

How Does a Search Engine Work? Part 1

Frank McCown
Introduction to Web Science
Harding University
Spring 2011

What we'll examine

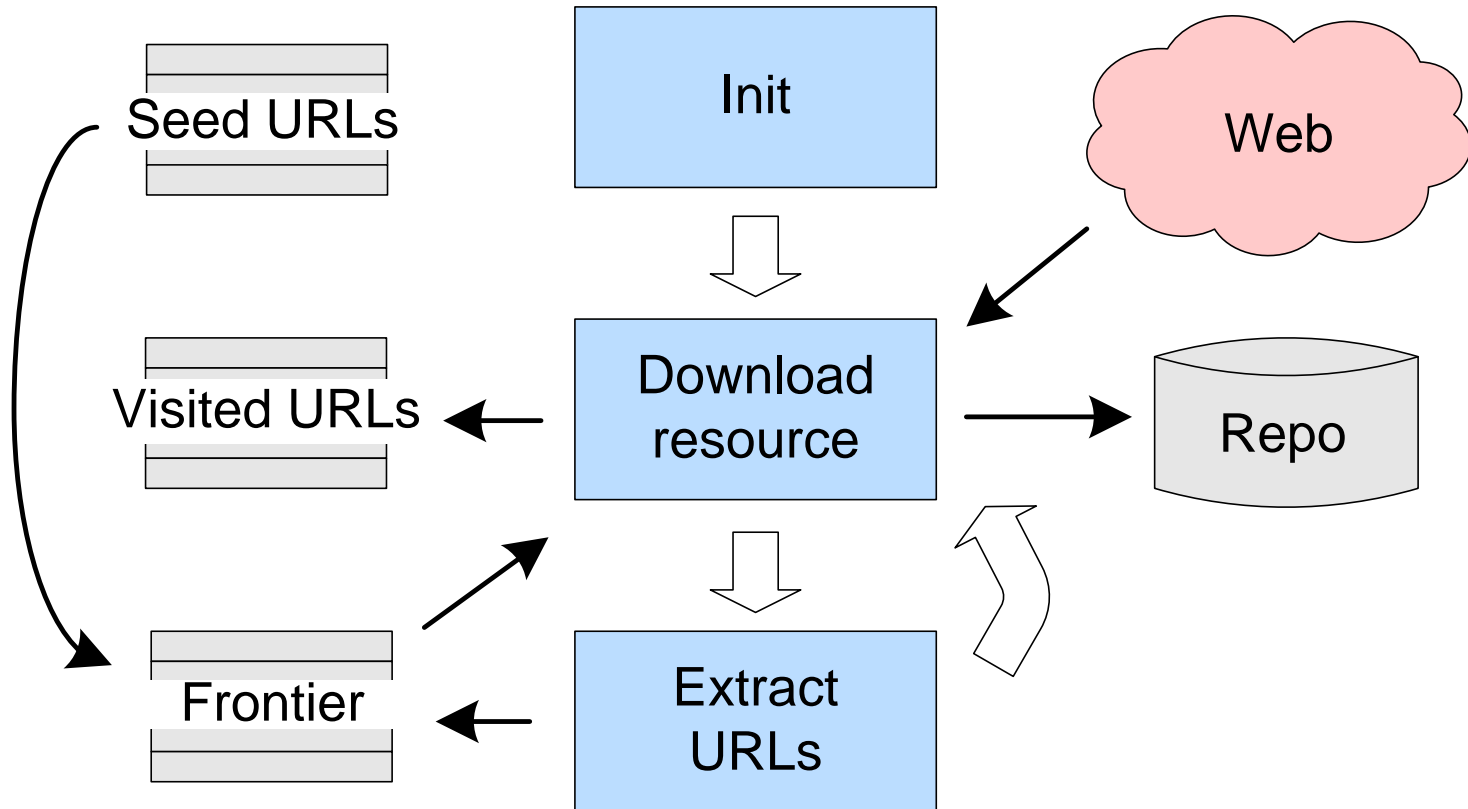
- Web crawling
- Building an index
- Querying the index
- Term frequency and inverse document frequency
- Other methods to increase relevance

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
```

Web Crawler Components



Crawling Issues

- Good source for seed URLs:
 - Yahoo or ODP web directory
 - 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

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>
```

Focused Crawling

- A vertical search engine focuses on a subset of the Web
 - Google Scholar – scholarly literature
 - ShopWiki – 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

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, PS, etc.) are converted into tokens
- Stop words may be removed
 - Frequently occurring words like a, the, and, to, etc.
 - 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

- Example:

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

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

↑
term list

Example Search

- Search for *what is it* is interpreted by search engines as *what AND is AND it*
- what: {1, 2} is: {1, 2, 3} it: {1, 2, 3}
- $\{1, 2\} \cap \{1, 2, 3\} \cap \{1, 2, 3\} = \{1, 2\}$
- Answer: Docs 1 and 2
- What if we want phrase “what is it”?

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

Phrase Search

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

Doc 1: It is what it was.

it (1,1) (1,4) (2,3) (3,1)

Doc 2: What is it?

is (1,2) (2,2) (3,2)

Doc 3: It is a banana.

what (1,3) (2,1)

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: Doc 2
- Position can be used to give higher scores to terms that are closer
 - “red cars” scores higher than “red bright cars”

Term Frequency

- Suppose page A and B both contain the search term *dog* but it appears in A three times and in B twice
- Which page should be ranked higher?
- What if page A contained 1000 words and B only 10?
- Term frequency is helpful, but it should be normalized by dividing by total number of words in the document (other divisors possible)
- TF is susceptible to spamming, so SEs look for unusually high TF values when looking for spam

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
- $IDF(\text{term}) = \frac{\text{total docs in corpus}}{\text{docs with term}}$
- Low frequency terms will have high IDF

Inverse Document Frequency

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

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

- 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
- $TF\text{-}IDF = TF \times IDF$
= 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:
 - $\text{Score} = TF\text{-}IDF(\text{term}_1) + TF\text{-}IDF(\text{term}_2) + \dots + TF\text{-}IDF(\text{term}_n)$

TF-IDF Example

- Using Bing, compute the TF-IDF scores for 2 documents that contain the words *harding* AND *university*
- Assume Bing has 20 billion documents indexed
- 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 image shows a screenshot of a Bing search results page for the query "harding university". The browser window title is "harding university - Bing" and the address bar shows the search URL. The search results are displayed under the "Web" tab, showing 1-10 of 9,390,000 results. The main result is for "Harding University - Faith, Learning, Living", with the official website "www.harding.edu". Below the main result, there are sections for "Quick Access" (Customer service 800-477-4407) and "School Information" (Location: Searcy, Arkansas; Setting: Distant Town; Type: Private not-for-profit; Level: Four or more years). To the left, there is a "RELATED SEARCHES" section with links to "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", and "Missouri Southern State University". To the right, there is a "Sponsored sites" section for "Harding University" with the text "Further your education University. Request fre www.collegesurfing.com See your message here".

harding university - Bing

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

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

bing

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

Sponsored sites

Harding University
Further your education
University. Request fre
www.collegesurfing.com
See your message here

Harding University - Faith, Learning, Living
www.harding.edu - Official site
Harding Campuses: Nursing graduates honored at pinning...

HARDING

Athletics
Students & Employees
Admissions & Aid

Harding University
News & Events
Feeds

Quick Access
Customer service 800-477-4407

School Information
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 Kirwa has won the last three NCAA II South Region individual

SEARCH HISTORY

2) Search for *harding*

The image shows a screenshot of a web browser displaying a Bing search results page for the query "harding". The browser's address bar shows the URL "www.bing.com/search?q=harding&go=&form=QBRE&q=n&sk=&sc=8-7". The search bar contains the text "harding". Below the search bar, there are tabs for "Web", "News", "Images", "Videos", and "More". The search results are displayed in a grid format. The first result is "Harding University - Faith, Learning, Living" with a sub-heading "Harding Campuses: Nursing graduates honored at pinning ceremony; Young Adult Author Series continues with Gary Schmidt ; New associate dean named for College of Business". The second result is "Warren G. Harding - Wikipedia, the free encyclopedia" with a sub-heading "Early life - Political career - Presidency: 1921-1923 - Personal controversies". The text "1-10 of 12,200,000 results" is circled in red. The browser window title is "harding - Bing".

harding - Bing

www.bing.com/search?q=harding&go=&form=QBRE&q=n&sk=&sc=8-7

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

bing

harding

Web News Images Videos More ▼

RELATED SEARCHES

- Tonya Harding
- Warren Harding
- Harding University
- Searcy Arkansas
- Harding's Marketplace
- Harding Real Estate
- Harding Connectors
- Harding Township
- Harding's Coaches

SEARCH HISTORY

- harding
- harding university
- frank mccown

See all

ALL RESULTS 1-10 of 12,200,000 results [Advanced](#)

[Harding University - Faith, Learning, Living](#)
Harding Campuses: Nursing graduates honored at pinning ceremony; Young Adult Author Series continues with Gary Schmidt ; New associate dean named for College of Business
[www.harding.edu](#) - Cached page

- Athletics
- Students & Employees
- Admissions & Aid
- Harding University
- News & Events
- Feeds
- Spiritual Life
- Majors & Minors

Show more results from [www.harding.edu](#)

[Warren G. Harding - Wikipedia, the free encyclopedia](#)
Early life - Political career - Presidency: 1921-1923 - Personal controversies
Warren Gamaliel **Harding** (November 2, 1865 – August 2, 1923) was the 29th President of the United States, from 1921 until his death in 1923. A Republican from Ohio, **Harding** ...
[en.wikipedia.org/wiki/Warren_G._Harding](#) - Cached page

2) Search for *university*

The screenshot shows a Bing search results page for the query "university". The browser address bar displays "www.bing.com/search?q=university&go=&form=QBRE&qsn=&sk=&sc=8-10". The search bar contains the text "university". Below the search bar, the "Web" tab is selected, and the results show "1-10 of 439,000,000 results" (the number "439,000,000" is circled in red). The results include sponsored sites such as "University of Phoenix", "Top Online Universities", and "Top Online University". A section titled "Listings for University near Searcy, Arkansas" is visible at the bottom, listing "Harding University" and "Arkansas State University".

university - Bing

www.bing.com/search?q=university&go=&form=QBRE&qsn=&sk=&sc=8-10

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

bing

university

Web Local News Videos Images More

RELATED SEARCHES

- Where Is Auburn University?
- University Of Phoenix Stadium
- Yale University
- Howard University
- Oxford University
- Karachi University
- Private University
- Princeton University

SEARCH HISTORY

- harding
- harding university
- frank mccown

ALL RESULTS 1-10 of 439,000,000 results Advanced

University of Phoenix Sponsored sites

Phoenix.edu - Accredited Online Degrees at **University** of Phoenix®. Learn More Today.

[Top Online Universities](#)

ClassesandCareers.com/Universities - You May Qualify For Grants To Earn Your Degree Online. Start Today!

[Top Online University](#)

www.Education180.com - Take Classes and Earn Your College Degree 100% Online! Financial Aid.

Sponsored sites

[University Degree](#)

Advance Your Career w Degree. Financial Aid A DegreeGuide.com/Univ

[Universities](#)

Find Local Universities Online Options Availabl Top.Schools.com/Unive

See your message here

Listings for **University** near **Searcy, Arkansas** Change location

1. **Harding University** - Website - (501) 279-4000
900 E Center Ave - Searcy- Directions

2. **Arkansas State University** - (501) 207-4090

Doc 1:

<http://www.harding.edu/pharmacy/>

- Copy and paste into MS Word or other word processor to obtain number of words and count occurrences
- $TF(\text{harding}) = 19 / 967$
- $IDF(\text{harding}) = \log_2(20B / 12.2M)$
- $TF(\text{university}) = 13 / 967$
- $IDF(\text{university}) = \log_2(20B / 439M)$
- $TF-IDF(\text{harding}) + TF-IDF(\text{university}) =$
 $0.020 \times 10.680 + 0.012 \times 5.510 = 0.280$

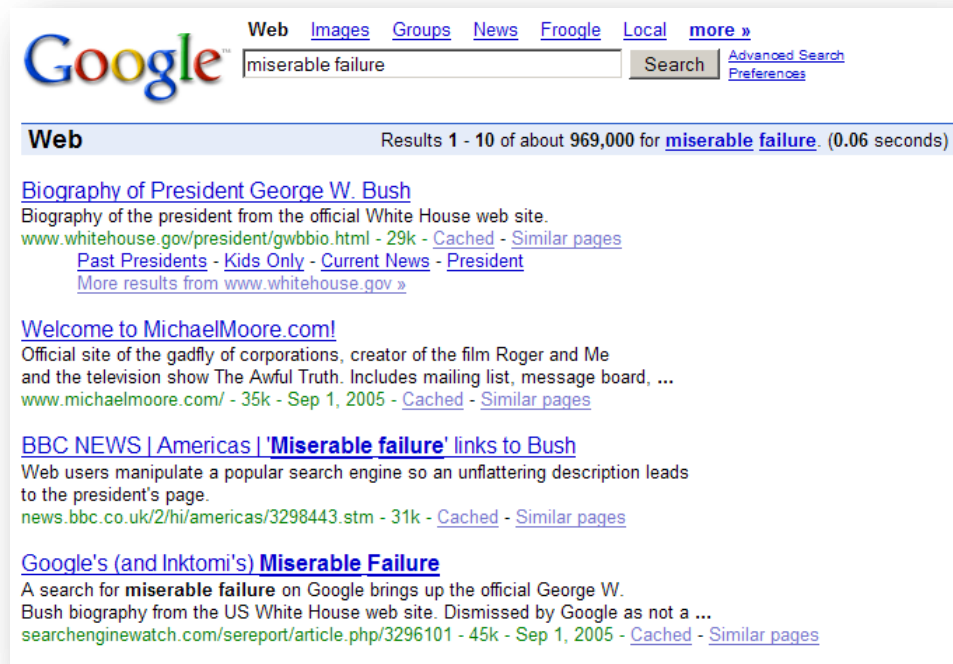
Doc 2:

http://en.wikipedia.org/wiki/Harding_University

- $TF(\text{harding}) = 44 / 3,135$
- $IDF(\text{harding}) = \log_2(20B / 12.2M)$
- $TF(\text{university}) = 25 / 3,135$
- $IDF(\text{university}) = \log_2(20B / 439M)$
- $TF-IDF(\text{harding}) + TF-IDF(\text{university}) =$
 $0.014 \times 10.680 + 0.008 \times 5.510 = 0.194$
- Doc 1 = 0.280 so it has higher score

Increasing Relevance

- Index link's anchor text with page it points to
 - `Ninja skills`
 - Watch out: Google bombs



The screenshot shows a Google search interface with the search term "miserable failure" entered. The search results are displayed under the "Web" tab, showing the first 10 results. The top result is the "Biography of President George W. Bush" from the official White House website. The second result is "Welcome to MichaelMoore.com!". The third result is a BBC News article titled "BBC NEWS | Americas | 'Miserable failure' links to Bush". The fourth result is a search engine watch article titled "Google's (and Inktomi's) Miserable Failure".

Google Web Images Groups News Froogle Local more »
miserable failure Search Advanced Search Preferences

Web Results 1 - 10 of about 969,000 for **miserable failure**. (0.06 seconds)

[Biography of President George W. Bush](#)
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - [Cached](#) - [Similar pages](#)
[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)
[More results from www.whitehouse.gov »](#)

[Welcome to MichaelMoore.com!](#)
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...
www.michaelmoore.com/ - 35k - Sep 1, 2005 - [Cached](#) - [Similar pages](#)

[BBC NEWS | Americas | 'Miserable failure' links to Bush](#)
Web users manipulate a popular search engine so an unflattering description leads to the president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - [Cached](#) - [Similar pages](#)

[Google's \(and Inktomi's\) Miserable Failure](#)
A search for **miserable failure** on Google brings up the official George W. Bush biography from the US White House web site. Dismissed by Google as not a ...
searchenginewatch.com/sereport/article.php/3296101 - 45k - Sep 1, 2005 - [Cached](#) - [Similar pages](#)

Increasing Relevance

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

¹<http://googlewebmastercentral.blogspot.com/2010/04/using-site-speed-in-web-search-ranking.html>