Topic 7: PageRank and Challenges in Web Search

COMP3009J: Information Retrieval

Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science
Beijing Dublin International College

Introduction

- Using the models we have seen so far (e.g. Vector Space, BM35), each document is treated completely **separately** to other documents.
- The ranking score is calculated based entirely on the **content** of the document itself.
- This does not take into account everything a human judge would use in deciding whether or not a document is suitable for returning in response to a particular query.
- Some of these additional factors are very difficult for a computer algorithm to figure out, some are impossible.

Introduction

- Some questions you might ask yourself are:
  - Though it contains the terms contained in the query, does the document actually satisfy the information need that is being expressed?
  - Is the document well-written and understandable?
  - How important and influential is this document?
- In most cases, these are not questions that a computer algorithm can answer, which is why we use human judges when evaluating IR systems.

Document Importance

- There are, however, some areas where it may be possible to estimate how **important** a document is.
- For instance, in academic writing a paper will generally cite other works that influenced it.
- Some organisations use "citation counting" to measure the influence of a piece of work.
- The intuition is that the more times your paper is cited by others, the more influence it must have had in its field.
- Also, if a paper appears in an influential journal, it is likely to be important.

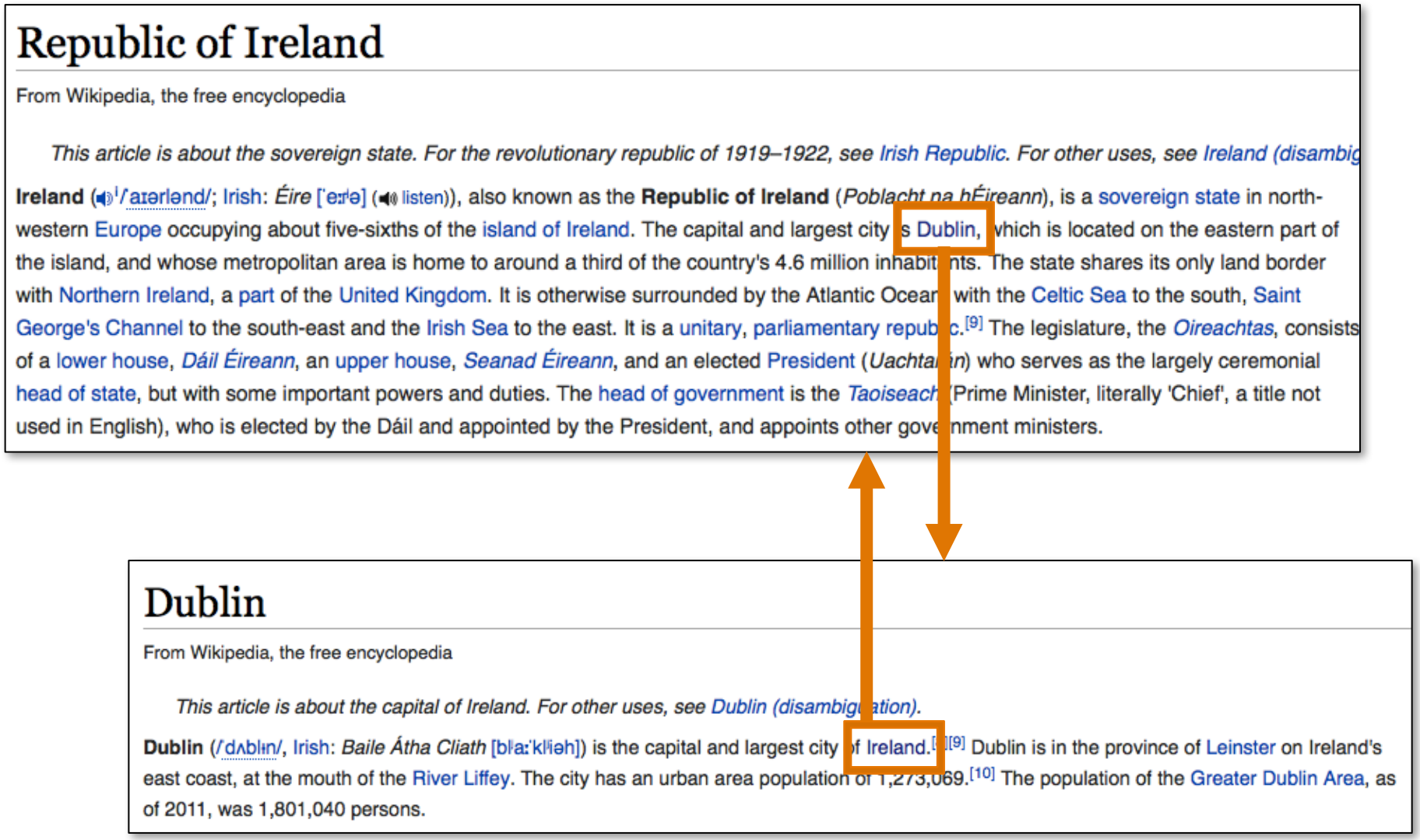
Citation Counts and the Web

- The World Wide Web contains the largest collection of documents in existence, and is a most suitable forum for IR.
- Unlike in traditional IR systems, the documents available on the web are **not standalone**.
- Because of the presence of **hypertext** (i.e. links) in documents on the web, there is an **interconnection** between these documents.
- A number of researchers have attempted to apply principles similar to citation counts to web pages.
- This operates on the assumption that a page that is linked to from many other pages is important and this should be reflected in search results.

Citation Counts and the Web

- There are, however, a few notable differences between academic publishing and web publishing that must be taken into account:
  - **Circular references** - in academic publishing, a paper can only cite a paper that has already been published. Later papers are not cited by earlier ones. On the web, two pages can link to one another.

Circular References



Republic of Ireland

From Wikipedia, the free encyclopedia

This article is about the sovereign state. For the revolutionary republic of 1919–1922, see Irish Republic. For other uses, see Ireland (disambig

Ireland (/ˈaɪərlənd/; Irish: *Éire* [ˈeːɾʲə] (listen)), also known as the **Republic of Ireland** (*Poblacht na hÉireann*), is a sovereign state in north-western Europe occupying about five-sixths of the island of Ireland. The capital and largest city is Dublin, which is located on the eastern part of the island, and whose metropolitan area is home to around a third of the country's 4.6 million inhabitants. The state shares its only land border with Northern Ireland, a part of the United Kingdom. It is otherwise surrounded by the Atlantic Ocean with the Celtic Sea to the south, Saint George's Channel to the south-east and the Irish Sea to the east. It is a unitary, parliamentary republic.[9] The legislature, the *Oireachtas*, consists of a lower house, *Dáil Éireann*, an upper house, *Seanad Éireann*, and an elected President (*Uachtarán*) who serves as the largely ceremonial head of state, but with some important powers and duties. The head of government is the *Taoiseach* (Prime Minister, literally 'Chief', a title not used in English), who is elected by the Dáil and appointed by the President, and appoints other government ministers.

Dublin

From Wikipedia, the free encyclopedia

This article is about the capital of Ireland. For other uses, see Dublin (disambiguation).

Dublin (/ˈdʌblɪn/, Irish: *Baile Átha Cliath* [bʲlaːˈklʲiəh]) is the capital and largest city of Ireland.[9] Dublin is in the province of Leinster on Ireland's east coast, at the mouth of the River Liffey. The city has an urban area population of 1,273,069.[10] The population of the Greater Dublin Area, as of 2011, was 1,801,040 persons.

Citation Counts and the Web

- **Quality control** - in academic publishing, a paper is peer-reviewed before it is approved for publication. Thus some quality is maintained in the papers that include citations. On the web, anybody can publish material and link to other documents. It would be a trivial task to write a program that would generate hundreds or thousands of pages containing links to somewhere else.

PageRank

- In 1998, Sergey Brin and Larry Page published the PageRank algorithm, which works on similar principles*.
- This later went on to be a core element in the success of the Google search engine.
- The algorithm itself has been modified since (secretly: Google rarely reveal anything about the search engine's inner workings anymore) to avoid situations where it was being exploited by malicious publishers.
- However, the way it functions has largely remained unchanged.

* S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine"

PageRank

- Again, it is based on the premise that documents that are linked to by many other documents are important, and should receive a boost in search engine rankings as a result.
- A document will tend to have a high PageRank score if:
  - It is linked to by many documents and/or
  - It is linked to by documents that themselves have a high PageRank.

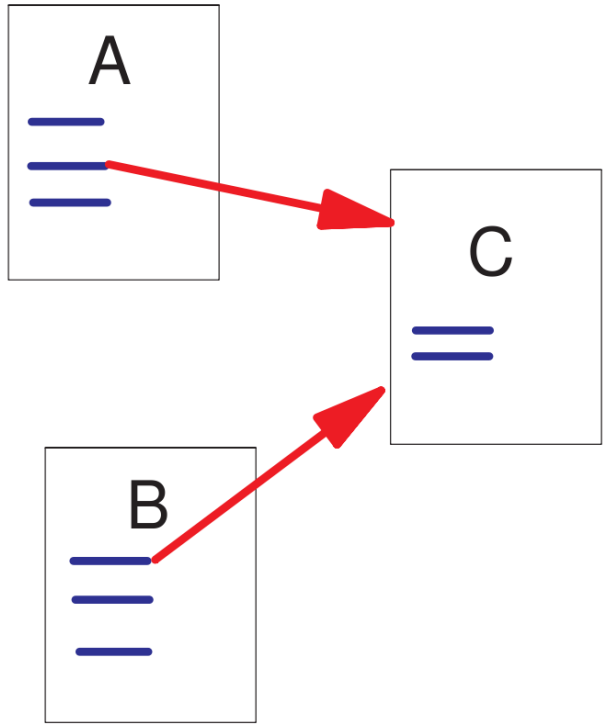
PageRank - Link Structure

The web is made up of HTML pages that are connected using hyperlinks.

We can think of this as a **directed graph**.

Pages A, B, C are vertices in the graph.

A and B have links (edges) to Page C.



PageRank - Link Structure

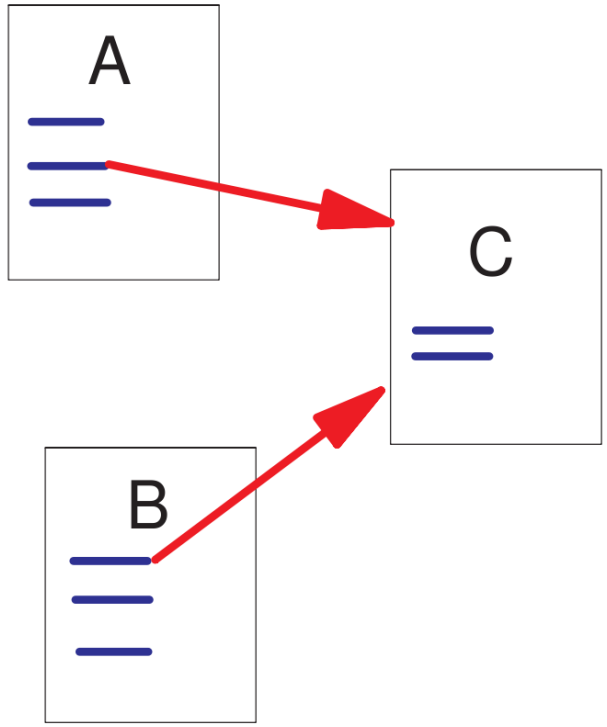
In the 1998 crawl, Brin and Page had 150 million vertices (pages) and 1.7 billion edges (links).

We say that A and B are **backlinks** of C.

We say that A and B have **outlinks** to C.

A document will have high PageRank if it has:

- Many backlinks
- Backlinks with high PageRank



PageRank: Simplified Version

- At a basic level, PageRank works using a formula similar to the following…
- $R(u) = \sum_{v \in B_u} \frac{R(v)}{N_v}$
  - R(u) is the PageRank score for document u.
  - B_u is the set of all backlinks of document u.
  - R(v) is the PageRank score for document v.
  - N_v is the number of outlinks in document v.

PageRank: Simplified Version

What does this mean?

A document contributes $\frac{R(v)}{N_v}$ to the PageRank of each document it links to.

That is, if a document links to 4 pages, its contribution to each of those pages is ¼ of its own PageRank.

- $R(u) = \sum_{v \in B_u} \frac{R(v)}{N_v}$
  - R(u) is the PageRank score for document u.
  - B_u is the set of all backlinks of document u.
  - R(v) is the PageRank score for document v.
  - N_v is the number of outlinks in document v.

PageRank: Simplified Version

So if a backlink has high PageRank (and few outlinks), this will have a beneficial effect.

A document's final PageRank score is the sum of each of these contributions from backlinks.

The more backlinks a document has, the higher its PageRank will be.

- $R(u) = \sum_{v \in B_u} \frac{R(v)}{N_v}$
  - R(u) is the PageRank score for document u.
  - B_u is the set of all backlinks of document u.
  - R(v) is the PageRank score for document v.
  - N_v is the number of outlinks in document v.

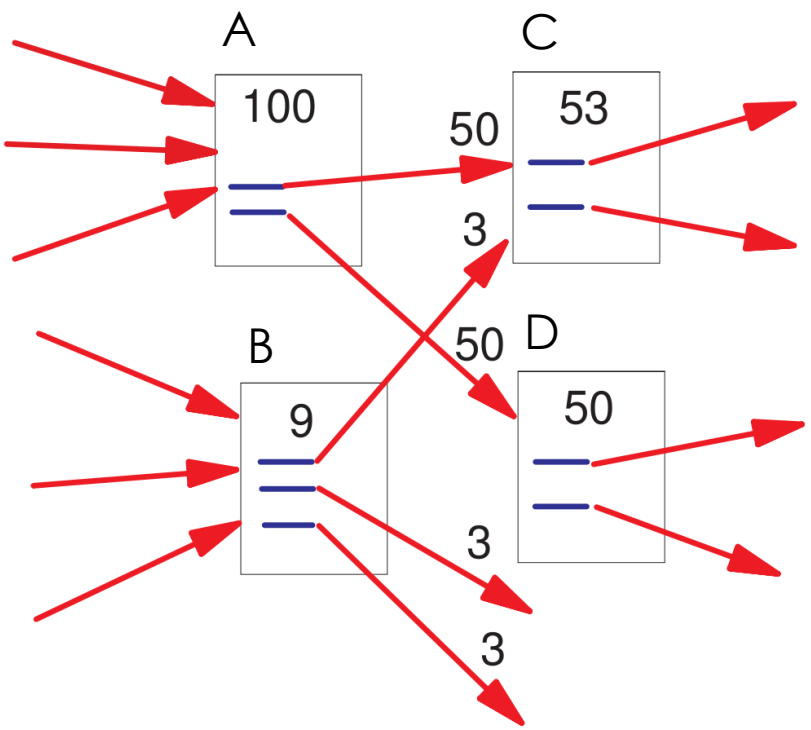
PageRank - Simplified Version

Document A has PageRank of 100 and 2 outlinks:

- It sends 50 to C and 50 to D

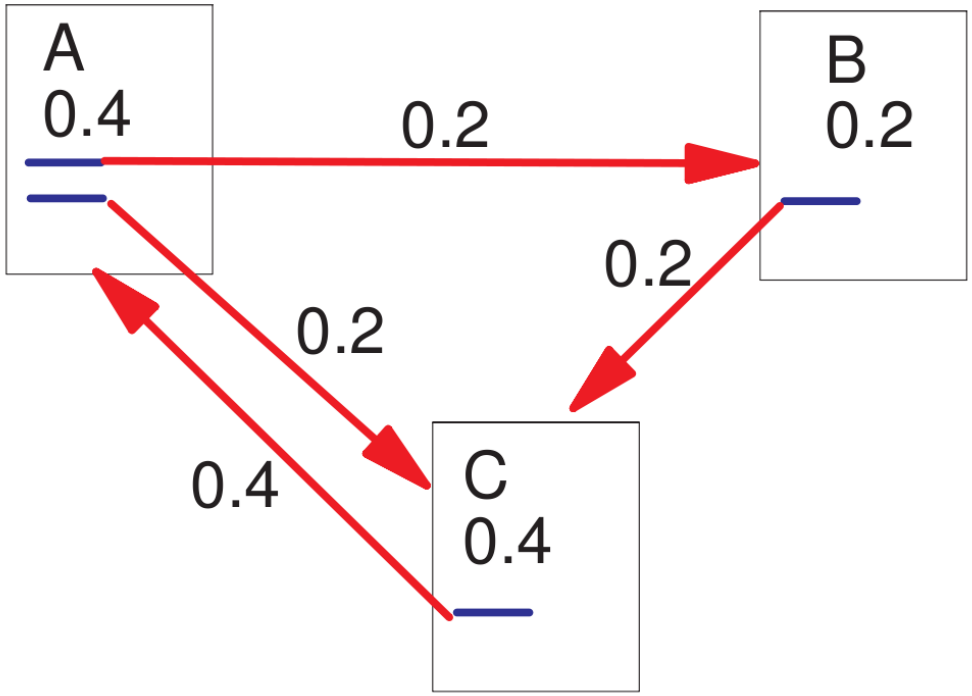
Document B has PageRank of 9 and 3 outlinks:

- It contributes 3 to the PageRank of each of its outlinks (including C)



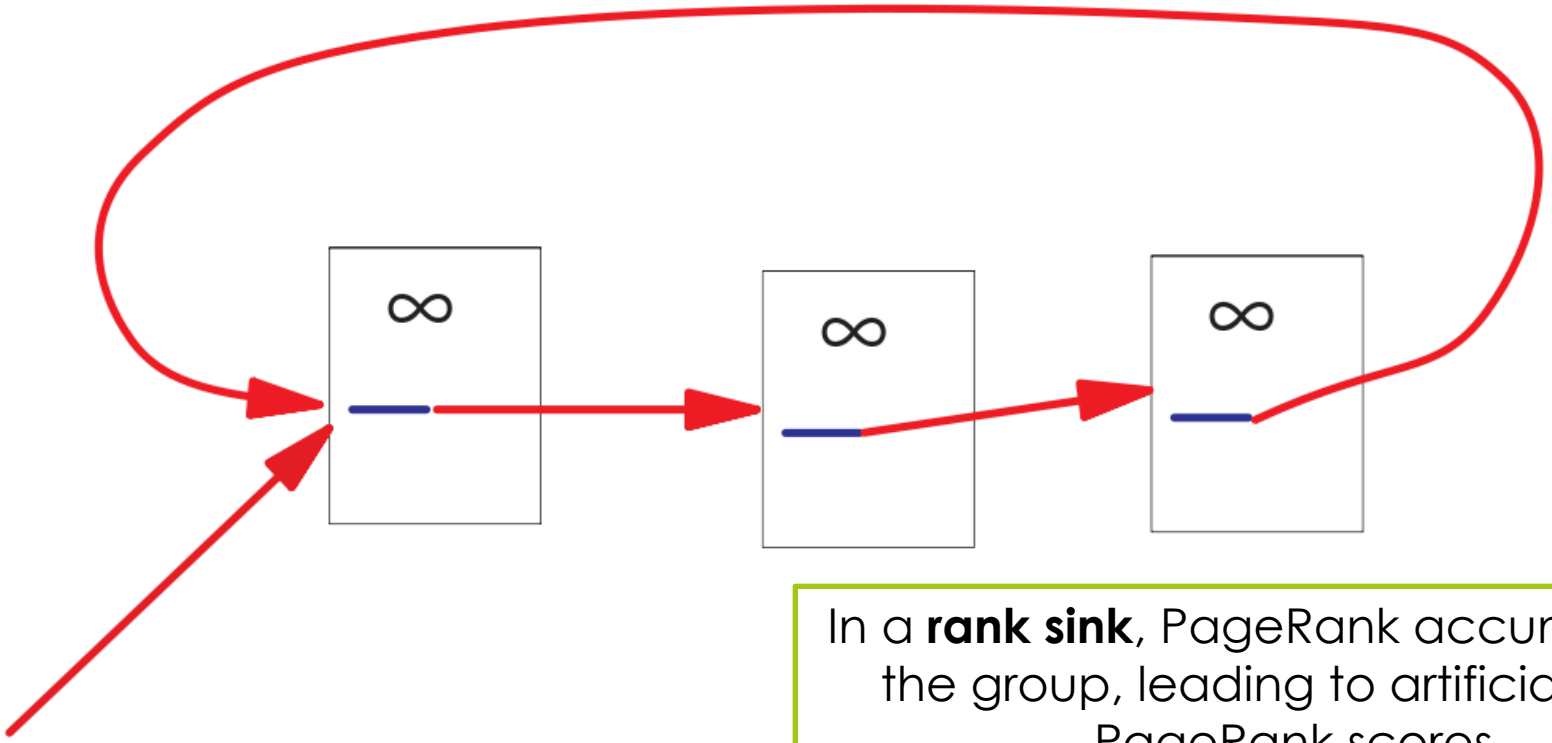
PageRank - Simplified Version

- Question: If we need PageRank to calculate PageRank, where does the initial PageRank come from?
  - At the beginning, we can give an **arbitrary score** to every document.
  - The formula we have seen can then be used to calculate new scores.
  - These **continue to be recalculated** until the scores **converge** (i.e. calculating again does not change the scores, or changes them very little).



PageRank - Simplified Version

This image shows a stable state: no matter how many times PageRank is recalculated, the scores for A, B and C will always be the same.



In a **rank sink**, PageRank accumulates in the group, leading to artificially high PageRank scores.

PageRank - Problems

Although this simple example illustrates how PageRank works, it does not deal with certain situations very well.

One such situation is a **rank sink** which refers to a group of pages that have at least one backlink and link to one another, but do not link to anywhere else outside the group.

Combating Rank Sinks

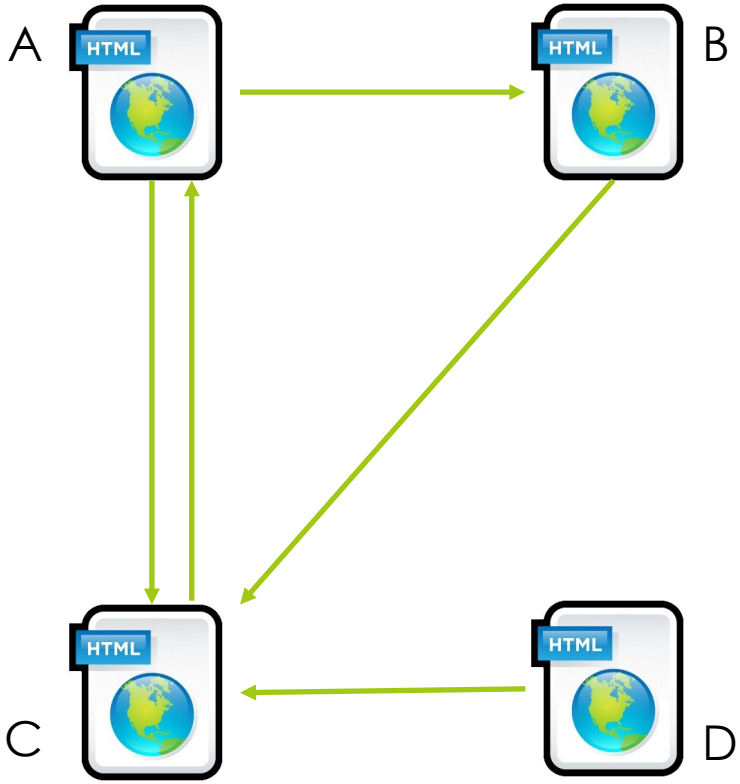
- To combat this type of situation, a new equation is used:
- $R(u) = (1 - d) + d \times \sum_{v \in B_u} \frac{R(v)}{N_v}$
- The formula is the very similar to the one we have seen before.
- The difference is the addition of a **damping factor** (d), which ensures that not all of a document's PageRank is passed on via its outlinks.

PageRank - No Backlinks

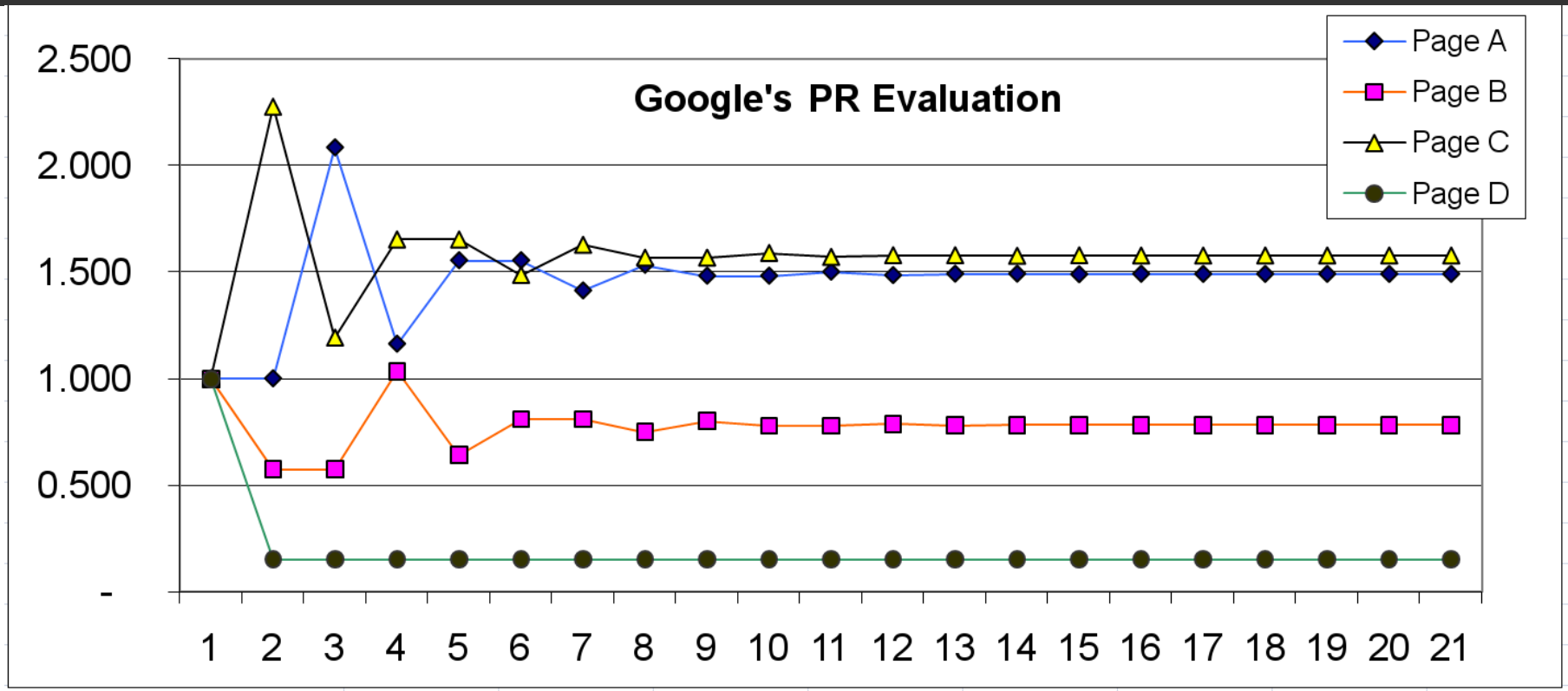
- In the original formula, the only source of PageRank for a document is from its backlinks.
- This meant that a document with no backlinks would have a PageRank of zero.
- This is perfectly acceptable when considering the importance of that document itself.
- However, this also means that it would not contribute anything to the PageRank of documents it ranks to.
- With the modified formula, a document with no backlinks has a PageRank of $(1-d)$ to contribute to the documents it links to.

PageRank - Example: 4 pages

- Consider the following simple page structure:
  - Page A: links to B and C
  - Page B: links to C
  - Page C: links to A
  - Page D: links to C
- Using a damping factor of 0.85 (which Google appears to use), calculate the PageRank of each document.

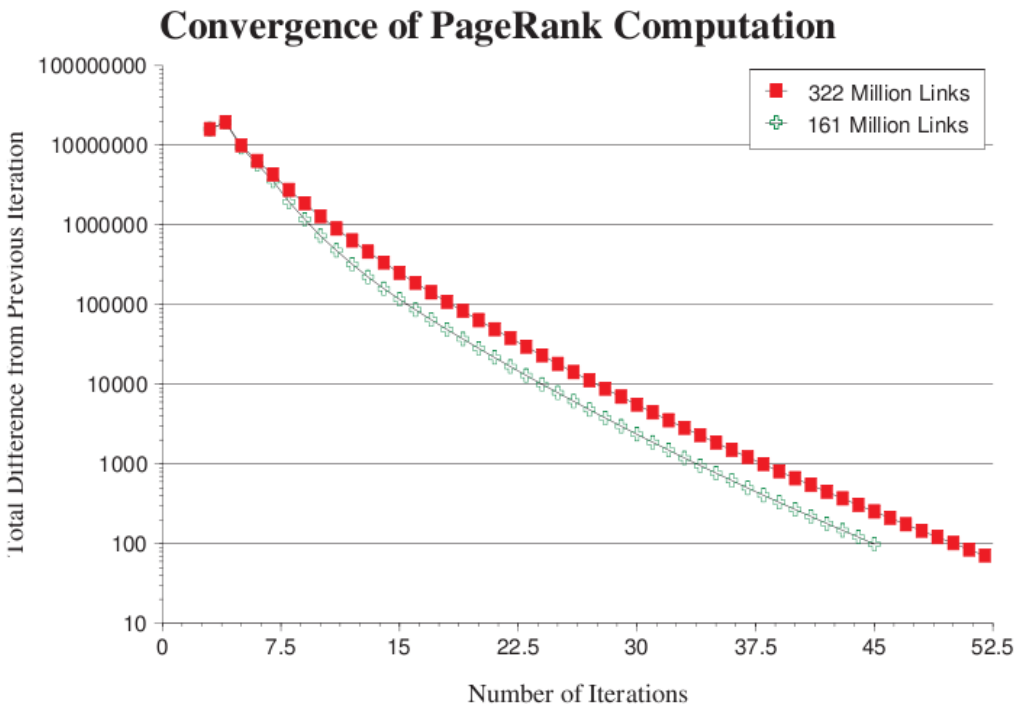


PageRank - Example: 4 Pages



PageRank - Example: 4 Pages

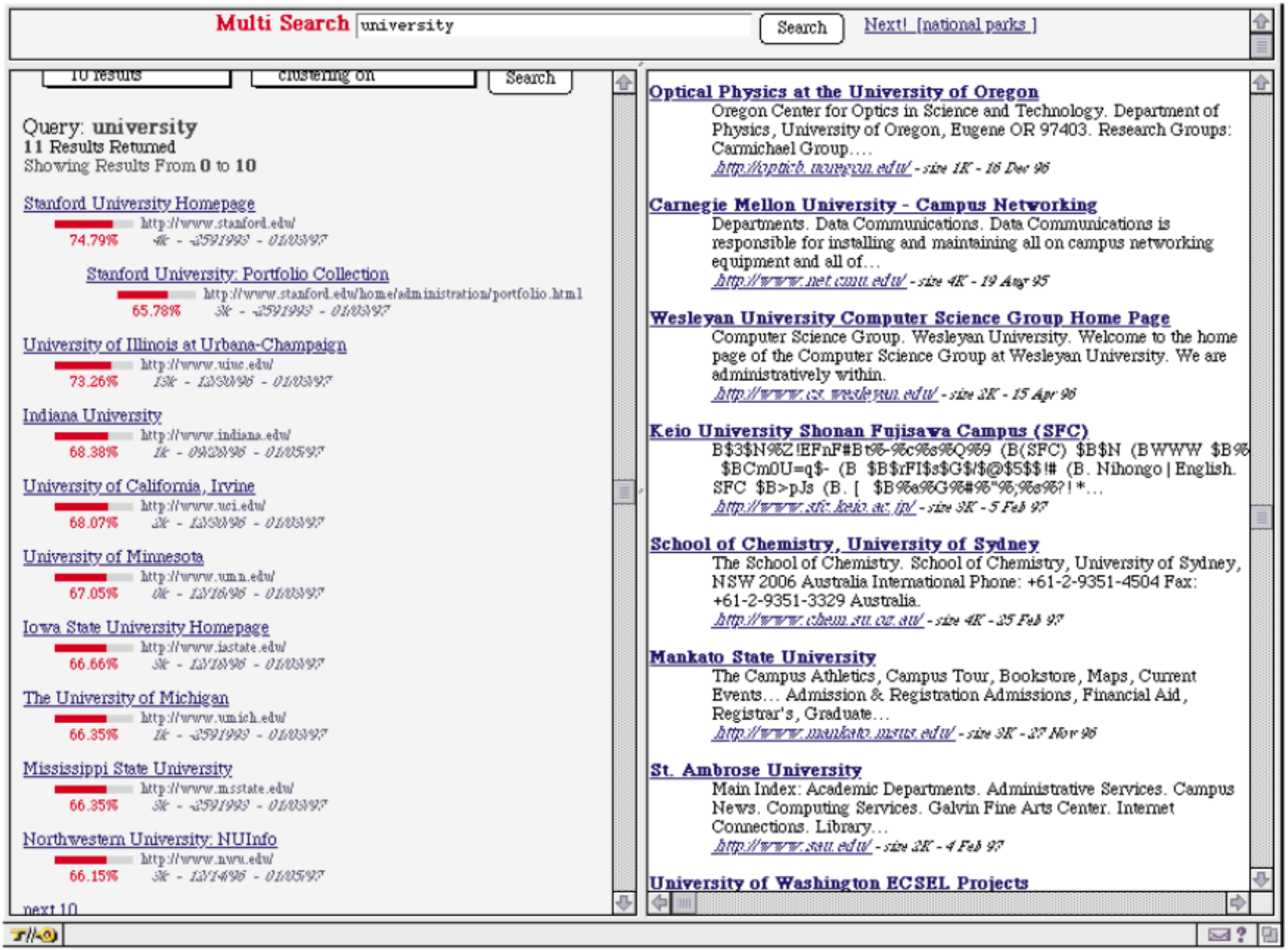
- With this simple system, the PageRank scores have **converged** perfectly after 20 iterations.
- Even after 8 or 9 iterations, the values are very similar to their ultimate values.
- Brin and Page were used a database of 322 million links and believed that the convergence had reached a reasonable level after 52 iterations.
- Calculations on half the data took 45 iterations to get to the same stage.
- This suggests that PageRank scales very well to very large scale data collections.



PageRank - Convergence

Doubling the size of the document collection does not double the time taken to converge.

In fact, the increase in the number of iterations required is very small (52 vs. 45): very efficient for larger collections.


Multi Search
university
Search
Next! [national parks]
10 results
clustering on
Search
Query: university
11 Results Returned
Showing Results From 0 to 10
Stanford University Homepage
http://www.stanford.edu/
74.79%
Stanford University: Portfolio Collection
http://www.stanford.edu/home/administration/portfolio.html
65.78%
University of Illinois at Urbana-Champaign
http://www.uiuc.edu/
73.26%
Indiana University
http://www.indiana.edu/
68.38%
University of California, Irvine
http://www.uci.edu/
68.07%
University of Minnesota
http://www.umn.edu/
67.05%
Iowa State University Homepage
http://www.iastate.edu/
66.66%
The University of Michigan
http://www.umich.edu/
66.35%
Mississippi State University
http://www.msstate.edu/
66.35%
Northwestern University: NUInfo
http://www.nwu.edu/
66.15%
next 10
Optical Physics at the University of Oregon
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group....
http://optich.uoregon.edu/ - size 1K - 16 Dec 96
Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
http://www.net.cmu.edu/ - size 4K - 19 Aug 95
Wesleyan University Computer Science Group Home Page
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
http://www.cs.wesleyan.edu/ - size 3K - 15 Apr 96
Keio University Shonan Fujisawa Campus (SFC)
B3N%Z!EFnF#Bt%-%c%s%Q%9 (B(SFC) BN (BWWW $B%
$BCm0U=q$- (B BrFIsG$/$@$5$$!# (B. Nihongo | English.
SFC $B>pJs (B. [$B%a%G%#%"%;%s%?! *...
http://www.sfc.keio.ac.jp/ - size 3K - 5 Feb 97
School of Chemistry, University of Sydney
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
http://www.chem.su.oz.au/ - size 4K - 25 Feb 97
Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
http://www.mankato.msus.edu/ - size 3K - 27 Nov 96
St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
http://www.sau.edu/ - size 2K - 4 Feb 97
University of Washington ECSEL Projects

PageRank - Consequences

- Google's use of PageRank to help rank documents led to them dominating web search in the English-speaking world, which they continue to do today.
- Other IR techniques are also used (full-text search, title search, proximity search etc.) and a fusion process is used to merge the results of these different kinds of search.
- Specific details about how exactly Google does its searching are not generally available anymore, such is the competitive nature of the online search business.
  - Other search companies have their own version of PageRank.

Challenges in Web Search

Introduction

- Many of the techniques we have look at, both for performing and evaluating Information Retrieval were developed before the World Wide Web (WWW) became popular.
- Nowadays, the WWW is the biggest source of information in existence and web search has become an extremely important aspect of using the web.
- With the shift of IR towards this new environment, a number of significant challenges have presented themselves. We will look at a number of these.

Finding Information

How do we find information?

- **Web Crawlers** or **Web Spiders** are programs that automatically look for documents on the web that will be included in the index so that users can find them when searching.
- They do this by **extracting the hyperlinks** from each page they find, which will in turn point them to new pages to index.
- Even the decision on what policy to use for following hyperlinks is a subject of debate: one study claims that using a Breadth-First Search strategy will result in finding higher-quality pages (as measured by PageRank)*.
- It is also considered to be more polite to use a Breadth-First Search, as it reduces load on servers.

* M. Najork and J. Weiner, Breadth-First Search Crawling Yields High-Quality Pages, 2000

How do we find information?

- Breadth-first spider search:
  - Manually add a list of "seed" URLs to the spider's list of pages to visit
  - While there are still URLs in the queue of pages to visit
    - Remove the first URL from the queue of pages to visit and download the page
    - Extract the links from the document
    - Add these links to the queue of pages to visit if they haven't already been visited and are not already in the queue.
- A depth-first search is the same, except you use a stack instead of a queue.

How big is the web?

- Google state that their index contains “hundreds of billions of webpages”.
- Technically, the size of the web is probably infinite, as dynamic pages give different results depending on the information they’re given:
  - Calendar applications can typically show data for an indefinite amount of time into the future.
  - Search engines themselves will give a different page in response to every different query they receive.

* http://www.google.com/insidesearch/howsearchworks/thestory/

Finding information: challenges

Pages designed to frustrate attempts at automated access (e.g. through the use of CAPTCHAs).

A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is typically an image containing text that computers find difficult to ready.

They are frequently to be seen when signing up for accounts with various web sites.

An automatic crawler can't access pages that are guarded by CAPTCHAs.

How big is the web?

- The **Deep Web*** (or invisible web or hidden web) refers to the fact that web crawlers cannot easily gain access to particular online resources.
  - Some pages or groups of pages may not have any links from elsewhere. Since web crawlers work by following links, it may not be possible to find these pages.
- Content served via AJAX/Asynchronous JavaScript or Flash can be hard for crawlers to index. AJAX content is fetched by JavaScript code running within a web page and is not accessible just using a URL like traditional content.

* M. K. Bergman, The Deep Web: Surfacing Hidden Value 2001

Polite Crawlers: The Robots Exclusion Standard

- The Robots Exclusion Standard is a method of asking web crawlers not to index certain portions of a web site.
- It is a **voluntary standard**, so there are no technical barriers stopping crawlers that ignore the standard from accessing the pages in question.
- Whenever a well-behaved crawler visits a web server, it will firstly check the root directory for a file called `robots.txt` (i.e. `http://www.example.com/robots.txt`).
- This is a plain text file that contains a list of resources that the crawler should not visit.
- Different lists can be provided for different crawlers (e.g. if you wanted a particular page to be searchable on Yahoo! but not on Google).

Polite Crawlers: noindex and nofollow

- Ignore pages that have `<meta name="robots" content="noindex" />`.
- Don't follow any links from pages that have `<meta name="robots" content="nofollow" />`.
- Individual links can be marked `nofollow` also:

```
<a href="http://example.com">Anchor Text</a>
<a href="http://example.com"
rel="nofollow">Anchor Text</a>
```

- The exact treatment of this varies for different search engines.

Non-static documents

- Another issue is the fact that the web is a **dynamic environment.**
- New pages are **added constantly** and many existing pages may **change frequently**.
- Even if a crawler has visited a large number of documents, it must still be **revisited** so the index does not get out-of-date.
  - For instance, news sites will be constantly changing and are typically re-indexed on a regular basis.
- On the other hand, there are huge numbers of pages that are essentially static over long periods of time (mailing list archives, individual news articles, etc.)
- A policy must be implemented that decides which pages should be re-indexed on a regular basis and which need not be.

Adversarial Information Retrieval

Adversarial IR

- One key difference between traditional IR and Web Search is that traditionally, works were published without IR systems in mind.
- An author writing a book or a newspaper, magazine or technical article would only concentrate on the work they were producing.
- The publishing of web pages is very different. An entire industry known as **Search Engine Optimisation (SEO)** has emerged.
- SEO involves taking steps to get your page high up in search engine rankings.
  - Higher rankings → more clicks → more money.
- Search engines need to differentiate between the **real** content of pages and the SEO content that is essentially trying to cheat the system.

Adversarial IR: Meta Tags

- HTML allows for page authors to describe the contents of their pages by using <meta> tags.
- These were designed to allow such information as a page's description, keywords and author to be specified in a standardised way.

```
<head>
  <title>Stamp Collecting World</title>
  <meta name="description" content="Everything you wanted to
     know about stamps, from prices to history." />
  <meta name="keywords" content="stamps, stamp
     collecting, stamp history, prices, stamps for sale" />
</head>
```

Adversarial IR: Meta Tags

- Some early search engines did not have the processing power to perform full-text search.
- Even for many that did, they initially gave a high weight to terms appearing in the description and keywords of a page.
- Theoretically, this would succinctly describe the key contents of their web sites.
- In reality, this was open to abuse by dishonest webmaster that would attempt to target certain search queries.
- A common example was people including the word “free” amongst their keywords, as many searchers would use this when they began searching for a particular service.
- Because of this behaviour, meta tags are mostly ignored by search engines nowadays.

Adversarial IR: Hidden Text

- Consider this example:

```
<p>A site for buying computer components</p>
<p><font color="#ffffff"> computer parts,
computer hardware, memory </font></p>
```

Adversarial IR: Hidden Text

- This is an example of manipulation using **hidden text**.
- The second paragraph in the previous slide will not be visible to the human eye, as the text is the same colour as the background.
- This is a method of changing the contents of the page without negatively affecting what a user sees.
- It has two main functions:

1. Add extra keywords for a search engine to pick up that aren't already contained in the content of the page.
2. Increase the term frequency of important search terms.

Another example.

- `<p class="class1">A site for buying computer components</p>`
- `<p class="class2">computer parts, computer hardware, memory</p>`

Adversarial IR: Hidden Text

- Here, the manipulation is not as obvious, as you'd have to check the Cascading Style Sheet class "class2" to see if there are any color changes defined there.
- You would also have to check any other elements that the second paragraph is contained in.
- This is not a trivial task for a crawler to carry out (remember that a crawler needs to operate very quickly to build up a sufficiently large index).
- Many search engines will remove your site from its index if this type of activity is found.

Adversarial IR: Cloaking

- **Cloaking** is the term given to the practice of serving different content to web crawlers than to users.
- A well-behaved crawler will identify itself to a web server when it visits, by way of a "user agent string".
- A server (or a script running on it) can easily be configured to give different content based on this user agent.
  - This technique is often used legitimately to show a mobile-optimised site for phone users.
- The idea is to serve highly-targeted text to the crawler, so as to gain a high ranking on particular searchers.
  - Regular users would be served the normal content.

Adversarial IR: Sneaky JavaScript

- Web crawlers (and old browsers) are not capable of executing JavaScript code.
- In this situation, there is a <noscript> that allows you to specify text content for users who are unable to run JavaScript (this is not so common anymore) or those who are unwilling to run it.
- This text is what the web crawler will see as the content of the page.
- As a result, this can be exploited in the same way as cloaking: web crawlers index the contents of the "noscript" tag, whereas real users will get redirected (via JavaScript) to a different page without them noticing.

Adversarial IR: Exploiting PageRank

- Since a high PageRank (or similar) results in high rankings in search results, there is a commercial benefit to having a high PageRank score.
- This has led to the creation of "link farms": sites (or networks of sites) with the sole purpose of linking to other sites so as to increase their PageRank.
- Each page has a small contribution to overall PageRank, so it is easy to create huge numbers of pages consisting solely of links.
- Additionally, pages that benefited from this PageRank would link back to the farm so as to boost the PageRank of the farm.
- Often, this link would be hidden by being the same colour as the rest of the page.
- Some owners of sites with high PageRank built up in this way would charge a fee from anybody who wanted to be linked to from the farm.

Adversarial IR: Exploiting PageRank

- Another approach is "comment spam", where pages that allow third-party commenting are abused to try and boost PageRank.
  - Advertising links are posted as comments in hundreds or thousands of places, to gain PageRank from each.
- This is why the "nofollow" attribute for links was created. This would means that search engines do not take the link into account when calculating PageRank. This allows site owners to have more control over which links from their site add to the PageRank of another page.

Summary

- The World Wide Web (WWW) has caused many challenges for the developers of IR systems.
- In particular:
  - PageRank has been successful in measuring the importance of web pages.
  - The web is **huge**, and finding things can be difficult.
  - People frequently try to **exploit** features of the system's ranking algorithm to promote particular pages, at the expense of users. This is known as **adversarial IR.**