PageRank

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, BM25), 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 easily, 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.

REFERENCES

Bloom, B. H. (1970). Space/Time Trade-offs in Hash Coding with Allowable Errors. *Communications of the ACM*, *13*(7), 422–426.

Breitinger, F., & Baier, H. (2012). Similarity Preserving Hashing: Eligible Properties and a New Algorithm MRSH-v2. In *International conference on digital forensics and cyber crime* (pp. 167–182). Springer.

Breitinger, F., Baier, H., & White, D. (2014). On the Database Lookup Problem of Approximate Matching. *Digital Investigation*, *11*, S1–S9. doi: 10.1016/j.diin.2014.03.001

Breitinger, F., Guttman, B., McCarrin, M., Roussev, V., & White, D. (2014). Approximate Matching: Definition and Terminology. *NIST Special Publication*, *800*, 168.

Breitinger, F., Rathgeb, C., & Baier, H. (2014, sep). An Efficient Similarity Digests Database Lookup - A Logarithmic Divide & Conquer Approach. *Journal of Digital Forensics, Security and Law*, *9*(2), 155–166.

Broder, A. Z. (1997). On the Resemblance and Containment of Documents. In *Compression and complexity of sequences 1997. proceedings* (pp. 21–29). doi: 10.1109/SEQUEN.1997.666900

Casey, E., Ferraro, M., & Nguyen, L. (2009). Investigation Delayed is Justice Denied: Proposals for Expediting Forensic Examinations of Digital Evidence. *Journal of forensic sciences*, *54*(6), 1353–1364.

de Braekt, R. I., Le-Khac, N. A., Farina, J., Scanlon, M., & Kechadi, T. (2016, April). Increasing Digital Investigator Availability Through Efficient Workflow Management and Automation. In *4th international symposium on digital forensic and security (isdfs)* (p. 68-73). doi: 10.1109/ISDFS.2016.7473520

Gupta, J. N., Kalaimannan, E., & Yoo, S.-M. (2016). A Heuristic for Maximizing Investigation Effectiveness of Digital Forensic Cases Involving Multiple Investigators. *Computers & Operations Research*, *69*, 1–9. doi: 10.1016/j.cor.2015.11.003

Gupta, V., & Breitinger, F. (2015). How cuckoo filter can improve existing approximate matching techniques. In J. James & F. Breitinger (Eds.), *Digital forensics and cyber crime* (Vol. 157, p. 39-52). Springer International Publishing. doi: 10.1007/978-3-319-25512-5_4

Harichandran, V. S., Breitinger, F., & Baggili, I. (2016). Bytewise Approximate Matching: The Good, The Bad, and The Unknown. *The Journal of Digital Forensics, Security and Law: JDFSL*, *11*(2), 59.

James, J. I., & Gladyshev, P. (2015). Automated Inference of Past Action Instances in Digital Investigations. *International Journal of Information Security*, *14*(3), 249–261. doi: 10.1007/s10207-014-0249-6

Kornblum, J. (2006). Identifying Identical Files Using Context Triggered Piecewise Hashing. *Digital investigation*, *3*, 91–97. doi: 10.1016/j.diin.2006.06.015

Lillis, D., Becker, B., O'Sullivan, T., & Scanlon, M. (2016). Current Challenges and Future Research Areas for Digital Forensic Investigation. In *11th ADFSL Conference on Digital Forensics, Security and Law (CDFSL 2016)*. Daytona Beach, FL, USA: ADFSL. doi: 10.13140/RG.2.2.34898.76489

Lillis, D., Breitinger, F., & Scanlon, M. (2017). Expediting MRSH-v2 Approximate Matching with Hierarchical Bloom Filter Trees. In *9th EAI International Conference on Digital Forensics and Cyber Crime (ICDF2C 2017)*. Prague, Czech Republic.

Oliver, J., Cheng, C., & Chen, Y. (2013). TLSH–A Locality Sensitive Hash. In *Fourth Cybercrime and Trustworthy Computing Workshop (CTC), 2013* (pp. 7–13). doi: 10.1109/CTC.2013.9

Quick, D., & Choo, K.-K. R. (2014). Impacts of Increasing Volume of Digital Forensic Data: A Survey and Future Research Challenges. *Digital Investigation*, *11*(4), 273–294. doi: 10.1016/j.diin.2014.09.002

Rogers, M. K., Goldman, J., Mislan, R., Wedge, T., & Debrota, S. (2006). Computer Foren-

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* [ˈeːrʲə] (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* [bʲlʲaːˈklʲiəh]) is the capital and largest city of Ireland.[8][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 to measure the importance of web pages*.
- 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 core of how it functions has largely remained unchanged.

* S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer networks and ISDN systems*, *30*(1-7), 107-117. 1998.

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.

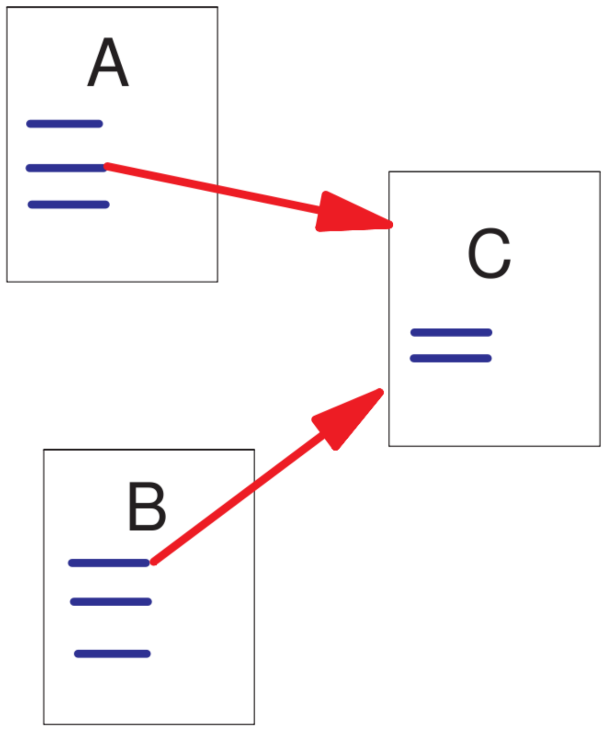
PageRank - Link Structure

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

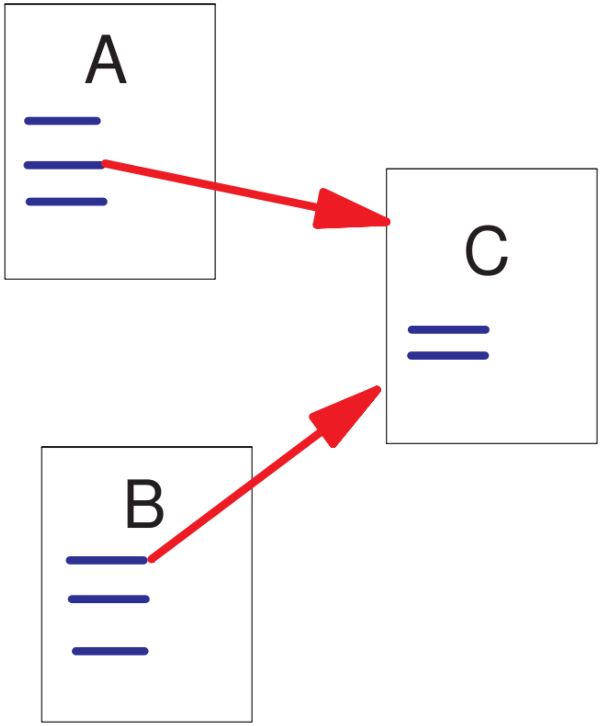
In their 1998 crawl, Brin and Page had 150 million vertices (pages) and 1.7 billion edges (links).

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.

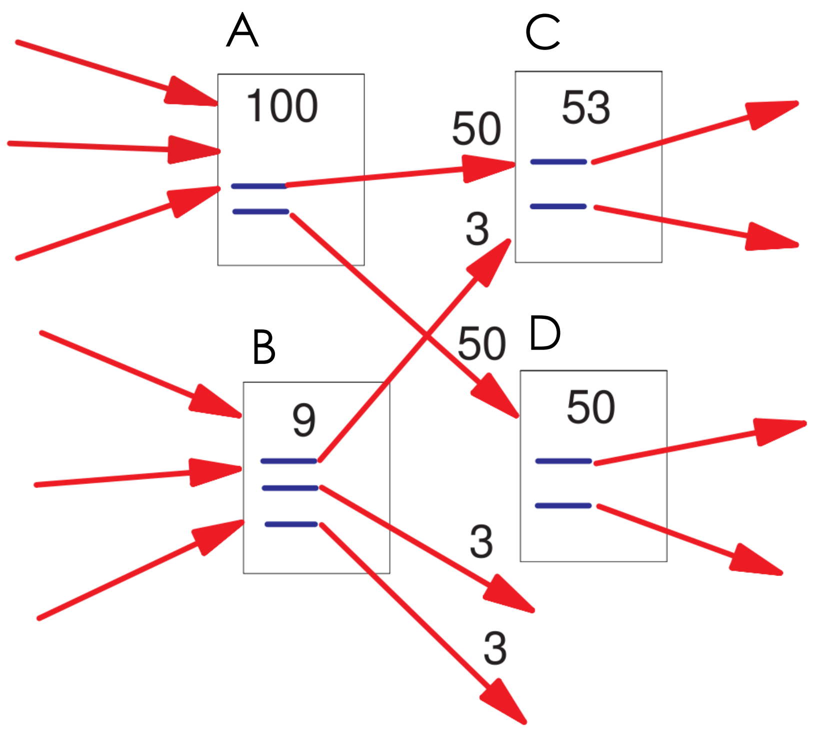
PageRank - Simplified Version

Document A has PageRank of 100 and 2 outlinks:

- It sends 50 to C and 50 to D

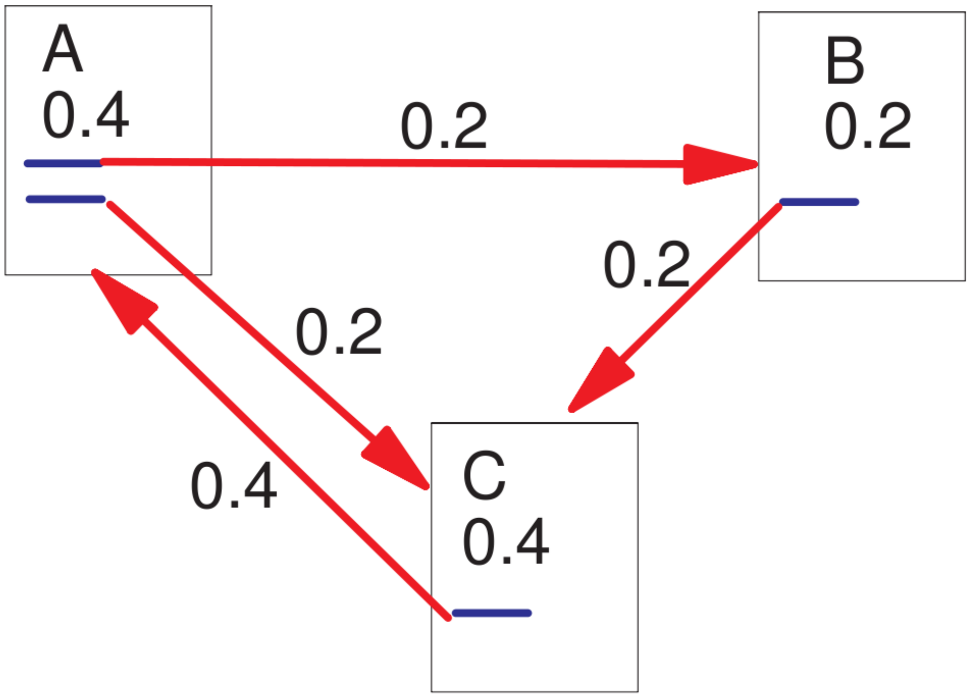
Document B has PageRank of 9 and 3 outlinks:

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



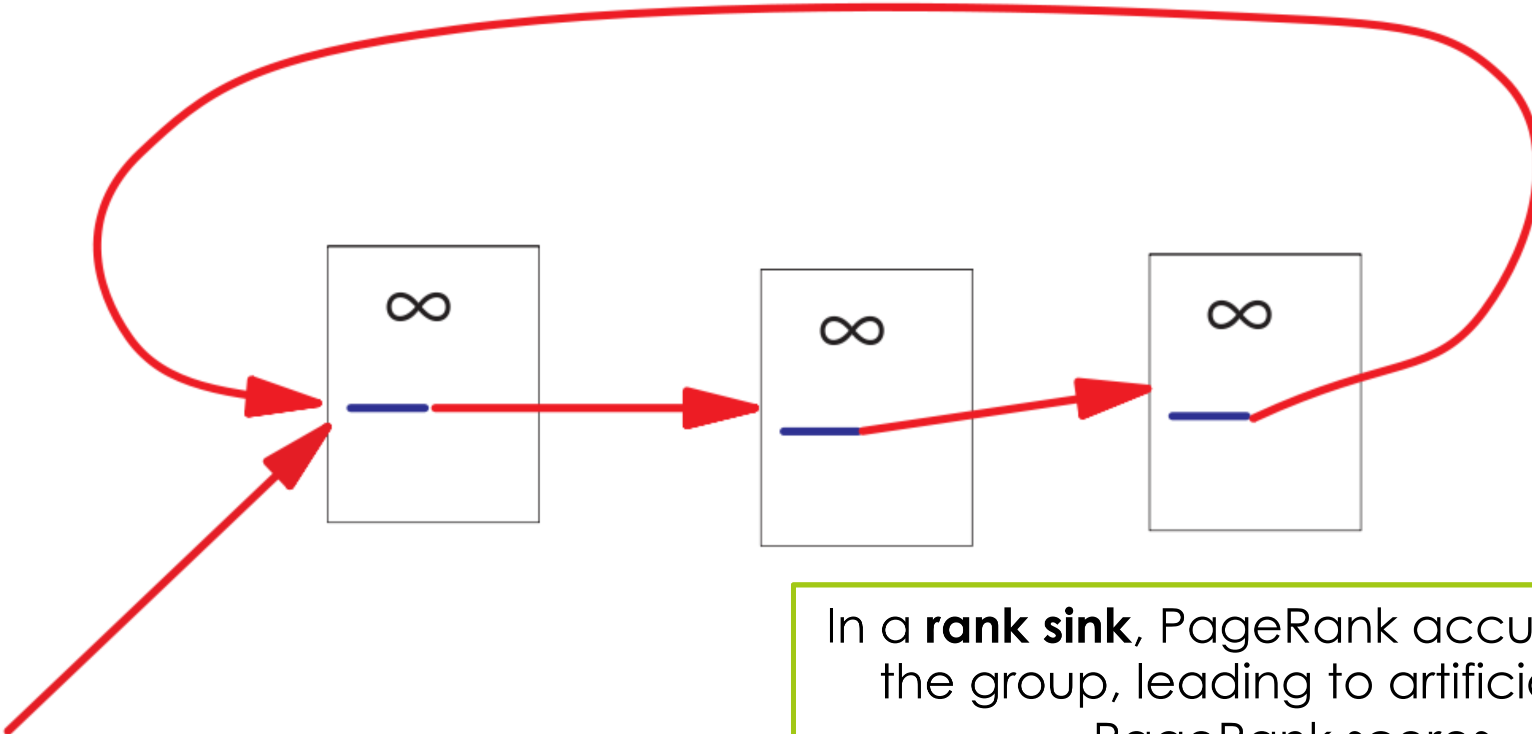
PageRank - Simplified Version

- 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).
    - The scores use as the inputs for each iteration of the algorithm are the scores from the previous iteration.



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

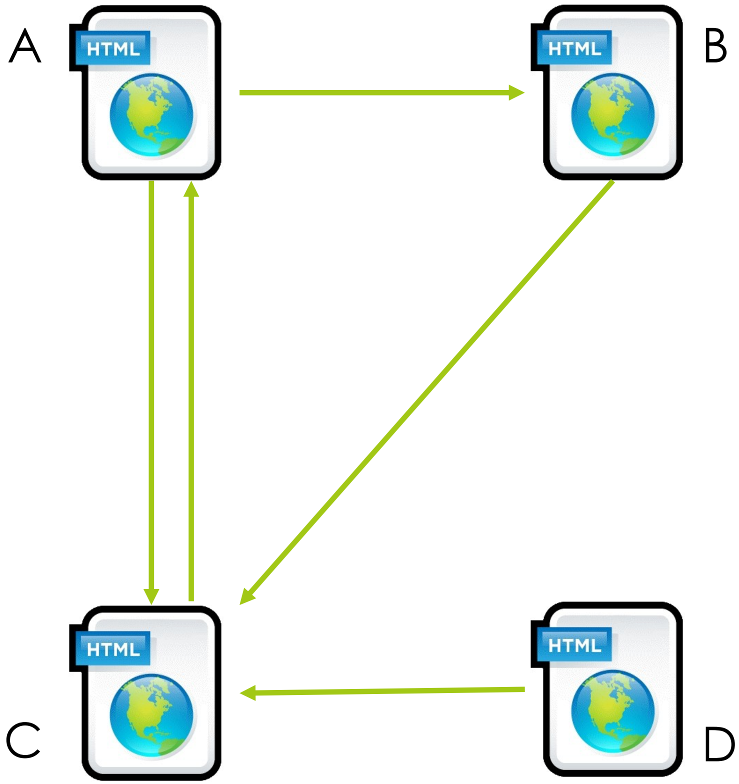
- To combat this type of situation, a new equation is used:
- $R(u) = (1 - d) + d \times \sum_{v \in B_u} \frac{R(v)}{N_v}$
- The formula is the very similar to the one we have seen before.
- The difference is the addition of a **damping factor** (d), which ensures that not all of a document's PageRank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

PageRank - No Backlinks

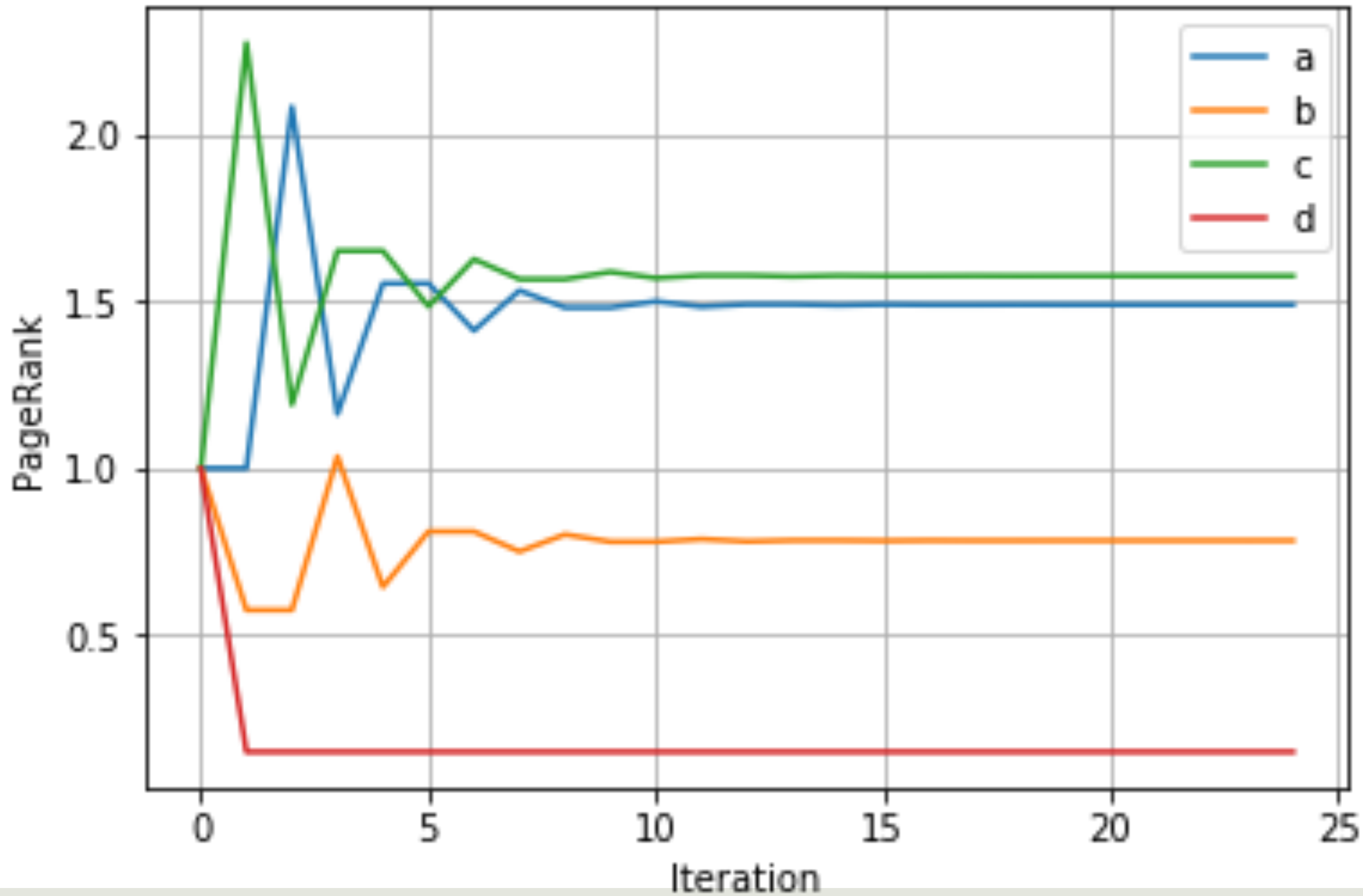
- In the original formula, the only source of PageRank for a document is from its backlinks.
- This meant that a document with no backlinks would have a PageRank of zero.
- This is perfectly acceptable when considering the importance of that document itself.
- However, this also means that it would not contribute anything to the PageRank of documents it links 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
- Starting with an initial PageRank of 1 and using a damping factor of 0.85 (which Google appears to use), calculate the PageRank of each document.

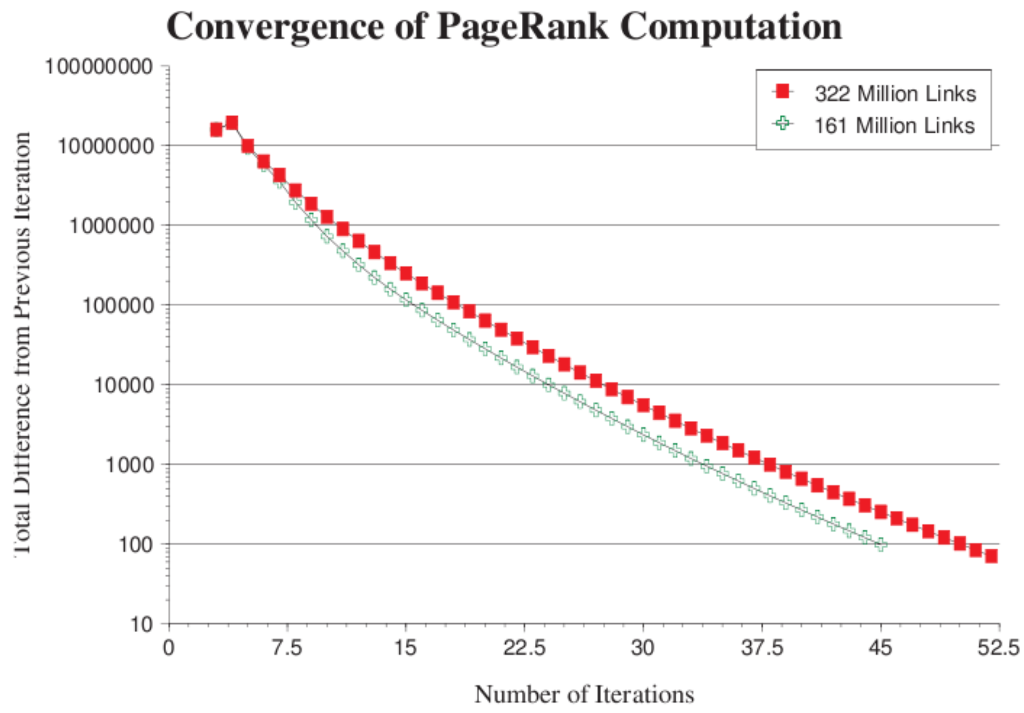


PageRank - Example: 4 Pages



PageRank - Example: 4 Pages

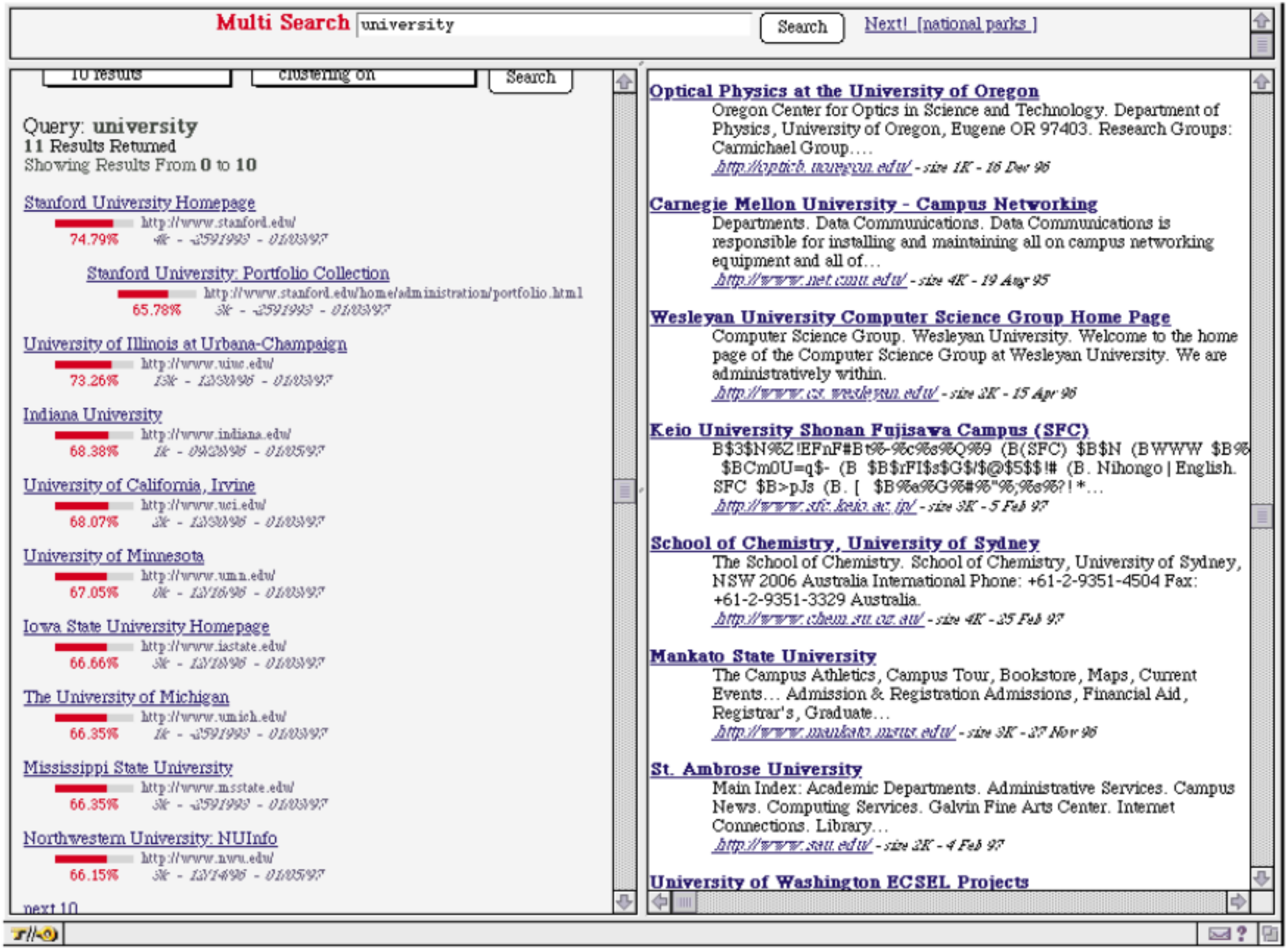
- With this simple system, the PageRank scores have **converged** perfectly after 20 iterations.
- Even after 8 or 9 iterations, the values are very similar to their ultimate values.
- Brin and Page were using a database of 322 million links and believed that the convergence had reached a reasonable 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....
Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
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.
Keio University Shonan Fujisawa Campus (SFC)
Nihongo | English.
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.
Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
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.