

Web Science: Searching the Web

(Part 1 - Crawling the Web)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0
Unported License](https://creativecommons.org/licenses/by-nc-sa/3.0/)

Search for “web science”

A screenshot of a Google search results page for the query "web science". The search bar at the top shows the query. Below it, the "All" tab is selected, along with other options like Images, News, Videos, Shopping, More, Settings, and Tools. The search results include:

- Web of Science**
www.webofknowledge.com
Web of Science - Web of Science Group - Clarivate Analytics
Web of Science is the world's most trusted publisher-independent global citation database. Guided by the legacy of Dr Eugene Garfield, inventor of the world's ...
Web of Science Core Collection · Web of Science Strategic ... · Contact Us
- Web Science Trust**
www.webscience.org
The Web Science Trust (WST) is a charity promoting the understanding of the Web, through education and research in the discipline of Web Science. It hosts the ...
Web Science Publications · Web Science Trust Board · Web Science Blog
- Web of Science - Wikipedia**
en.wikipedia.org · wiki · Web_of_Science
Web of Science is a website which provides subscription-based access to multiple databases that provide comprehensive citation data for many different ...
Temporal coverage: 1900 to present No. of records: 90 million +
Producer: Clarivate Analytics (United States) Disciplines: Science, social science, arts, hum...
- People also ask**
 - What is Web of Science used for?
 - Is Web of Science free?
 - How do I find Web of Science articles?
 - How do I create a Web of Science account?
- Web science - Wikipedia**
en.wikipedia.org · wiki · Web_science
Web science is an emerging interdisciplinary field concerned with the study of large-scale socio-technical systems, particularly the World Wide Web. ... Web Science combines research from disciplines as diverse as sociology, computer science, economics, and mathematics.

Overview of Web Search

- Must collect the documents
 - **crawl** the web
- Then **index** the documents
 - what words are in each document?
- Finally, return a ranked list of documents from a query
 - **ranking** is where the hard part comes in

What we'll examine

- Web crawling
- Building an index
- Querying the index
- Term frequency and inverse document frequency
- Other methods to increase relevance
- How links between web pages can be used to improve ranking search engine results
- Link spam and how to overcome it

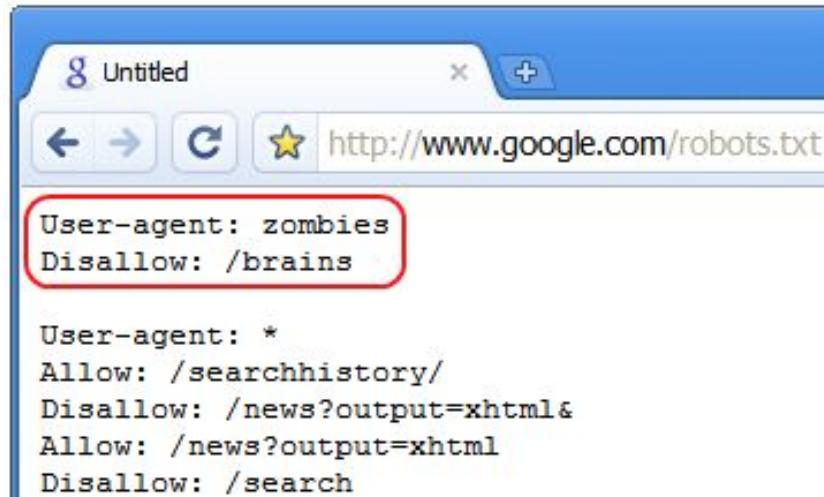
Web Crawling

- Large search engines use thousands of continually running web crawlers to discover web content
- Web crawlers fetch a page, place all the page's links in a queue, fetch the next link from the queue, and repeat
- Web crawlers are usually polite
 - Identify themselves through the HTTP User-Agent request header (e.g., googlebot)
 - Throttle requests to a web server, crawl at off-peak times
 - Honor robots exclusion protocol (robots.txt). Example:

```
User-agent: *
Disallow: /private
```

more about robots.txt: [The Web Robots Pages](#)

Robots.txt Humor



The screenshot shows a web browser window titled "Untitled". The address bar displays the URL "http://www.google.com/robots.txt". The page content is the Google robots.txt file. A red box highlights the following two lines:

```
User-agent: zombies
Disallow: /brains
```

Below this, the rest of the file is shown:

```
User-agent: *
Allow: /searchhistory/
Disallow: /news?output=xhtml&
Allow: /news?output=xhtml
Disallow: /search
```

Halloween 2008

"Halloween easter egg: Google protects itself from zombies", which links to live version
of robots.txt (archived version of robots.txt)

Web Crawler Components

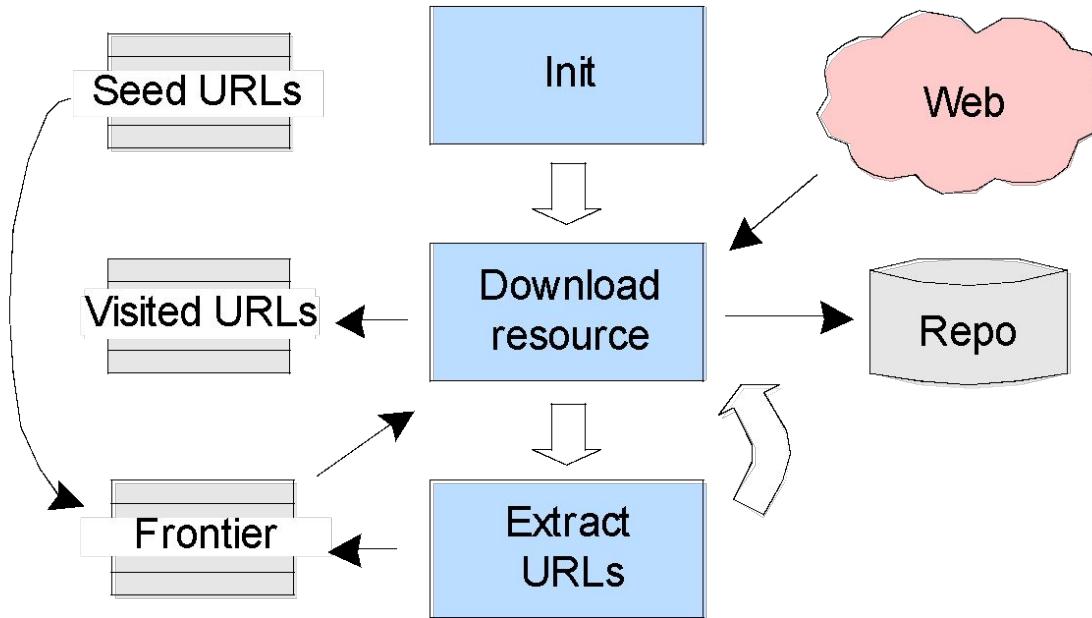
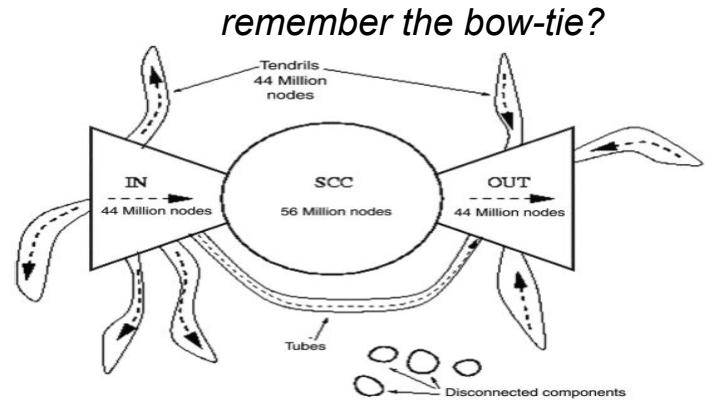


Figure: McCown, *Lazy Preservation: Reconstructing Websites from the Web Infrastructure*, Dissertation, 2007

Crawling Issues

- Good source for seed URLs:
 - [DMOZ - The Directory of the Web](#)
 - archived version of dmoz.org
 - Previously crawled URLs
- Search engine competing goals:
 - Keep index fresh (crawl often)
 - Comprehensive index (crawl as much as possible)
- Which URLs should be visited first or more often?
 - Breadth-first (FIFO)
 - Pages which change frequently & significantly
 - Popular or highly-linked pages



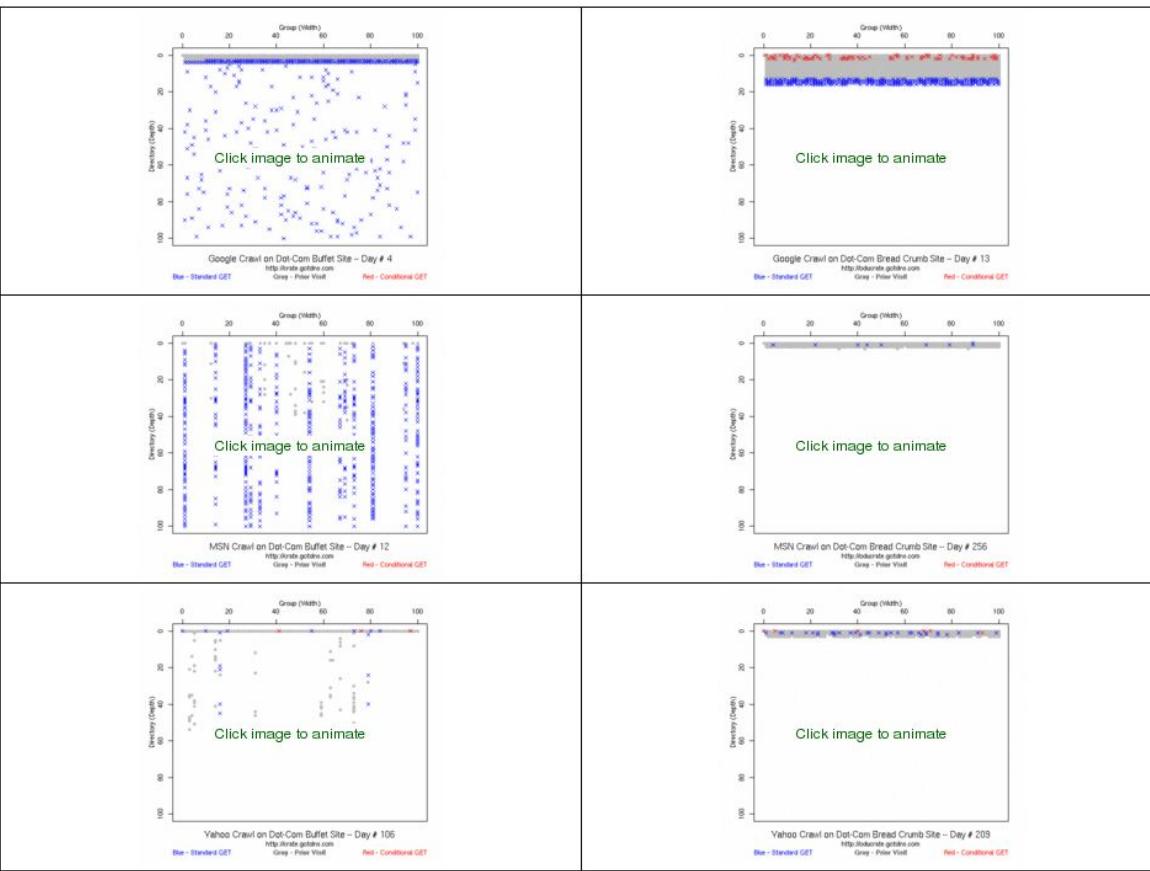


Table 3 Links to Animations of Crawler Activity at the Dot-Com Sites

web crawler animations - see Table 3 of Smith and Nelson, [Site Design Impact on Robots: An Examination of Search Engine Crawler Behavior at Deep and Wide Websites](#), 2008

Crawling Issues - Part 2

- Should avoid crawling duplicate content
 - Convert page content to compact string (*fingerprint*) and compare to previously crawled fingerprints
- Should avoid crawling spam
 - Content analysis of page could make crawler ignore it while crawling or in post-crawl processing
- Robot traps
 - Deliberate or accidental trail of infinite links (e.g., calendar)
 - Solution: limit depth of crawl
- Deep Web
 - Throw search terms at interface to discover pages¹
 - Sitemaps allow websites to publish URLs that might not be discovered in regular web crawling

¹Madhavan et al., Google's Deep Web crawl, Proc. VLDB 2008

Example Sitemap

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
    <url>
        <loc>http://www.example.com/</loc>
        <lastmod>2009-10-22</lastmod>
        <changefreq>weekly</changefreq>
        <priority>0.8</priority>
    </url>
    <url>
        <loc>http://www.example.com/specials.html</loc>
        <changefreq>daily</changefreq>
        <priority>0.9</priority>
    </url>
    <url>
        <loc>http://www.example.com/about.html</loc>
        <lastmod>2009-11-4</lastmod>
        <changefreq>monthly</changefreq>
    </url>
</urlset>
```

Find Sitemaps in robots.txt

```
$ curl -i www.cnn.com/robots.txt
HTTP/1.1 200 OK
Date: Wed, 05 Feb 2020 20:29:35 GMT
Content-Type: text/plain; charset=utf-8
Content-Length: 1061
Accept-Ranges: bytes
Date: Wed, 05 Feb 2020 20:29:35 GMT
Transfer-Encoding: chunked
Connection: keep-alive

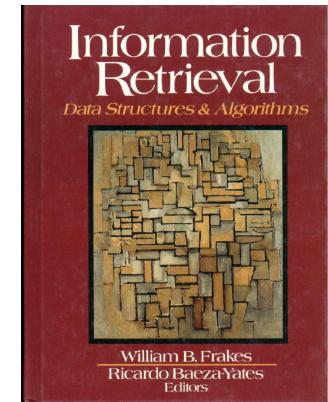
Sitemap: https://www.cnn.com/sitemaps/cnn/index.xml
Sitemap: https://www.cnn.com/sitemaps/cnn/news.xml
Sitemap: https://www.cnn.com/sitemaps/sitemap-section.xml
Sitemap: https://www.cnn.com/sitemaps/sitemap-interactive.xml
Sitemap: https://www.cnn.com/ampstories/sitemap.xml
Sitemap: https://edition.cnn.com/sitemaps/news.xml
User-agent: *
Allow: /audio/podcasts/
Allow: /partners/ipad/live-video.json
Disallow: /*.jsx$
Disallow: /ads/
Disallow: /aol/
[deletia...]
```

Focused Crawling

- A *vertical search engine* focuses on a subset of the Web
 - Google Scholar – scholarly literature
 - Shopzilla – Internet shopping
- A topical, or *focused, web crawler* attempts to download only pages about a specific topic
 - Has to analyze page content to determine if it's on topic and if links should be followed
 - Usually analyzes anchor text as well

Is What I've Found On Topic?

- Precision
 - "ratio of the number of relevant documents retrieved over the total number of documents retrieved" (p. 10)
 - *how much extra stuff did you get?*
- Recall
 - "ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in the database" (p. 10)
 - note: assumes a priori knowledge of the denominator!
 - *how much did you miss?*



[FBY]

Precision and Recall

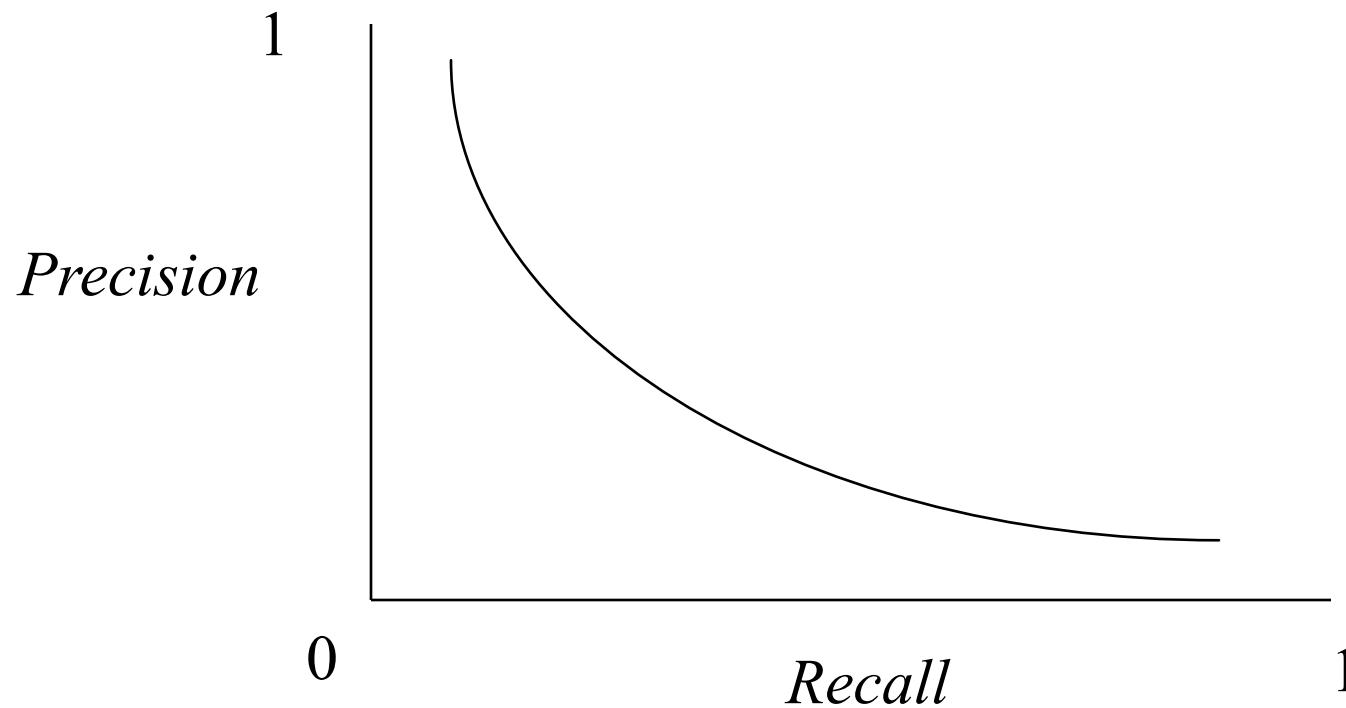


Figure 1.2
in FBY

shadow - Google Search

Web Images Maps News Shopping Gmail more ▾ RhodeWarri

Google™ shadow Search Advanced search Preferences

Web Video Images Groups Results 1 - 10 of about 151,000,000 for shadow [definition]. (0.20 seconds)

Shadow - Wikipedia, the free encyclopedia
A **shadow** is an area where direct light from a light source cannot reach due to obstruction by an object. It occupies all of the space behind an opaque ...
en.wikipedia.org/wiki/Shadow - 40K - [Cached](#) - [Similar pages](#) -

The Shadow - Wikipedia, the free encyclopedia
The **Shadow** as depicted on the cover of the July 15, 1939 issue of The **Shadow** Magazine. The story, "Death From Nowhere," was one of the magazine plots ...
en.wikipedia.org/wiki/The_Shadow - 95k - [Cached](#) - [Similar pages](#) -

T-Mobile Shadow
T-Mobile Sites T-Mobile My T-Mobile Wi-Fi HotSpot myFaves T-Mobile G1 T-Mobile Sidekick HotSpot @Home T-Mobile **Shadow** T-Mobile NBA T-Mobile Invitational ...
www.t-mobileshadow.com/ - 12k - [Cached](#) - [Similar pages](#) -

See results for: **shadow the hedgehog**

SHADOW the HEDGEHOG
www2.sega.com/gamesite/shadow/base.html

Shadow the Hedgehog - Wikipedia, the free encyclopedia
Shadow the Hedgehog (シャドウ・ザ・ヘッジhog, Shadō Za Hejihoggu?) is a character from the Sonic the Hedgehog series, an artificially created life form ...
en.wikipedia.org/wiki/Shadow_the_Hedgehog

Shadow the Hedgehog (video game) - Wikipedia, the free encyclopedia
Nov 21, 2005 ... **Shadow the Hedgehog** is a video game starring **Shadow the Hedgehog** of Sega's Sonic the Hedgehog series. It was revealed at the 2005 inauguration of Sonic the ...
[en.wikipedia.org/wiki/Shadow_the_Hedgehog_\(video_game\)](http://en.wikipedia.org/wiki/Shadow_the_Hedgehog_(video_game))

Video results for shadow

The Rasmus-In the Shadows
3 min 44 sec
www.youtube.com

Wonderful World shadow puppet
2 min 17 sec
www.youtube.com

shadow
Manufacturers of BMX apparel, parts and accessories - USA

Find: Next Previous Highlight all

Done

Why Isn't Precision Always 100%?

What were we really searching for?
Science? Games? Music?

Why Isn't Recall Always 100%?



Virginia Agricultural and Mechanical College?

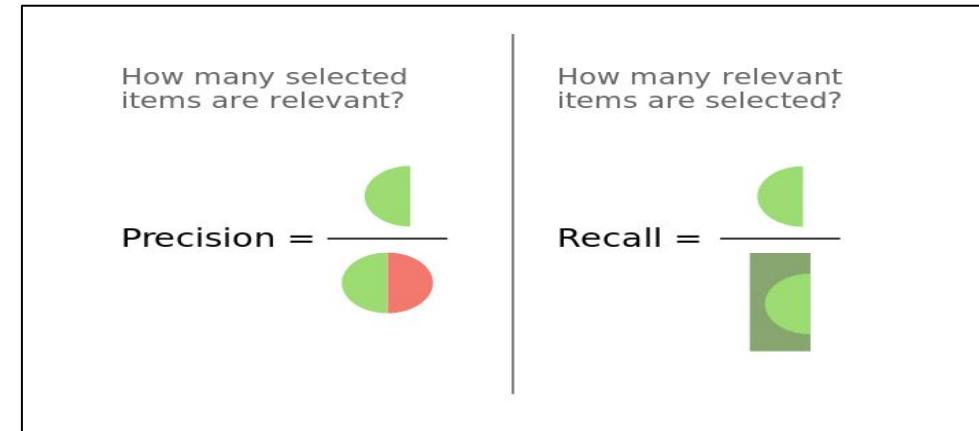
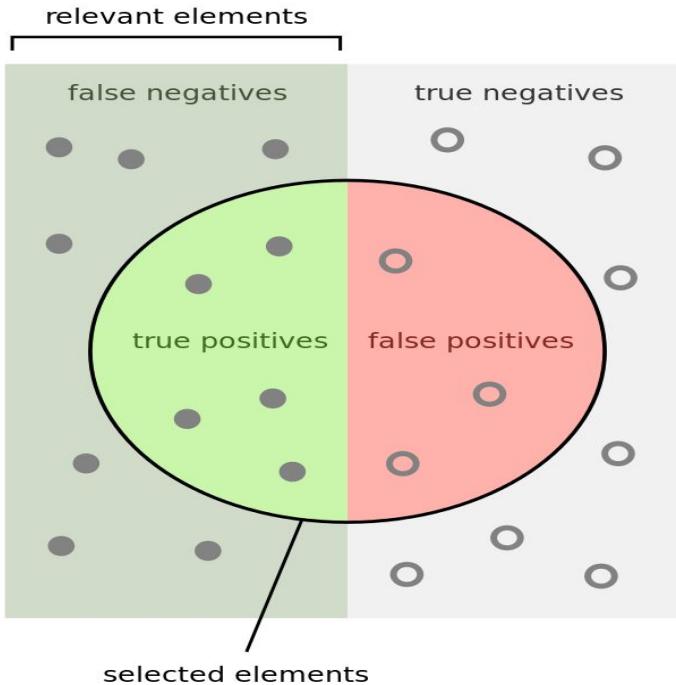
Virginia Agricultural and Mechanical College
and Polytechnic Institute?

Virginia Polytechnic Institute?

Virginia Polytechnic Institute
and State University?

Virginia Tech?

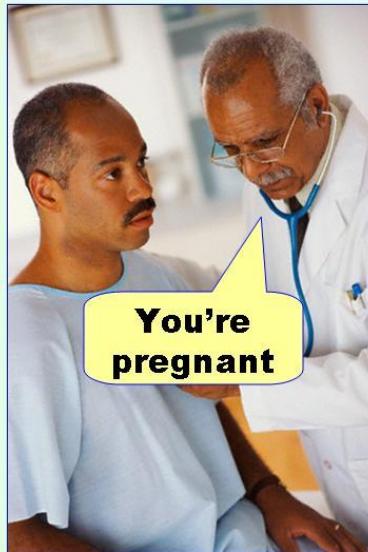
Precision and Recall in Diagrams



From: [Precision and recall](#) (Wikipedia)

Mnemonic for I vs. II

Type I error
(false positive)



Type II error
(false negative)



From: [Posted in Type II error](#)

More than just P&R...

		True condition		Prevalence $= \frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	
Total population		Condition positive	Condition negative		
Predicted condition	Predicted condition positive	True positive	False positive (Type I error)	Positive predictive value (PPV), Precision $= \frac{\sum \text{True positive}}{\sum \text{Test outcome positive}}$	False discovery rate (FDR) $= \frac{\sum \text{False positive}}{\sum \text{Test outcome positive}}$
	Predicted condition negative	False negative (Type II error)	True negative	False omission rate (FOR) $= \frac{\sum \text{False negative}}{\sum \text{Test outcome negative}}$	Negative predictive value (NPV) $= \frac{\sum \text{True negative}}{\sum \text{Test outcome negative}}$
$\text{Accuracy (ACC)} = \frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$		True positive rate (TPR), Sensitivity, Recall $= \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out $= \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$
		False negative rate (FNR), Miss rate $= \frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$	

From: [Precision and recall](#) (Wikipedia)

Adapting Precision & Recall for Web Scale

- Precision @ n (often n=10)
 - previous query for shadow had ~151M hits!
 - no one will check more than just a few
 - "ten blue links"
 - [10 blue links: are they dead or alive in search? – Econsultancy](#)
- Ex:
 - Prec@3 of 2/3
 - Prec@4 of 2/4
 - Prec@5 of 3/5
- Relative recall
 - recall was hard enough on bounded collections; all but impossible for Web
 - relative recall is a way of judging SE overlap
 - $(\text{recall of system A}) / (\text{recall of A} + \text{B} \dots + \text{N})$

Web Science: Searching the Web

(Part 2 - Indexing the Documents)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](#)

Processing Pages

- After crawling, content is *indexed* and links stored in link database for later analysis
- Text from text-based files (HTML, PDF, MS Word, etc.) are converted into *tokens*
- Stop words may be removed
 - Frequently occurring words like *a, the, and, to*, etc.
 - [Stop word](#) (Wikipedia)
 - Most traditional IR systems remove them, but most search engines do not ("to be or not to be")
- Special rules to handle punctuation
 - e-mail → email?
 - Treat "O'Connor" like "boy's"?
 - 123-4567 as one token or two?

Processing Pages

- *Stemming* may be applied to tokens
 - Technique to remove suffixes from words (e.g., *gamer*, *gaming*, *games* → *gam*)
 - Porter Stemmer very popular algorithmic stemmer
 - Can reduce size of index and improve recall, but precision is often reduced
 - Google and Yahoo use partial stemming
- Tokens may be converted to lowercase
 - Most web search engines are case insensitive

Inverted Index

- *Inverted index*, or *inverted file*, is the data structure used to hold tokens and the pages they are located in

documents		
it	1, 2, 3	
is	1, 2, 3	
what	1, 2	
was	1	
a	3	
banana	3	

↑
term list

- Example:
 - Doc 1: It is what it was.
 - Doc 2: What is it?
 - Doc 3: It is a banana.

Example Search

- Search for *what is it* is interpreted by search engines as *what* AND *is* AND *it*

– <i>what</i> : {1, 2}	<i>is</i> : {1, 2, 3}	<i>it</i> : {1, 2, 3}	it	1, 2, 3
			is	1, 2, 3
– {1, 2} \cap {1, 2, 3} \cap {1, 2, 3} = {1, 2}			what	1, 2
– Answer: Docs 1 and 2			was	1
			a	3
			banana	3

- What if we want the phrase "*what is it*"?

Phrase Search

- Phrase search requires *position* of words to be added to inverted index

Doc 1: It is what it was.	it	(1,1) (1,4) (2,3) (3,1)
	is	(1,2) (2,2) (3,2)
Doc 2: What is it?	what	(1,3) (2,1)
Doc 3: It is a banana.	was	(1,5)
	a	(3,3)
	banana	(3,4)

Example Phrase Search

- Search for "*what is it*"
- All items must be in same doc with position in increasing order
- **what:** (1,3) (2,1) **is:** (1,2) (2,2) (3,2) **it:** (1,1) (1,4) (2,3) (3,1)
- **Answer:** Document 2
- Position can be used to give higher scores to terms that are closer
 - "red cars" scores higher than "red bright cars"

What About *Large* Indexes?

- When indexing the entire Web, the inverted index will be too large for a single computer
- Solution: Break up index onto separate machines/clusters
- Two general methods:
 - Document-based partitioning
 - Term-based partitioning

Google's Data Centers

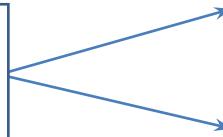


Image source: [10 of the coolest photos from inside Google's secret data centres](#)

Partitioning Schemes

Document-based Partitioning

apple	1, 3, 5, 10
banana	2, 3, 5
...	



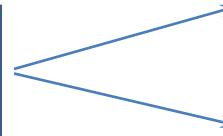
apple 1, 10
banana 2

apple 3, 5
banana 3, 5

apple 1, 3, 5, 10

Term-based Partitioning

apple	1, 3, 5, 10
banana	2, 3, 5
...	



banana 2, 3, 5

Ref: [Distributing indexes](#), *Introduction to Information Retrieval*

Comparing the Two Schemes

Event	Document-Based	Term-Based
Fetch query results	All computers fetch local results and merge	Single machine fetches result for each term
Machine goes down	Some docs not in results	Some terms cannot be processed
Index new doc	Add new terms/docs to single machine	Rebuild index

Guess which scheme Google uses?

Web Science: Searching the Web (Part 3 - Ranking: TF-IDF)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](#)

If two documents contain the same query terms, how do we determine which one is *more* relevant?



Term Frequency

Query for "dogs" results in two pages:

TF = 3

Dogs, dogs, I
love them
dogs!

Doc A

Dogs are
wonderful
animals.

TF = 1

Doc B

- Which page should be ranked higher?
- Simple method called **term frequency**: Count number of times the term occurs in the document
- Pages with higher TF get ranked higher

Normalizing Term Frequency

Query for "dogs" results in two pages:

$$TF = 3/6 = 0.5$$

Dogs, dogs, I
love them
dogs!

Doc A

Dogs are
wonderful
animals.

$$TF = 1/4 = 0.25$$

Doc B

- To avoid penalizing shorter documents, TF should be normalized
 - Divide by total number of words in the document
 - Other divisors possible

TF Can Be Spammed!

Watch out!

Dogs, dogs, I
love them
dogs!
dogs
dogs
dogs
dogs
dogs
dogs

TF is susceptible to
spamming, so SEs look for
unusually high TF values
when looking for spam

$$\text{TF} = 8/11 = 0.72$$

Inverse Document Frequency

- Problem: Some terms are frequently used throughout the corpus and therefore aren't useful when discriminating docs from each other
- Less frequently used terms are more helpful
- $\text{IDF}(\text{term}) = \text{total docs in corpus} / \text{docs with term}$
- Low frequency terms will have high IDF

Inverse Document Frequency

- To keep IDF from growing too large as corpus grows:

$$\text{IDF(term)} = \log_2(\text{total docs in corpus} / \text{docs with term})$$

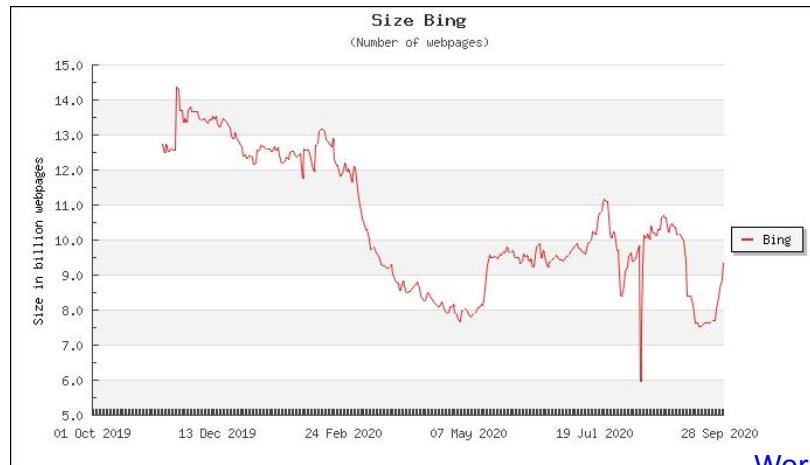
- IDF is not as easy to spam since it involves all docs in corpus
 - Could stuff rare words in your pages to raise IDF for those terms, but people don't often search for rare terms

TF-IDF

- TF and IDF are usually combined into a single score
- $\text{TF-IDF} = \text{TF} \times \text{IDF}$
= occurrence in doc / words in doc \times
 $\log_2(\text{total docs in corpus} / \text{docs with term})$
- When computing TF-IDF score of a doc for n terms:
 - Score = $\text{TF-IDF}(\text{term}_1) + \text{TF-IDF}(\text{term}_2) + \dots + \text{TF-IDF}(\text{term}_n)$

TF-IDF Example

- Using Bing, compute the TF-IDF scores for 2 documents that contain the words *harding* AND *university*
- Assume Bing has **10 billion documents** indexed



WorldWideWebSize.com (accessed Sep 2020)

TF-IDF Example

- Actions to perform:
 1. Query Bing with *harding university* to pick 2 docs
 2. Query Bing with just *harding* to determine how many docs contain *harding*
 3. Query Bing with just *university* to determine how many docs contain *university*

1) Search for *harding university* and choose two results

The screenshot shows a Bing search results page for the query "harding university". The search bar at the top contains the query. Below the search bar, there are tabs for Web, Facts, Local, Images, and More. The main content area displays search results for Harding University. The first result is a sponsored link for "Harding University - Faith, Learning, Living" with a thumbnail image of a building and the word "HARDING" below it. This result includes links for Athletics, Students & Employees, Admissions & Aid, and School Information. The School Information section provides details about the location (Searcy, Arkansas), setting (Distant Town), type (Private not-for-profit), and level (Four or more years). Below this result, there is a section for "Harding University Athletics" which mentions the university's success in NCAA II South Region championships.

harding university - Bing

www.bing.com/search?q=harding+university&go=&form=QBLH&qs=n&sk=&sc=8-18

Web Images Videos Shopping News Maps More | MSN Hotmail Sign in ▾ Searcy, Arkansas

bing Web

harding university

Web Facts Local Images More

RELATED SEARCHES

- Harding University
- Searcy AR
- Harding University Choir
- Harding University Graduate School of Religion
- Harding Academy
- Searcy
- Hardin University
- Harding College of Pharmacy
- Heritage Christian University
- Missouri Southern State University

ALL RESULTS

1-10 of 9,390,000 results - Advanced

Harding University - Faith, Learning, Living

www.harding.edu · Official site

Harding Campuses: Nursing graduates honored at pinning...

Athletics Harding University

Students & Employees News & Events

Admissions & Aid Feeds

Quick Access School Information

Customer service 800-477-4407 Location: Searcy, Arkansas

Setting: Distant Town

Type: Private not-for-profit

Level: Four or more years

Harding University Athletics

Harding Men's Cross Country has won 10 of the last 11 NCAA II South Region championships. Daniel Kinwa has won the last three NCAA II South Region individual

SEARCH HISTORY

2) Search for *harding*

A screenshot of a Microsoft Bing search results page. The search term 'harding' is entered in the search bar. The results are categorized under 'Web'. The first result is 'Harding University - Faith, Learning, Living', which is a Cached page. Below this result are links for Athletics, Students & Employees, Admissions & Aid, Harding University, News & Events, Feeds, Spiritual Life, Majors & Minors, and a link to show more results from www.harding.edu. A red arrow points to the text '1-10 of 12,200,000 results' with the caption 'Gross exaggeration' written in red. The left sidebar lists 'RELATED SEARCHES' including Tonya Harding, Warren Harding, Harding University, Searcy Arkansas, Harding's Marketplace, Harding Real Estate, Harding Connectors, Harding Township, and Harding's Coaches. The bottom section shows 'SEARCH HISTORY' with entries for 'harding', 'harding university', 'frank mccown', and 'See all'.

Gross exaggeration

2) Search for *university*

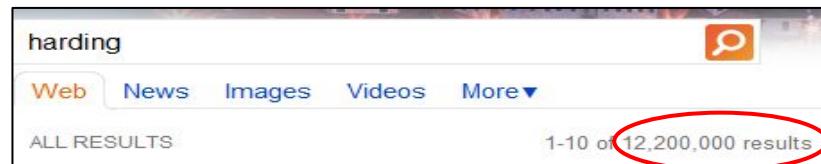
A screenshot of a Bing search results page. The search bar at the top contains the query "university". Below the search bar, there are tabs for Web, Local, News, Videos, Images, and More. The "Web" tab is selected. To the left, under "RELATED SEARCHES", are links to "Where Is Auburn University?", "University Of Phoenix Stadium", "Yale University", "Howard University", "Oxford University", "Karachi University", "Private University", and "Princeton University". On the right, there are sponsored site links for "University Degree" and "Universities". The main search results list includes "University of Phoenix®" (Sponsored sites), "Top Online Universities" (Phoenix.edu), "Top Online University" (ClassesandCareers.com), and "Listings for University near Searcy, Arkansas" (Change location). The search results indicate 1-10 of 439,000,000 results. A red circle highlights the search count "1-10 of 439,000,000 results".

Doc 1: Harding - College of Pharmacy

- webpage contains **967** words
 - "harding" appears **19** times
 - "university" appears **13** times

Assume Bing has **10 billion** documents indexed

- $\text{TF}(\text{harding}) = 19 / 967$
- $\text{IDF}(\text{harding}) = \log_2(10\text{B} / 12.2\text{M})$
- $\text{TF}(\text{university}) = 13 / 967$
- $\text{IDF}(\text{university}) = \log_2(10\text{B} / 439\text{M})$



Doc 1: Harding - College of Pharmacy

- $\text{TF}(\text{harding}) = 19 / 967 = \textcolor{red}{0.020}$
- $\text{IDF}(\text{harding}) = \log_2(10B / 12.2M) = \textcolor{green}{9.679}$
- $\text{TF}(\text{university}) = 13 / 967 = \textcolor{brown}{0.013}$
- $\text{IDF}(\text{university}) = \log_2(10B / 439M) = \textcolor{purple}{4.510}$
- $\text{TF-IDF}(\text{harding}) + \text{TF-IDF}(\text{university}) =$
 $\textcolor{red}{0.020} \times \textcolor{green}{9.679} + \textcolor{brown}{0.013} \times \textcolor{purple}{4.510} = \textcolor{black}{0.252}$

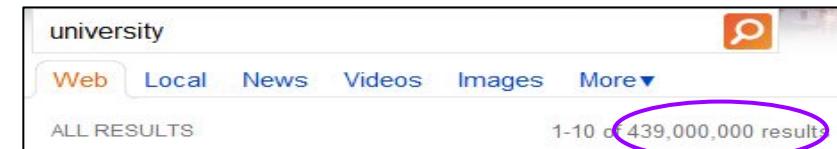
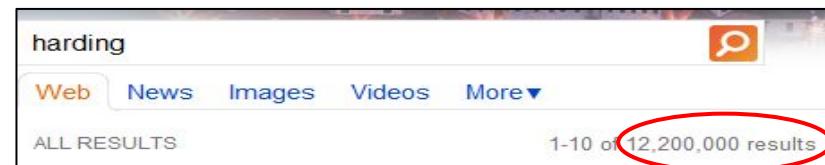
$\log_2 x$ is
 $\log(x, 2)$ in R

Doc 2: Harding University (Wikipedia)

- webpage contains 3135 words
 - "harding" appears 44 times
 - "university" appears 25 times

Assume Bing has **10 billion** documents indexed

- $\text{TF}(\text{harding}) = 44 / 3135$
- $\text{IDF}(\text{harding}) = \log_2(10\text{B} / 12.2\text{M})$
- $\text{TF}(\text{university}) = 25 / 3135$
- $\text{IDF}(\text{university}) = \log_2(10\text{B} / 439\text{M})$



Doc 2: Harding University (Wikipedia)

- $\text{TF}(\text{harding}) = 44 / 3135 = \textcolor{red}{0.014}$
- $\text{IDF}(\text{harding}) = \log_2(10B / 12.2M) = \textcolor{green}{9.679}$
- $\text{TF}(\text{university}) = 25 / 3135 = \textcolor{brown}{0.008}$
- $\text{IDF}(\text{university}) = \log_2(10B / 439M) = \textcolor{purple}{4.510}$
- $\text{TF-IDF}(\text{harding}) + \text{TF-IDF}(\text{university}) = \textcolor{red}{0.014} \times \textcolor{green}{9.679} + \textcolor{brown}{0.008} \times \textcolor{purple}{4.510} = \textcolor{black}{0.172}$

Doc 1 = **0.252**

Doc 2 = 0.172

Web Science: Searching the Web

(Part 4 - Ranking: Link-based Metrics, PageRank, and HITS)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](#)

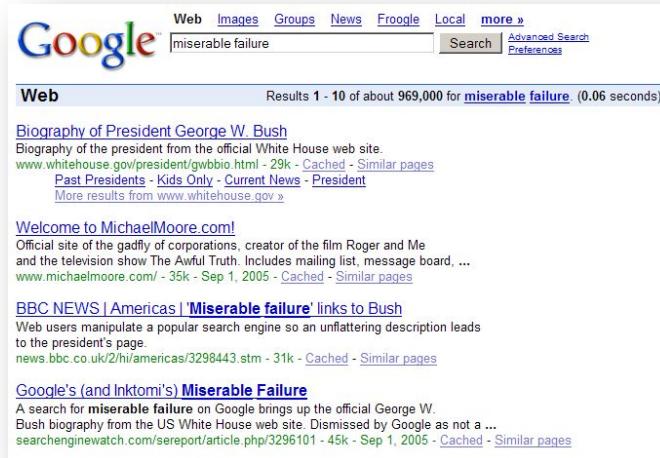
Increasing Relevance

- Index words in URL
- Weigh *importance* of terms based on HTML or CSS styles
- Website responsiveness¹
- Account for last modification date
- Allow for misspellings
- Link-based metrics
- Popularity-based metrics

¹[Using site speed in web search ranking](#)

Increasing Relevance

- Index a link's *anchor text* along with the page it points to
`Ninja skills`
 - Watch out: Google bombs ([Google bombing](#) (Wikipedia))



More info: [Google Kills Bush's Miserable Failure Search & Other Google Bombs](#)

Increasing Relevance

- Index words in URL
- Weigh *importance* of terms based on HTML or CSS styles
- Website responsiveness¹
- Account for last modification date
- Allow for misspellings
- **Link-based metrics**
- Popularity-based metrics

¹[Using site speed in web search ranking](#)

Link Analysis

- Content analysis is useful, but combining with link analysis allow us to rank pages much more successfully
- 2 popular methods
 - Sergey Brin and Larry Page's *PageRank*
 - Jon Kleinberg's Hyperlink-Induced Topic Search (*HITS*)

What Does a Link Mean?



- A recommends B
- A specifically does **not** recommend B
- B is an authoritative reference for something in A
- A & B are about the same thing (topic locality)

PageRank



- Developed by Brin and Page (Google) while Ph.D. students at Stanford
- Links are a recommendation system
 - The more links that point to you, the more important you are
 - Inlinks from important pages are weightier than inlinks from unimportant pages
 - The more outlinks you have, the less weight your links carry

Page et al., [The PageRank citation ranking: Bringing order to the web](#), 1998.

Image credit: <http://scrapetv.com/News/News%20Pages/Technology/images/sergey-brin-larry-page.jpg>

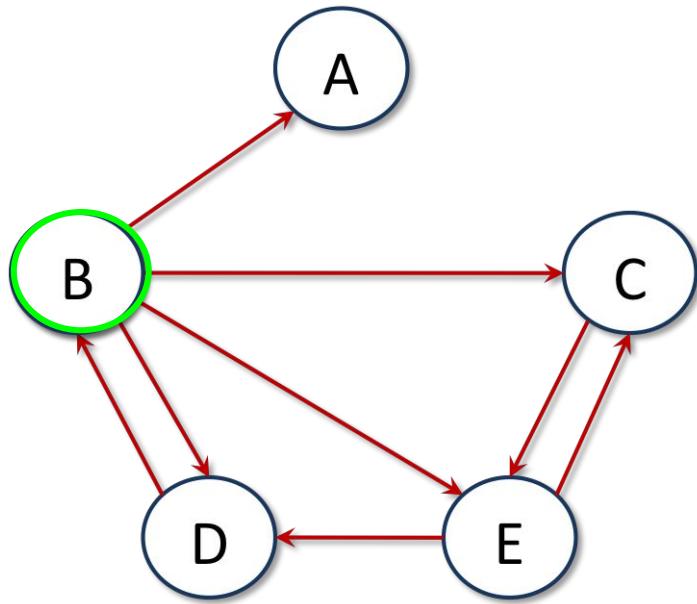
Random Surfer Model

- Model helpful for understanding PageRank
- The Random Surfer starts at a randomly chosen page and selects a link at random to follow
- PageRank of a page reflects the probability that the surfer lands on that page after clicking any number of links



Image credit: <http://missloki84.deviantart.com/art/Random-Surfer-at-Huntington-Beach-319287873>

Example of Random Surfer



Start at: B

$\frac{1}{4}$ probability of going to A

$\frac{1}{4}$ probability of going to C

$\frac{1}{4}$ probability of going to D

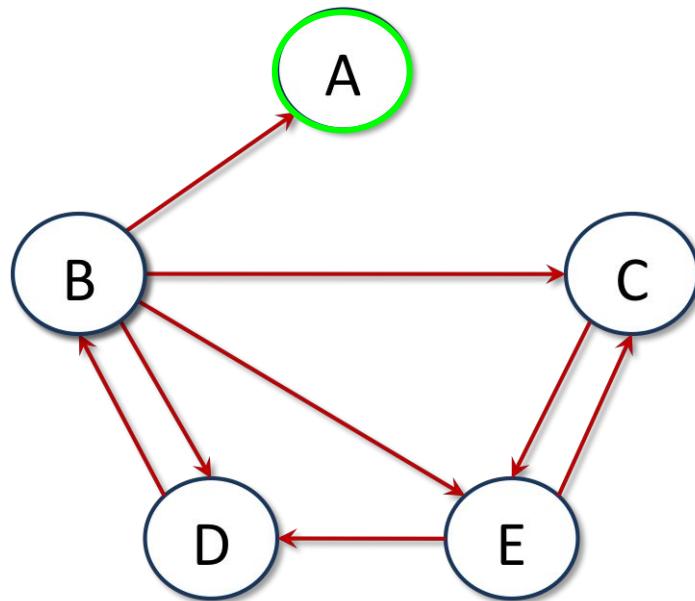
$\frac{1}{4}$ probability of going to E

Choose: E

$\frac{1}{2}$ probability of going to C

$\frac{1}{2}$ probability of going to D

Problem 1: Dangling Node



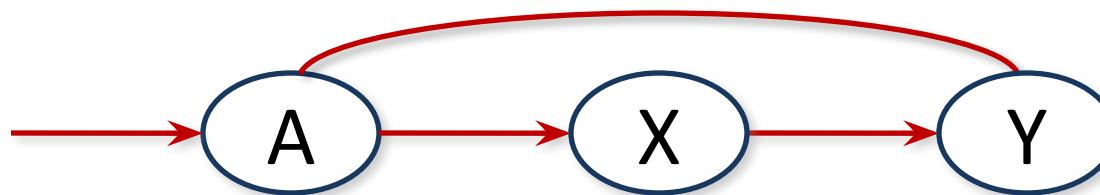
What if we go to A? We're stuck at a dead-end!

Solution: **Teleport** to any other page at random

Problem 2: Infinite Loop

What if we get stuck in an a **cycle**?

Solution: Teleport to any other page at random



Rank Sinks

- Dangling nodes and cycles are called **rank sinks**
- Solution is to add a **teleportation probability α** to every decision
- $\alpha\%$ chance of getting bored and jumping somewhere else, $(1-\alpha)\%$ chance of choosing one of the available links
- $\alpha = 0.15$ is typical

PageRank Definition

Teleportation
probability

$$PR(P_i) = \frac{\alpha}{|P|} + (1 - \alpha) \cdot \sum_{P_j \in B_{P_i}} \frac{PR(P_j)}{|P_j|}$$

PageRank of
page P_i

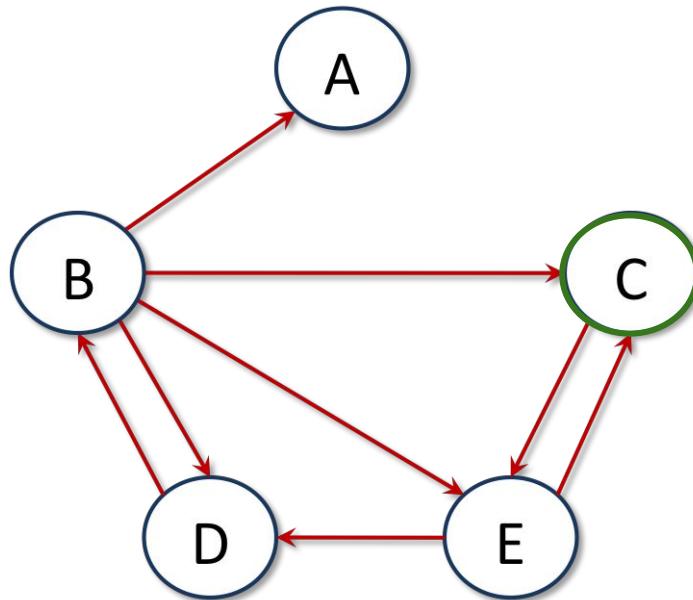
Total num
of pages

B_{P_i} is set of all pages
pointing to P_i

Number of
outlinks from P_j

PageRank Example

$$PR(P_i) = \frac{\alpha}{|P|} + (1 - \alpha) \cdot \sum_{P_j \in B_{P_i}} \frac{PR(P_j)}{|P_j|}$$



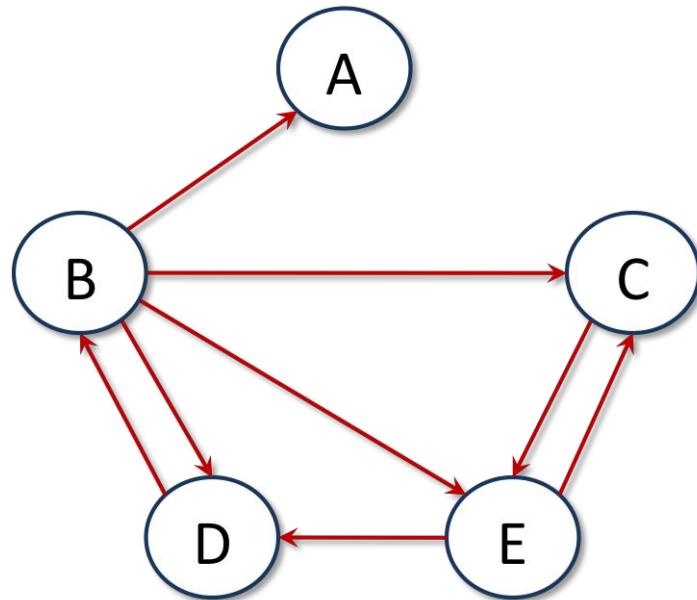
$$PR(C) = .15/5 + .85 \times (PR(B)/4 + PR(E)/2)$$

Problem: What is $PR(B)$ and $PR(E)$?

Solution: Give all pages same PR to start ($1/|P|$) & iteratively calculate new PR

PageRank Example

$$PR(P_i) = \frac{\alpha}{|P|} + (1 - \alpha) \cdot \sum_{P_j \in B_{P_i}} \frac{PR(P_j)}{|P_j|}$$



$$\begin{aligned} PR(A) &= .03 + .85 \times PR(B)/4 \\ &= .03 + .85 \times .2/4 = .0725 \end{aligned}$$

$$\begin{aligned} PR(B) &= .03 + .85 \times PR(D)/1 \\ &= .03 + .85 \times .2/1 = .2 \end{aligned}$$

$$\begin{aligned} \alpha/|P| &= .15 / 5 = .03 \\ \text{initial PR} &= 1/|P| = 1/5 = .2 \end{aligned}$$

$$\begin{aligned} PR(C) &= .03 + .85(PR(B)/4 + PR(E)/2) \\ &= .03 + .85(.2/4 + .2/2) = .1575 \end{aligned}$$

$$PR(D) = .03 + .85(PR(B)/4 + PR(E)/2) = .1575$$

$$PR(E) = .03 + .85(PR(B)/4 + PR(C)/1) = .2425$$

Calculating PageRank

- PageRank is computed over and over until it converges, around 20 times
- Can also be calculated efficiently using matrix multiplication
 - we'll see this a little later

Computing PageRank

original paper
version, sums to N
(number of
pages in graph)

$$PR(A) = 1 - d + d \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \dots \right)$$

more common
version, sums to 1

$$PR(A) = \frac{1 - d}{N} + d \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \dots \right).$$

d = damping factor

$L()$ = out-degree of a page

$PR()$ = PageRank of a page (all nodes start with $PR() = 1$ or $1/N$)

Calculating PageRank for a Page, One Iteration

Ref: [Programming Collective Intelligence](#), Ch 4

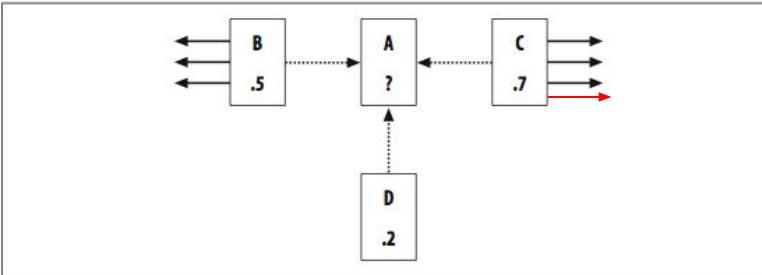
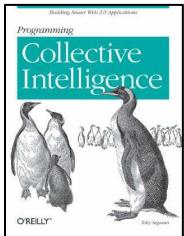


Fig 4-3 needs an extra link from C to match the text

Figure 4-3. Calculating the PageRank of A

$$\begin{aligned} \text{PR}(A) &= (1-0.85) + 0.85 * (\text{PR}(B)/\text{links}(B) + \text{PR}(C)/\text{links}(C) + \text{PR}(D)/\text{links}(D)) \\ &= 0.15 + 0.85 * (0.5/4 + 0.7/5 + 0.2/1) \\ &= 0.15 + 0.85 * (0.125 + 0.14 + 0.2) \\ &= 0.15 + 0.85 * 0.465 \\ &= 0.54525 \end{aligned}$$

damping factor (d) = 0.85 (probability surfer landed on page by following a link)

$1-d = 0.15$ (probability surfer landed on page at "random")

since this is the original version where PR sums to N , and we've only accounted for ~1.95 of total PR, pages not shown must be holding PR

PageRank Definition as Matrix

Stated as a matrix equation where \mathbf{R} is the vector of PageRank values and \mathbf{T} the matrix for transition probabilities:

$$\mathbf{R} = \mathbf{T}\mathbf{R}$$

where T_{ij} is the probability of going from page j to i :

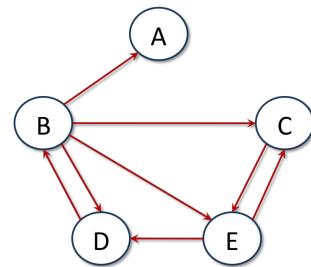
$$T_{ij} = \frac{\alpha}{|P|} + (1 - \alpha) \frac{1}{|P_j|}$$

Total num of pages

Total outlinks from page j

if a link from
page j to i
exists, otherwise

$$T_{ij} = \frac{\alpha}{|P|}$$



PageRank Matrix Example

$$\begin{bmatrix} PR_A \\ PR_B \\ PR_C \\ PR_D \\ PR_E \end{bmatrix} = \begin{bmatrix} \cdot & \cdot & .03 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & .243 & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} .2 \\ .2 \\ .2 \\ .2 \\ .2 \end{bmatrix}$$

No link from C to A, so value = .15/5

$$\begin{aligned}
 T_{EB} &= \frac{\alpha}{|P|} + (1 - \alpha) \frac{1}{|P_B|} \\
 &= .15/5 + (1 - .15)/4 = .243
 \end{aligned}$$

Init PR values
 $1/|P|$

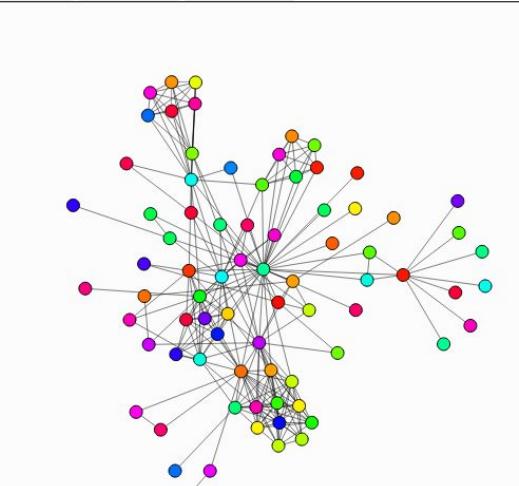
Visualizing the PageRank Model X

Elijah Meeks's Block f448eeff77b5fe94b1c0
Updated June 7, 2017 Popular / About

Visualizing the PageRank Model

X

run once run 100 times run 1000 times



[Visualizing the PageRank Model](#)

This is a simple attempt to visualize the model that the PageRank algorithm is based on. PageRank is a method for discovering central nodes in a network by treating nodes as web pages and edges as links between them, upon which a simulated web surfer starts on a random page and clicks a random link, navigating to a new page, with a 15% chance that the surfer ends that session. In this example, you can step through the process by clicking "run once". The first random node is colored green and "start" appears at the top. If the random walk goes to a new node, the link is colored red and the new node is stroked in black and "step" appears as the top. If the random walk is ended, "end" will appear at the top. After each tick, new PageRank values are calculated for each node by totaling the number of visits to that node and dividing it by summing the number of total

Check Your PageRank...

- [Check Google page rank instantly](#)
- 9/10: google.com, cnn.com
- 7/10: www.cs.odu.edu
- 5/10: ws-dl.blogspot.com
- 2/10: lsufootball.net

updated Feb 2020

PageRank Issues

- Rich-get-richer phenomenon
 - May be difficult for new pages with few inlinks to compete with older, highly linked pages with high PageRank
 - Could promote small fraction of new pages at random¹ or add decay factor to links
- Study² showed just counting number of inlinks gives similar ranking as PageRank
 - Study was on small scale and pages were not necessarily "typical"
 - Counting inlinks more susceptible to spamming

¹Pandey et al., [Shuffling a stacked deck](#), VLDB 2005,

²Amento et al., [Does "authority" mean quality?](#), SIGIR 2000

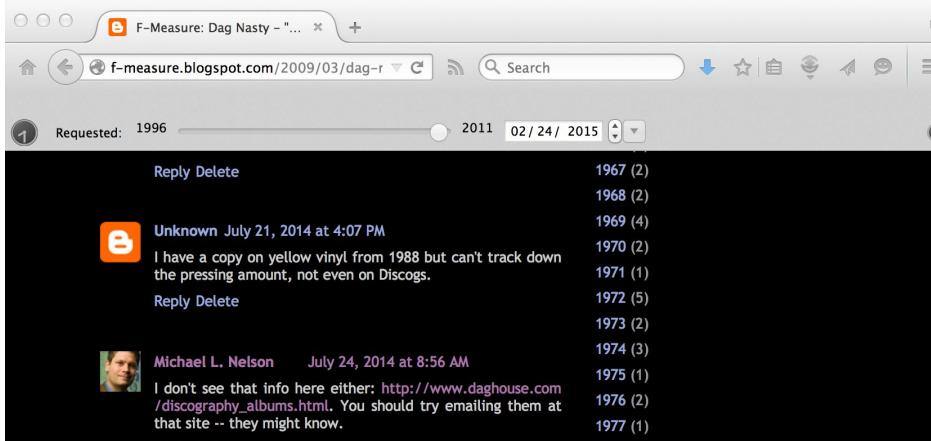
Link Spam



- There is a strong economic incentive to rank highly in a SERP
- “White hat” SEO firms follow published guidelines to improve customer rankings¹
- To boost PageRank, “black hat” SEO practices include:
 - Building elaborate link farms
 - Exchanging reciprocal links
 - Posting links on blogs and forums

¹[Google's Webmaster Guidelines](#)

Trying to Stealing PageRank...



...I don't see that info here either:

<u href="<http://www.daghous...>"
rel="nofollow"><http://www.daghous...></u>. You should try emailing them at that site...

[nofollow](#) (Wikipedia) / [Comment spam](#) (Wikipedia)

Combating Link Spam

- Sites like Wikipedia can discourage links that only promote PageRank by using "nofollow"

```
<a href="http://somesite.com/" rel="nofollow">Go here!</a>
```

- Davison¹ identified 75 features for comparing source and destination pages
 - Overlap, identical page titles, same links, etc.
- TrustRank²
 - Bias teleportation in PageRank to set of trusted web pages

¹Davison, Recognizing nepotistic links on the web, 2000

²Gyöngyi et al., Combating web spam with TrustRank, VLDB 2004

Combating Link Spam

- SpamRank¹
 - PageRank for whole Web has power-law distribution
 - Penalize pages whose supporting pages do not approximate power-law distribution
- Anti-TrustRank²
 - Give high weight to known spam pages and propagate values using PageRank
 - New pages can be classified spam if large contribution of PageRank from known spam pages, or if high Anti-TrustRank

¹Benczúr et al., SpamRank – fully automated link spam detection, AIRWeb 2005

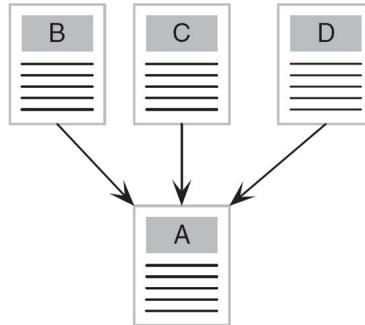
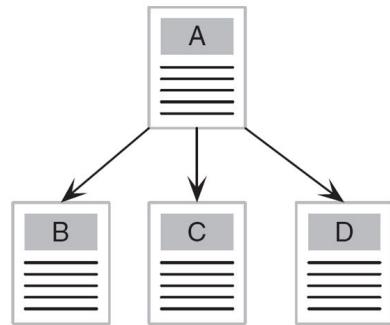
²Krishnan & Raj, Web spam detection with Anti-TrustRank, AIRWeb 2006

HITS

- Hyperlink-induced topic search (HITS) by Jon Kleinberg¹
- **Hub:** page with outlinks to informative web pages
- **Authority:** informative/authoritative page with many inlinks



Img credit: <http://scgp.stonybrook.edu/archives/6084>

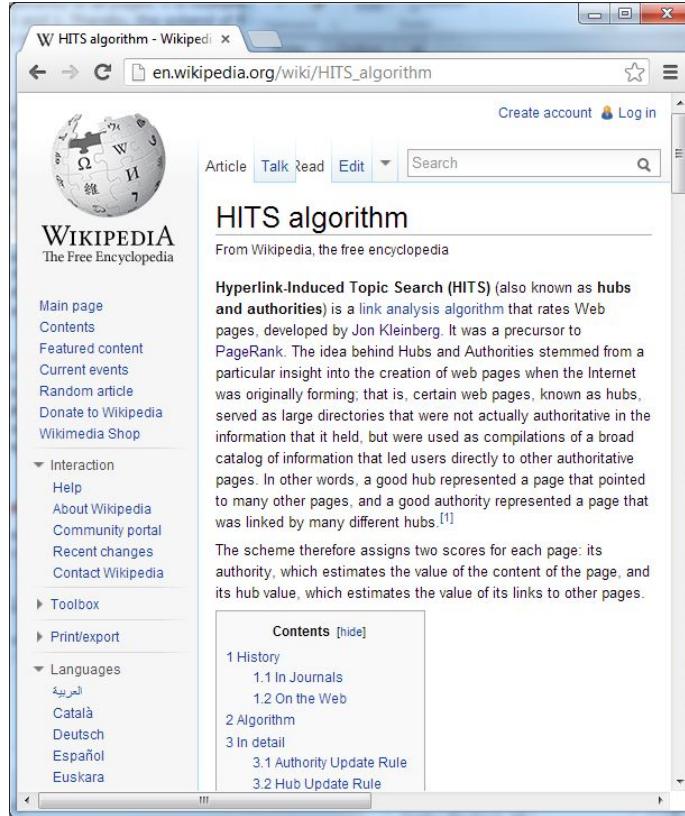


Figures from Levene (2010), *An Introduction to Search Engines and Web Page Navigation*

- Recursive definition:
 - Good *hubs* point to good *authorities*
 - Good *authorities* are pointed to by good *hubs*

¹Kleinberg, [Authoritative sources in a hyperlinked environment](#), J. ACM, 1999

Good Authority & Hub?



This screenshot shows the Wikipedia article on the HITS algorithm. The page title is "HITS algorithm". The content starts with a brief introduction: "Hyperlink-Induced Topic Search (HITS) (also known as **hubs** and **authorities**) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It was a precursor to PageRank. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs.^[1]" Below this, there is a paragraph about the scoring scheme: "The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages." At the bottom of the page, there is a "Contents" sidebar with sections like "1 History", "2 Algorithm", and "3 In detail".



This screenshot shows the same Wikipedia article on the HITS algorithm, but with some additional content. At the top, there is a code snippet in pseudocode:

```
10     p.auth += q.hub
11     for each page p in G do // then update all
12         p.hub = 0
13         for each page r in p.outgoingNeighbors do
14             p.hub += r.auth
```

Below the code, there is a "References" section with a list of sources:

1. ^ "Introduction to Information Retrieval" (HTML). Cambridge University Press. 2008. Retrieved 2008-11-09.
2. ^ Kleinberg, Jon (1999-12). "Hubs, Authorities, and Communities" (PDF). Cornell University. Retrieved 2008-11-09.
3. ^ von Ahn, Luis (2008-10-19). "Hubs and Authorities" (PDF). 15-396: Science of the Web Course Notes. Carnegie Mellon University. Retrieved 2008-11-09.
- Kleinberg, Jon (1999). "Authoritative sources in a hyperlinked environment" (PDF). *Journal of the ACM* 46 (5): 604–632. doi:10.1145/324133.324140
- Li, L.; Shang, Y.; Zhang, W. (2002). "Improvement of HITS-based Algorithms on Web Documents" (PDF). *Proceedings of the 11th International World Wide Web Conference (WWW 2002)*. Honolulu, HI. ISBN 1-880672-20-0.

At the bottom, there is a "External links" section with a single item: "U.S. Patent 6,112,202".

A Good Hub

The Best 10 Pizza Places in Norfolk, VA

Showing 1-10 of 10 results

Businesses > Restaurants > Pizza

\$ \$ \$\$ \$\$\$ Open Now Order Pickup or Delivery All Filters

Order Pickup or Delivery Pickup 1 Yelp St., San Francisco, CA 94105

World Pizza Ad 5 reviews
Lambert's Point, Norfolk, VA 23508 (757) 489-0800
Italian, Pizza

This restaurant accepts pickup and delivery Start Order

El Rey 2 Ad 60 reviews
849 E Little Creek Rd, Norfolk, VA 23518 (757) 932-1239
Mexican

I would give this place 5 stars ... But a few of the servers and staff are more concerned with facebook or twitter or whatever else is on their phone then checking on the... [read more](#)

Mo' Map Redo search when map moves

Map data ©2016 Google Terms of Use Report a map

Ads by Google related to: Pizza Norfolk, VA

PapaJohns.com
Papa John's® Online - Better Pizza, Better Deals
Introducing Our New Pan Pizza! Try One With

A Good Authority

The screenshot shows a web browser displaying the Amazon.com homepage. The search bar at the top contains the query "3 wolf moon". The main navigation bar includes links for Clothing, Shoes & Jewelry, 12 DAYS OF DEALS, Hello Michael Your Account, Prime, Lists, and Cart. Below the navigation, there are categories for Amazon Fashion, WOMEN, MEN, GIRLS, BOYS, BABY, LUGGAGE, SALES & DEALS, YOUR FASHION & S, and FREE RE. The page features a "App Only Flash Deal" for the Amazon app. A section titled "What do customers buy after viewing this item?" displays two recommended products: "Best Selling" and "Top Rated • Lowest Price". The "Best Selling" product is "The Mountain This item: Three Wolf Moon Short Sleeve T-Shirt" with a price of \$10.37. The "Top Rated • Lowest Price" product is "The Mountain Wolf Spirit Adult Balsam Green T-shirt" with a price of \$8.99. The main content area shows the product details for "The Mountain Three Wolf Moon Short Sleeve T-Shirt", including its price range (\$10.37 - \$45.99), customer reviews (3,167 reviews), and purchase options like "Add to Cart" and "Add to List".

In the Bad Old Days, Discovery Was Difficult...

- ford.com, toyota.com, etc. don't describe themselves as "automobile manufacturers", though a query for those terms arguably should return those companies

```
<title>Ford Motor Company: Cars, Trucks, SUVs, Hybrids, Parts - Ford </title>
```

```
<meta http-equiv="Content-Type" content="charset=utf-8" />
<meta name="KEYWORDS" content="ford, ford motor company, ford vehicles, ford dealers, ford motors, fo
<meta name="DESCRIPTION" content="Ford Motor Company maker of cars, trucks, SUVs and other vehicles.
```

- harvard.edu is clearly canonical for a query of "Harvard", even though it uses the term less frequently than many other pages
- Many search engines of the day (ca. late 90s) could not "find themselves"

HITS Algorithm

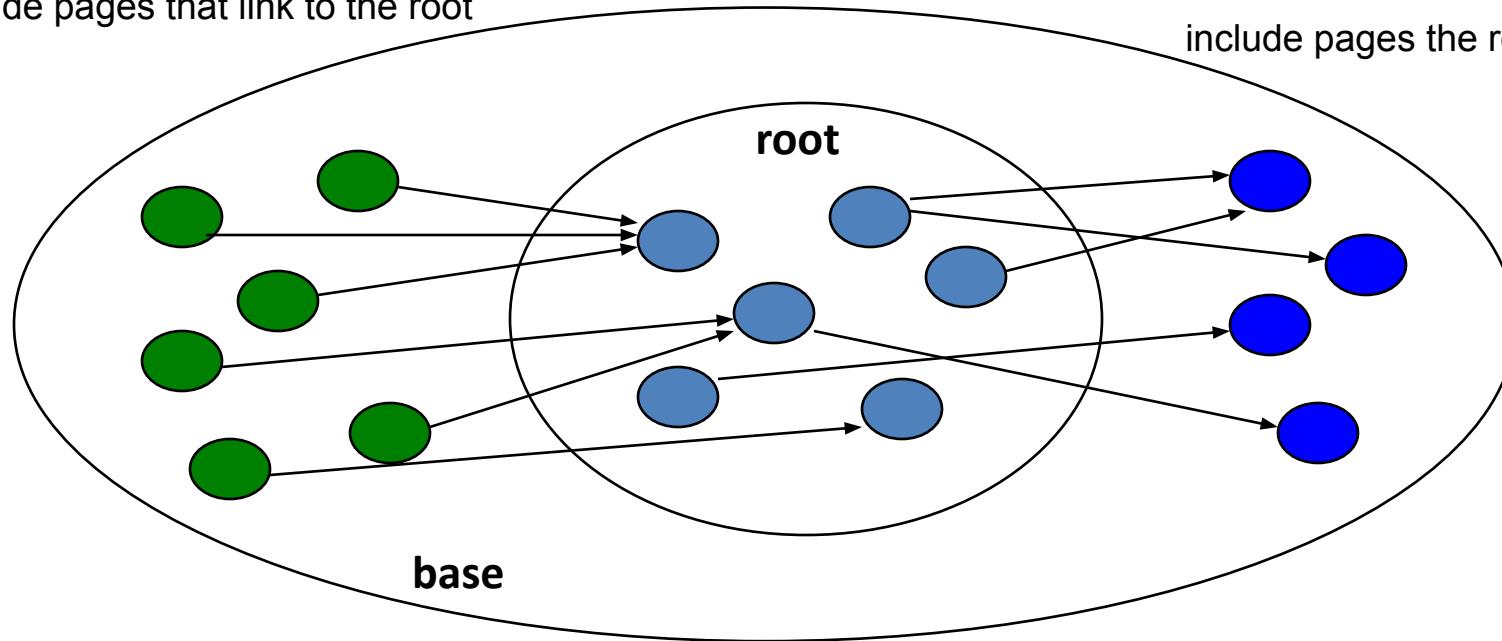
1. Retrieve pages most relevant to search query¹ → *root set*
2. Retrieve all pages linked to/from the root set → *base set*
3. Perform authority and hub calculations iteratively on all nodes in the subgraph
4. When finished, every node has an authority score and hub score

¹ note: you apply HITS to other search engines

Idea: Use Initial Search Results as "Root"

include pages that link to the root

include pages the root links to



now you have a subgraph to work with

Calculate H and A

E is set of all directed edges in subgraph

e_{qp} is edge from page q to p

$$H(p) = \sum_{q: e_{pq} \in E} A(q)$$

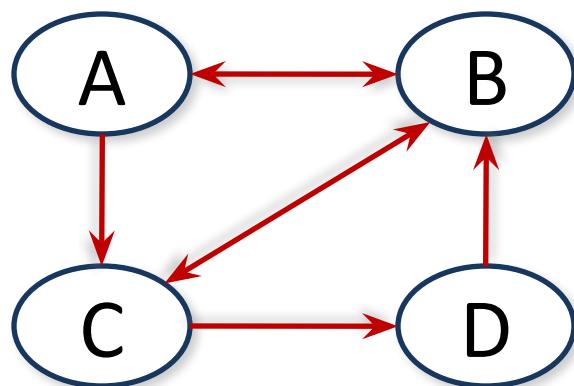
outlinks

$$A(p) = \sum_{q: e_{qp} \in E} H(q)$$

inlinks

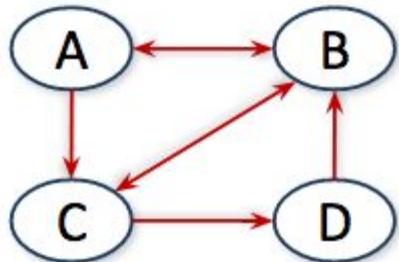
- H and A scores computed repetitively until they converge, about 10-15 iterations
- Can also be calculated efficiently using matrix multiplication

Example Subgraph



	A	B	C	D
A	0	1	1	0
B	1	0	1	0
C	0	1	0	1
D	0	1	0	0

Adjacency matrix



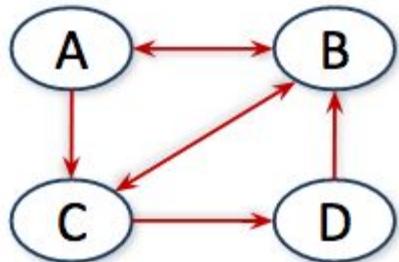
Authority Scores

$$\begin{bmatrix} A_A \\ A_B \\ A_C \\ A_D \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

Transposed adjacency matrix

Authority scores Transposed adjacency matrix Initial hub scores Resulting auth scores

Best authority



Hub Scores

$$\begin{bmatrix} H_A \\ H_B \\ H_C \\ H_D \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ 4 \\ 3 \end{bmatrix}$$

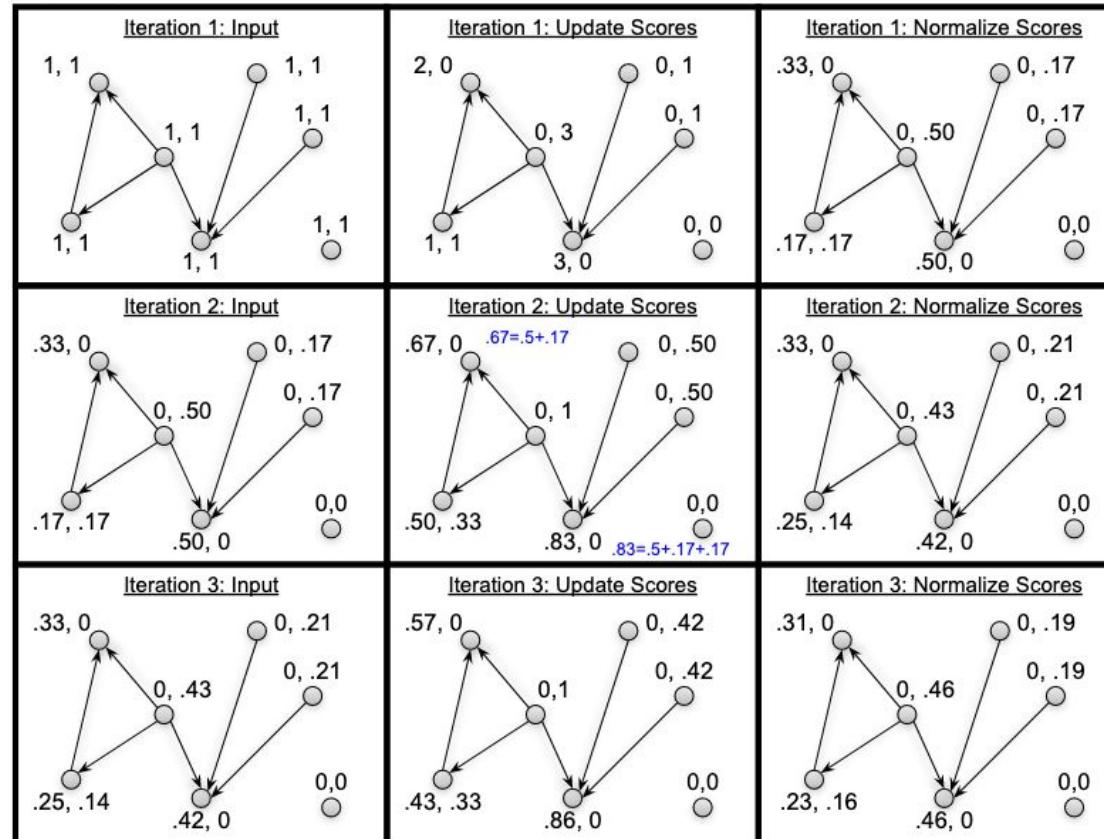
Hub scores Adjacency matrix Init auth scores Resulting hub scores

Best hub →

Fig 10.3 from [Search Engines: Information Retrieval in Practice](#)

HITS Example

(Authority, Hub)



Problems with HITS

- Has not been widely used
 - IBM holds patent
- Query dependence
 - Later implementations have made query independent
- Topic drift
 - Pages in expanded base set may not be on same topic as root pages
 - Solution is to examine link text when expanding

Objectives

- Describe the main steps needed for web search.
- Describe what a web crawler does.
- Explain the difference between precision and recall.
- Explain the difference between false positives and false negatives.
- Explain TF-IDF and how it is used to determine relevance in web search.
- Explain the importance of inlinks and outlinks in the Page Rank algorithm.
- Given a query term, compute TF-IDF for the term in a set of documents.