	07. d511	12. d48
03. d56 *	08. d129	13. d250
04. d6	09. d187	14. d113
05. d8	10. d25 *	15. d3 *

$$\text{Recall} = \frac{|Ra|}{R} = \frac{2}{10}$$

$$\text{Precision} = \frac{|Ra|}{A} = \frac{2}{3}$$

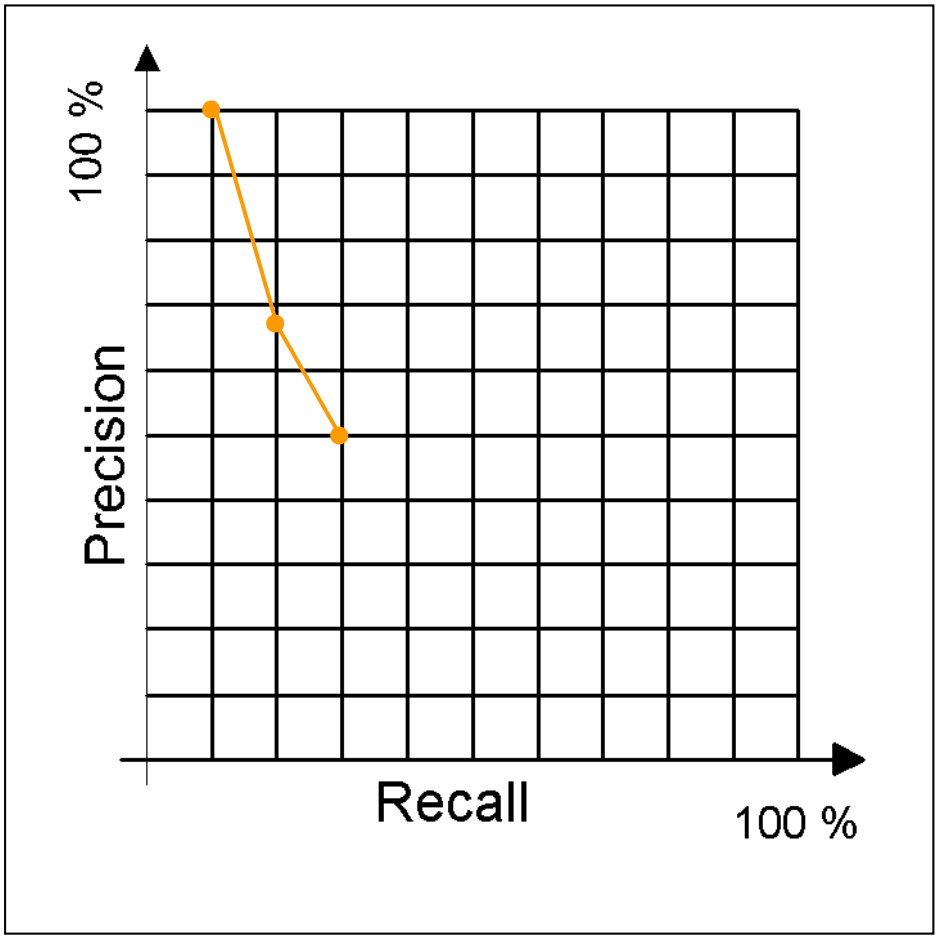


Recall and Precision

01. d123 *	06. d9 *	11. d38
02. d84	07. d511	12. d48
03. d56 *	08. d129	13. d250
04. d6	09. d187	14. d113
05. d8	10. d25 *	15. d3 *

Precision = ?

Recall = ?

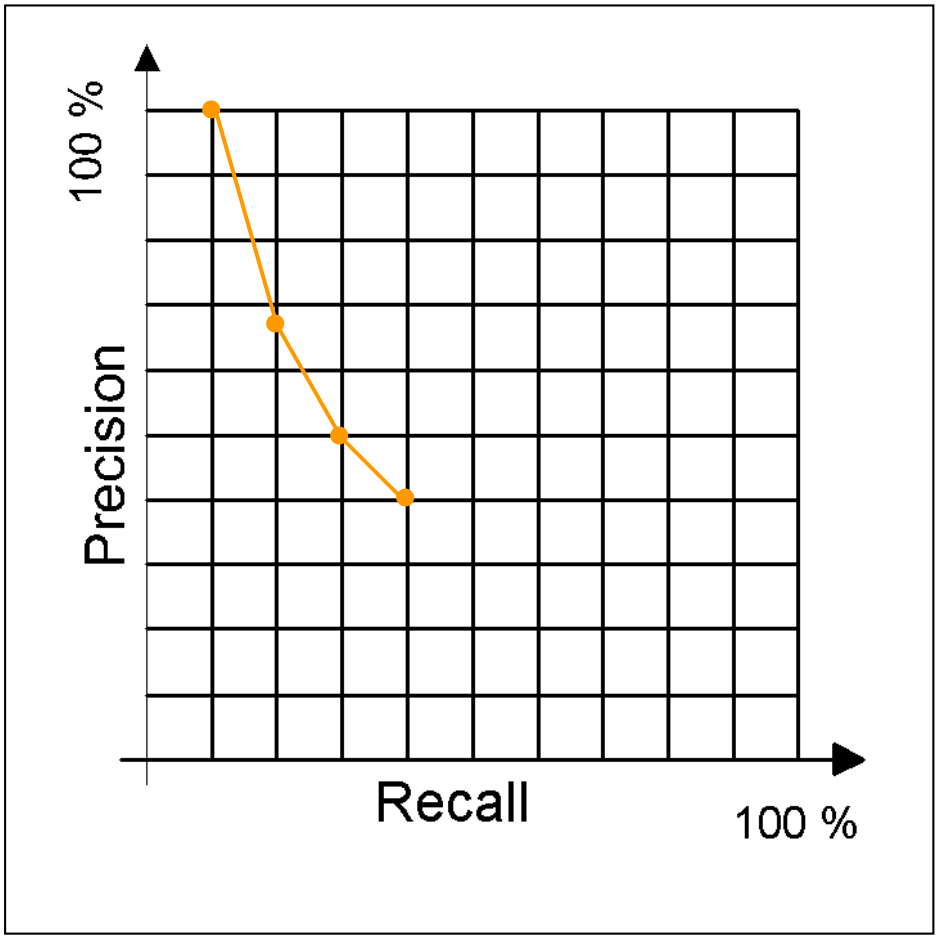


Recall and Precision

01. d123 *	06. d9 *	11. d38
02. d84	07. d511	12. d48
03. d56 *	08. d129	13. d250
04. d6	09. d187	14. d113
05. d8	10. d25 *	15. d3 *

Precision = ?

Recall = ?



Confusion Matrix

Query	In Query (positiv)	Nicht in Query (negativ)
Relevant	TP	FN
Nicht Relevant	FP	TN

$$R = \frac{TP}{TP + FN}$$

Recall

Precision

$$P = \frac{TP}{TP + FP}$$

Kombination im F-Measure

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

$$\beta = 1 \Rightarrow F_1 = \frac{2PR}{P + R}$$

Problems

The Deep Web

What is the deep web?

⇨Pages crawlers do not currently find.

Example: http://www.aekstmk.or.at/

Communications of the ACM

Volume 50, Number 5 (2007), Pages 94-101

"Accessing the deep web", Bin He, Mitesh Patel, Zhen Zhang, Kevin Chen-Chuan Chang

Problems

Spam

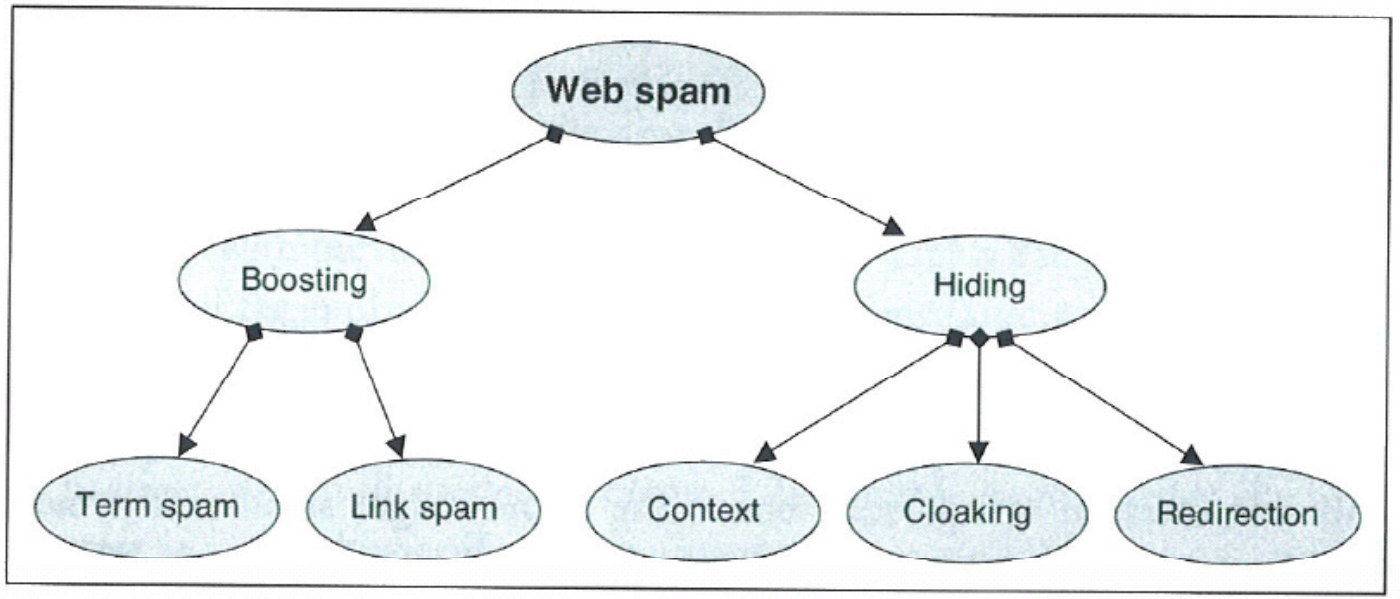


Figure 5.1 The taxonomy of web spam.

Problems: Spam

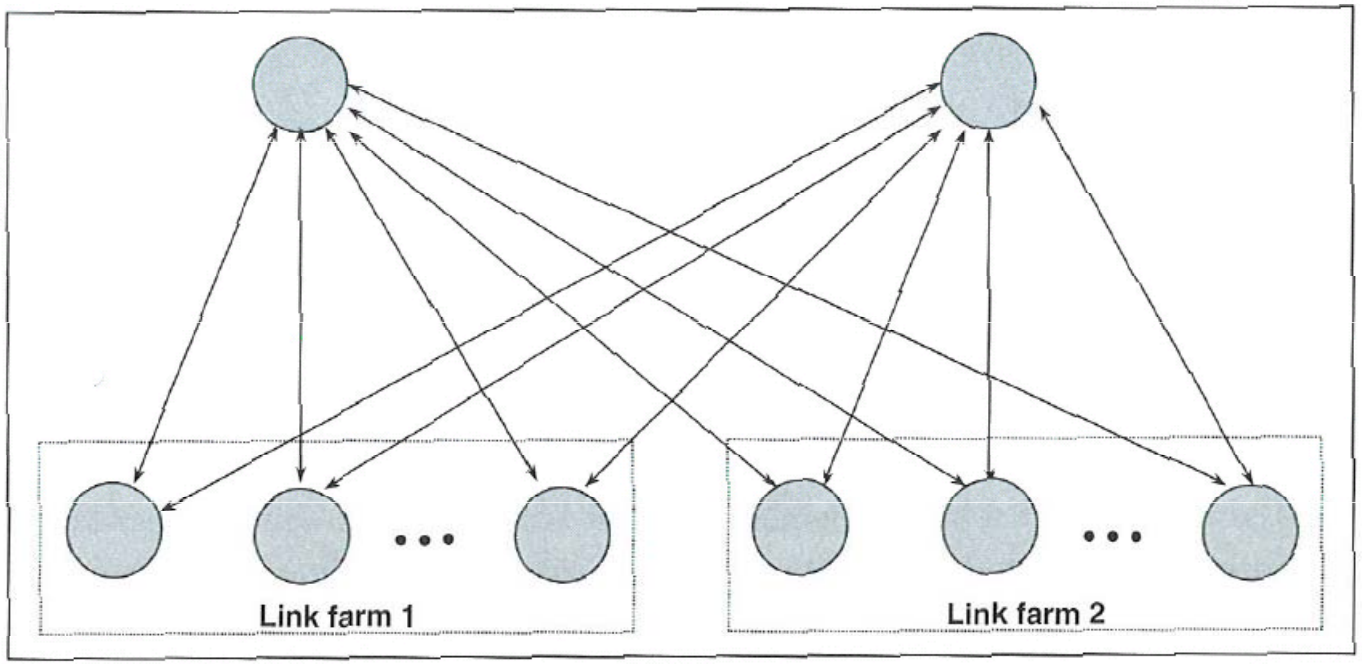


Figure 5.4 A spam alliance in which two link farms jointly boost two target pages.

Any questions?

See you THURSDAY!

(check TUGonline for details)