

Web Spam, SEO and AIR

COMP {3410,6341}
IT in e-commerce
Tim Jones

In recent years, the web has become more and more popular

- General information (wikipedia, etc)
- Corporate information (apple iphone, political campaigns, etc)
- Shopping (amazon, mp3s, travel, etc)

For most people, search is now the main entry point to the web

But people don't look past the first page of results

People click the first result twice as often as the second

Source: Joachims et al. 2005

So, if you want people to find your page, having a high rank in search results is important

Lecture Overview

- Introduction
- Ranking functions
 - And how people cheat
- Web spam vs Search Engine Optimisation
- Adversarial Information Retrieval
 - Some examples

A quick history of web search ranking:

- Tf.Idf
- In-link counting
- PageRank / HITS

We'll talk a little about each

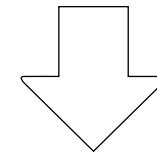
Tf.Idf

- Term frequency . inverse document frequency
- IDF is the log of the percentage of documents that contain the term
- Used for ranking documents relative to a query

The number of times a query term appears in a document, times the rarity of that term in the collection

Tf.Idf is easy to attack

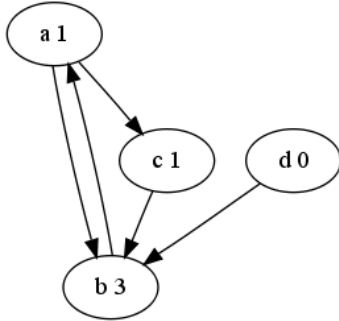
Cheap Flights from Sydney to London



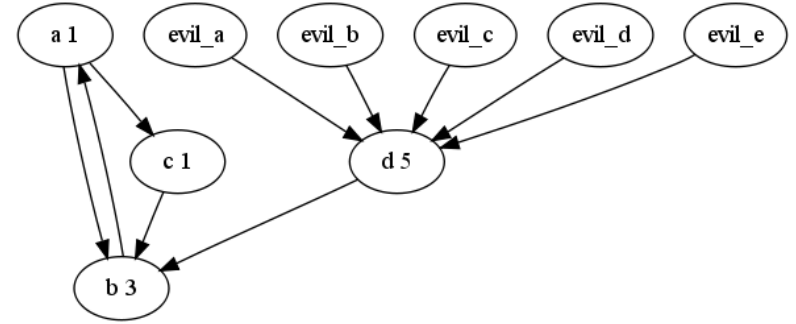
Cheap Flights from Sydney to London
cheap flights cheap flights cheap flights
cheap flights cheap flights cheap flights

In-link counting

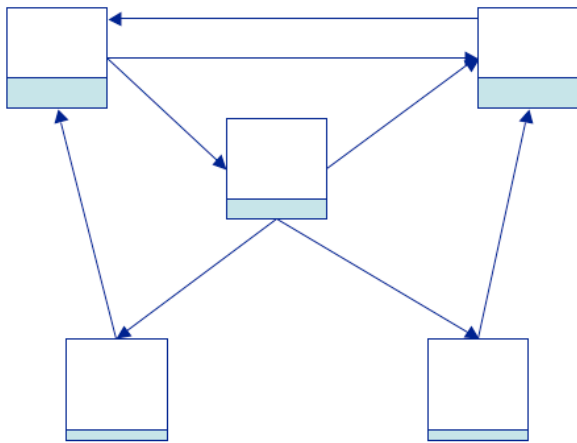
- Each link to a page could be a "vote" for that page
- Sometimes called in-degree counting or link voting
- Query-independent
- Lycos did this in 1996



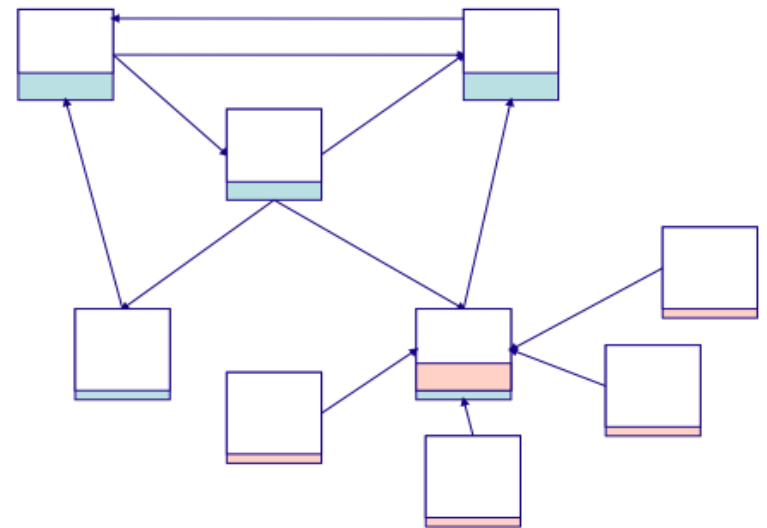
Sites can cheat at in-link counting too



PageRank



Cheating at PageRank



Web Spam

“any deliberate action that is meant to an unjustifiably favorable relevance or importance for some web page, considering the page’s true value”

- Web Spam Taxonomy (Z. Gyöngyi, H. Garcia-Molina, 2005)

Web spam is bad for a search engine

- Result quality degrades
- Index size increases
- Response time increases
- Crawlers can get lost in spam

Web Spam vs SEO

- **Web Spam:**
Deliberately taking advantage of the way a search engine works to increase search ranking
 - Usually at the expense of content
 - Sometimes no real content at all!
- **Search Engine Optimisation:**
Altering a page to make it more accessible to a search engine
 - Not at the expense of content
 - Often helps users when they visit a page

Examples of web spam

- Google webmaster guidelines advise against
 - Hidden text and links
 - "Cloaking" pages
 - Filling pages with keywords
 - Duplicating content
 - Meaningless link exchanges
- If sites are caught by a search engine
 - They can be removed from the index
 - This is very bad for the site!

Examples of SEO

- Search engines are just machines
 - They can't read everything (little flash, JavaScript, etc)
 - Making sure link text is descriptive
 - "campus map" instead of "click here for map"
 - Making sure the right HTML tags are used
 - H1 tags for headers, p tags for paragraphs, etc
 - Making sure pages have meaningful titles
 - Get links from other sites related to your site
- Most search engine optimisation is just good site design anyway

Let's have a look at some pages

0845householdinsurance.co.uk		health-spa-weekends.co.uk		Search the V
Navigation	Auto Insurance Quotes	Car	Navigation	
Auto Insurance Quotes	Direct Line Car Insurance	Categories	HP Computers	With HP's Hot Offers You Get The Latest Computers. See HP A
Car Insurance	We check our prices every week and www.directline.com	Catalogue Shopping	www.hp.com	
Car Insurance Rate		Golf	Home Computer Services	Solutions to Network, Broadband, Hardware & Software Proble
Instant Car Insurance	Cheaper Car Insurance	Online Gambling	www.geekssquad.co.uk	
Categories	Surprisingly unposh prices. Cheaper www.privilege.com	Car Insurance		
Car Insurance	Churchill Car Insurance	Home Insurance	Computer	Save up to £200 on selected Dell PCs from only £269. Buy onli
Content Insurance	You could save up to £140 on our aw www.churchill.com	Mortgage	www.dell.co.uk	
General Insurance	Esure Car Insurance		Quality Computer Desks	Computer workstation desks for home or office. Free UK deliv
Health Insurance	Policies made to suit your needs. Re www.esure.com		www.equip4work.co.uk	
Life Insurance			Home Office Furniture	Stylish desks, storage and chairs. Many exclusive designs.
Travel Insurance	Compare Cheap UK Car Insur		www.cotswoldco.co.uk	
	We compare 50+ car insurers. Find y moneysupermarket.com/insurance			

More spam

all the cheaptravels.co.uk

Travels Hotels Flights Holidays Car Hire Travel Insurance

Search

All the Cheap Travels
Welcome to **alltheCheapTravels.co.uk!** We are a **travel** website where you will find the best system **offer** in terms of the **cheapest hotels, flights, and car hire**.

We are the reference point for all your **travels**. Everyday, we select and provide you the best alternatives in: **holidays, travels insurance, travel agencies, last minute travel, adventure travels, hotels, package holidays, bed & breakfast**... A big world of information and travel services with the cheapest prices that you have ever thought.

Don't forget it!! In **alltheCheapTravels.co.uk** we help you to organise your **travels, hotels, flights and car hire** with the best prices and the best companies.

SpecialOffers
Write your personal email and you will receive more information about the best travel services in hotels, flights, and car hire.

Search Flights Online
Find cheap flights with our Flight Search.

Car Hire from £10 a Day
Find great offers provided by great companies. Here!

Related Links
+ Adventure Travel + Alicante Car Hire + Bed Breakfast + Car Hire + Car Hire France + Car Hire Portugal + Cheap Car Hire + Cheap Flight + Cheap Holiday + Flight to Australia + London Hotel + Paris Hotel + Rome Hotel + Travel Insurance

Travels Hotels Flights Holidays

© 2005 alltheCheapTravels.co.uk

| Contact us | Site Index | About us | Links exchange | Your link here |

Adversarial Information Retrieval

- Detection and removal of web spam
- Developing hard to abuse ranking schemes
- Detecting automated queries
- Detecting click fraud

Pretty much anything where someone is trying to abuse an IR system. It's a cool field!

Some examples of AIR research

- Optimal attack for PageRank (Adali *et al*, 2005)
- Detecting spam pages by content (Ntoulas *et al*, 2006)
- Detecting spam pages by links (Wu, Davison, 2006)

- TrustRank (Gyöngyi *et al*, 2005)
- Detecting cloaking (Wu, Davison, 2005)

Cloaking

What is cloaking?

Sending different pages when a human visits a URL, and when a search engine crawler visits that same URL

Why?

So that users don't spot the spam content

How does it work?

Detecting the "user agent" of the browser
Detecting the IP address of the visitor

Cloaking detection

Download two copies of the page

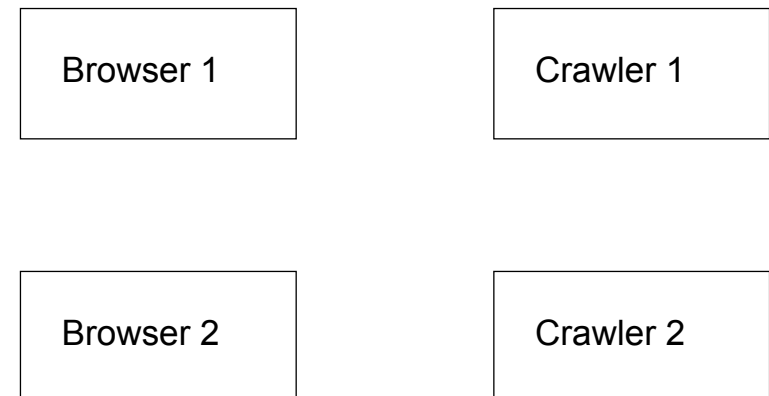
One from a browser
One from a crawler

And then compare them.

Why might this not work very well?

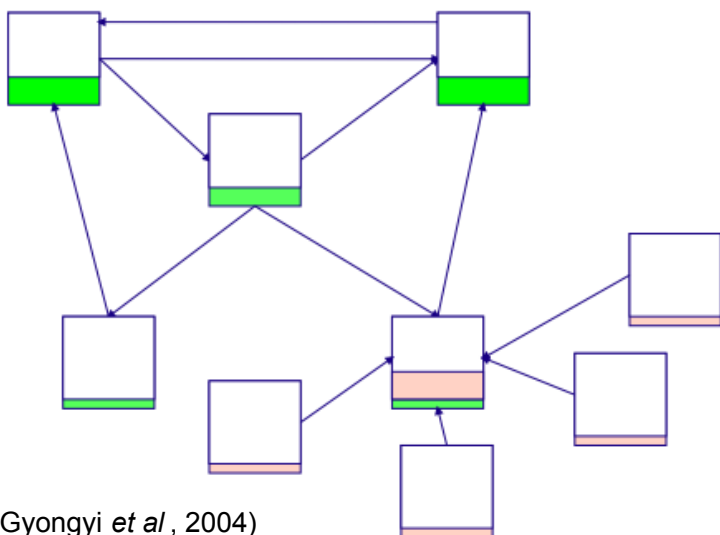
A better way to detect cloaking

Download four copies of a page



(Wu and Davison, 2004)

Detecting PageRank cheating



TrustRank (Gyongyi *et al*, 2004)

Here's an experiment I did

Queens University - Two-panel search tool

Note: The search engine is UK-centric, which means that some well known sites (not hosted in the uk) will not appear in the results. This also means that if you get stuck dreaming up queries, just imagine you're going to visit the UK soon.)

Search for:

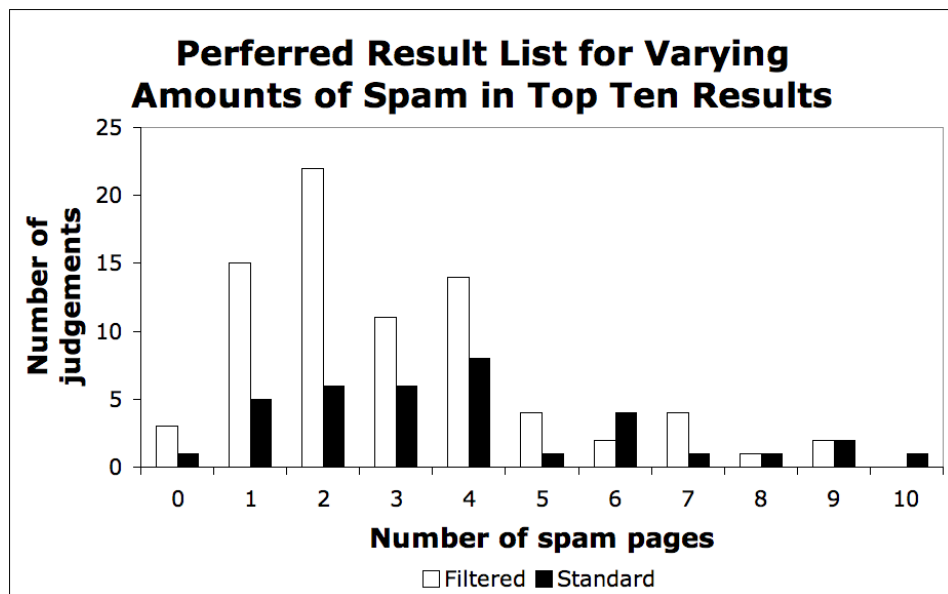
[About this experiment](#)

These are better These are better

No difference

<p>The Queen's University of Belfast - Leading, Inspiring, D... The Queen's University of Belfast - Leading, Inspiring, Delivering http://www.qub.ac.uk/</p> <p>Durham University Search, A-Z, Index, Accessibility, Business, Colleges, Departments, Research, Students, About Us, Home Sunday Times University of the Year 2005 Welcome to Durham University. A supercomputer at Durham University http://www.dur.ac.uk/</p> <p>The Colleges: Contact information Index of addresses and telephone numbers for the colleges in the University of Cambridge. http://www.cam.ac.uk/cambuni/finding/addresses/college.html</p> <p>Money off the things you buy everyday Find Money off the things you buy everyday at money-off.co.uk, the UK's only search engine designed exclusively to find you the best deals on the internet! http://www.money-off.co.uk/</p> <p>Gallerv :: University of Ulster v Queens Sports Day 5 Apr...</p>	<p>The Queen's University of Belfast - Leading, Inspiring, D... The Queen's University of Belfast - Leading, Inspiring, Delivering http://www.qub.ac.uk/</p> <p>Durham University Search, A-Z, Index, Accessibility, Business, Colleges, Departments, Research, Students, About Us, Home Sunday Times University of the Year 2005 Welcome to Durham University. A supercomputer at Durham University http://www.dur.ac.uk/</p> <p>The Colleges: Contact information Index of addresses and telephone numbers for the colleges in the University of Cambridge. http://www.cam.ac.uk/cambuni/finding/addresses/college.html</p> <p>Gallery :: University of Ulster v Queens Sports Day 5 Apr... University of Ulster v Queens Sports Day 5 April 2006 18 images in this album on 2 pages. Gallery: Gallery Album: Public Affairs Images Album: Sports Album: All folders 1 2 640x416 2979x1937 rugby2. http://gallery.publicaffairs.ulster.ac.uk/album541</p> <p>KN Key to Key Skills Continuation Project. Case study: Ou...</p>
--	--

It had some surprising results



Some problems:

Search result quality was bad (many unanswerable queries)

Not enough judgements in >5 spam pages present

Perhaps only a few users care about spam, but they may care more when there's more spam
And we couldn't see this from the experiment

A question not answered:

How **much** does having more spam pages present in the results affect user preference?

Another experiment:

How **much** does having more spam pages present in the results affect user preference?

SPAM #1	SPAM #2
Yahoo #1	Yahoo #2
SPAM #3	Yahoo #4
Yahoo #3	Yahoo #6
SPAM #4	Yahoo #8
Yahoo #5	Yahoo #9
...	...

Search for:

[About this experiment](#)

Left results are...	There is...	Right results are...
<input type="button" value="way better"/> <input type="button" value="better"/> <input type="button" value="a little better"/>	<input type="button" value="no difference"/>	<input type="button" value="a little better"/> <input type="button" value="better"/> <input type="button" value="way better"/>

The Queen (Movie) Official site for the movie The Queen , starring Helen Mirren as Queen Elizabeth II in the immediate aftermath of the Princess of Wales' death in August 1997, as a grieving nation waits to see what its leaders will do. Also starring	The Queen (Film) - Wikipedia User-edited article about the movie The Queen , the 2006 Academy Award-winning British film directed by Stephen Frears. http://en.wikipedia.org/wiki/The_Queen_(film)
--	---

Shameless plug

I'm running an experiment at the moment

it's at:

<http://retrieval.anu.edu.au/pub/>

It takes about 30 minutes, and we'll give you a Dendy movie ticket for your time!

In summary

- High rank in search is important for e-commerce
- Search ranking has had to change because of cheating
 - It might never be perfect
- Cheating and optimising are not the same
 - Web spam vs SEO
- Adversarial IR research
 - to detect spam
 - to prevent spam from affecting results
 - to improve search quality