



Information Retrieval

BITS Pilani
Pilani Campus

Abhishek
March 2020



CS F469, Information Retrieval

Lecture topics: Web search basics and Web crawlers

This Lecture



- Web search basics (Chapter 19, IIR)
 - Big picture
 - Ads
 - Duplicate detection
 - Spam detection
 - Web IR

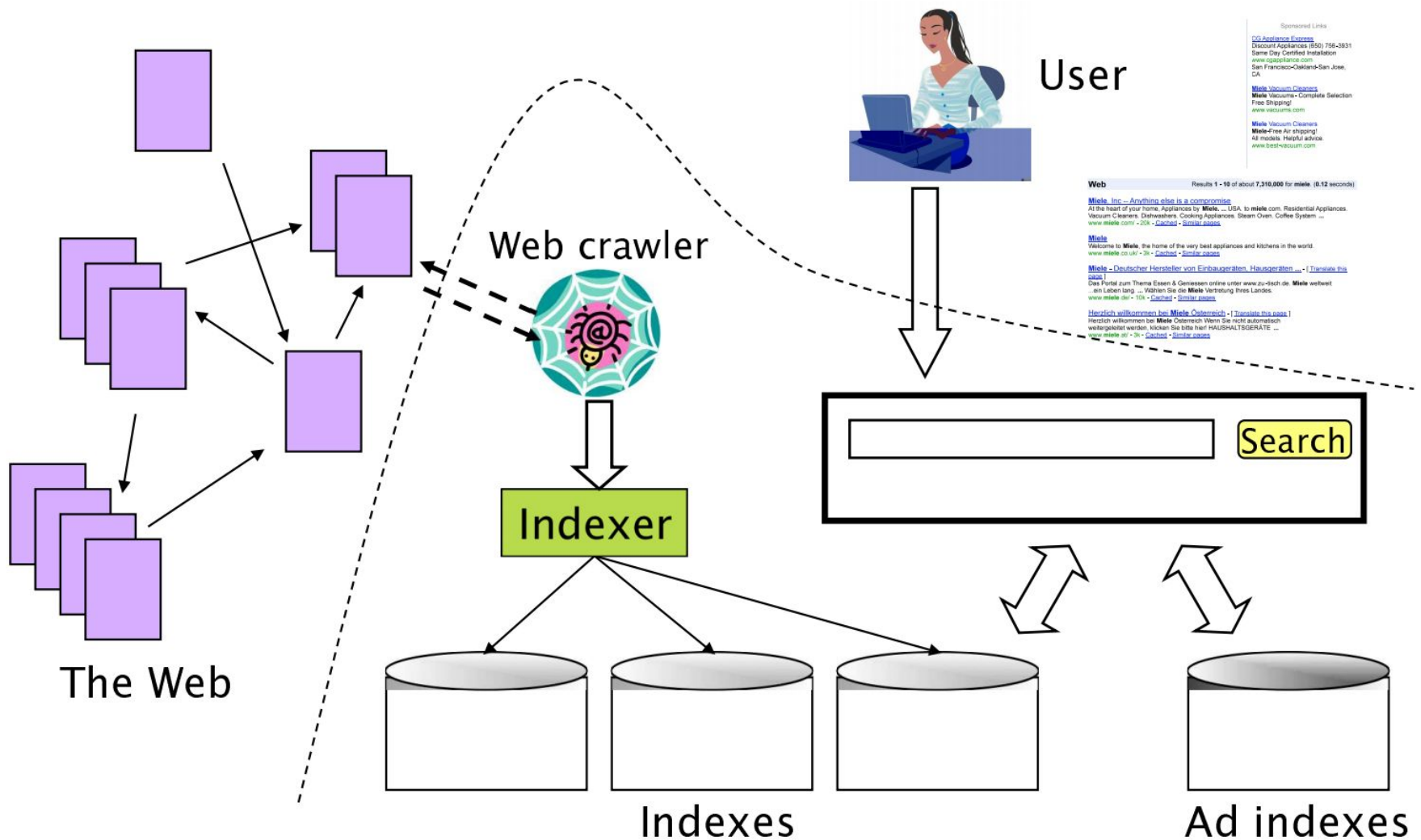
Web Search



Web search is far harder than searching “traditional” documents

- Scale
- Lack of coordination in its creation
- Diversity of backgrounds and motives of its participants

Web Search Overview





Without search engines, the web wouldn't work

- Without search, **content is hard to find**.
- → Without search, there is **no incentive to create content**.
 - Why publish something if nobody will read it?
 - Why publish something if I don't get ad revenue from it?
- Somebody needs to pay for the web.
 - Servers, web infrastructure, content creation
 - A large part today is paid by search ads.
 - **Search pays for the web.**



Web Search Facilitates Interest Aggregation

- Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
 - Elementary school kids with hemophilia.
 - People interested in minimalist and simple Linux distribution (Arch Linux).
 - People making Kiki Challenge videos.
- Search engines are a key enabler for interest aggregation.

Web Characteristics

- Server Client architecture
- HTTP, HTML, URL
- Static vs Dynamic web pages
- Anchor text
- Web as graph
 - In-link
 - Out-link
 - Strongly connected components

Ads - they pay for the web

Advertising as the economic model



- Web-servers have a associated cost
 - Hardware
 - Internet access
 - Data-center management
- Cost is directly proportional to the server load
- If it is is free for user → Someone else is paying

Ads on Search Pages



www.amazon.in > Smartphones ▾

Smartphones: Buy Smart Mobile Phones Online at Best Prices ...

Results 1 - 24 of 96 - OnePlus 7T (Glacier Blue, 8GB RAM, Fluid AMOLED Display, 128GB Storage, 3800mAh Battery) ... Samsung Galaxy A50s (Prism Crush Black, 4GB RAM, 128GB Storage) with No Cost EMI/Additional Exchange Offers. ... 1-24 of 96 results for Electronics : Mobiles & Accessories : **Smartphones** & Basic ...




Redmi Note 8 (Neptune Blue ... · Glacier Blue, 8GB RAM, Fluid ... · Mi · 1000 - ₹5000

Top stories



See smartphone

Sponsored ⓘ

		
Samsung Galaxy S20 Cosmic gra...	Samsung Galaxy S20 Ultra Cosmi...	Samsung Galaxy S9 Midnight...
₹66,999	₹92,999	₹26,999
Samsung.com	Samsung.com	Samsung.com
★★★★★ (107)	★★★★★ (367)	★★★★★ (9k+)

Mobile Klean UV Sanitizer - Clean Your Tech Gadgets Today AD

mobileklean.com | Report Ad

Mobile Klean Clean & Sanitize All Your Gadgets. Special 50% Off For Limited Time

Promotional Smartphone Wallet W/ Stock Card- \$1.59 AD

anypromo.com | Report Ad

Imprinted Giveaways Low Price. 6% Off All New Customers







Smartphone - Wikipedia

W <https://en.wikipedia.org/wiki/Smartphone>

The development of the **smartphone** was enabled by several key technological advances. The exponential scaling and miniaturization of MOSFETs (MOS transistors) down to sub-micron levels during the 1990s-2000s (as predicted by Moore's law) made it possible to build portable smart devices such as **smartphones**, as well as enabling the transition from analog to faster digital wireless mobile ...

See smartphone

Ads ⓘ

					
Apple iPhone 11 Pro - 256G...	Apple iPhone 11 Pro - 64GB...	Apple iPhone 11 - 128GB - ...	Samsung Galaxy A10s...	Samsung Galaxy A10s...	BLU G70 Dual SIM 32GB GS...
\$38.34/mo	\$33.34/mo	\$25.00/mo	\$139.99	\$139.99	\$109.00
For 30 months	For 30 months	For 30 months	✓ B&H Photo...	✓ B&H Photo...	✓ B&H Photo...
AT&T	AT&T	AT&T	★★★★★ 5	★★★★★ 2	Free shipping

Create and Edit All Day - More Hours of Power

<https://www.samsung.com> ▾

AD Power Back Up in a Snap Thanks to the Fast-Charging Battery. Learn More Today. Your Ideas Don't Stop, and Neither Does Your 2-in-1. Purchase the Galaxy Tab S6.

Ads Pricing Schemes










- CPM: Cost per mile or cost per thousand
 - More towards promotions
- CPC: Cost per click
 - More towards transactions

How are ads ranked?



See smartphone

Sponsored ⓘ

 <p>Samsung Galaxy S20 Cosmic gra... ₹66,999 Samsung.com ★★★★★ (107)</p>	 <p>Samsung Galaxy S20 Ultra Cosmi... ₹92,999 Samsung.com ★★★★★ (367)</p>	 <p>Samsung Galaxy S9 Midnight... ₹26,999 Samsung.com ★★★★★ (9k+)</p>
 <p>Galaxy M30 Black 16.21c... ₹9,999 Samsung.com ★★★★★ (267)</p>	 <p>Samsung Galaxy M31 6GB RAM 6... ₹15,999 Samsung.com Free delivery</p>	 <p>Lenovo Tab V7- Qualcomm... ₹12,999 Lenovo India Free delivery</p>
 <p>Samsung Galaxy S10 8GB... ₹54,900 Samsung.com</p>	 <p>Galaxy M30s Black 4GB... ₹13,999 Samsung.com</p>	 <p>Speria P30 Pro 6'15 Inch With 6gb... ₹4,499 bookmyphone</p>

How are ads ranked?

- Advertisers bid for keywords – **sale by auction**.
- Open system: Anybody can participate and bid on keywords.
- How does the auction determine an ad's **rank** and the **price paid** for the ad?
- Several auction schemes ...
- For the bottom line, this is perhaps the most important research area for search engines – computational advertising.
 - Squeezing an additional fraction of a **cent** from each ad means **billions** of additional revenue for the search engine.

How are ads ranked?

- First cut: according to bid price: Goto
 - Bad idea: open to abuse
 - Example: query [treatment for cancer?] → how to write your last will
 - We don't want to show nonrelevant or offensive ads.
- Instead: rank based on **bid price and relevance**
- Key measure of ad relevance: clickthrough rate
 - clickthrough rate = CTR = clicks per impressions
- Result: A nonrelevant ad will be ranked low.
 - Even if this decreases search engine revenue short-term
 - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.
- Other ranking factors: location, time of day, quality and loading speed of landing page
- The main ranking factor: the query

Google's Second Price Auction



advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

- **bid**: maximum bid for a click by advertiser
- **CTR**: click-through rate: when an ad is displayed, what percentage of time do users click on it? **CTR is a measure of relevance.**
- **ad rank**: $\text{bid} \times \text{CTR}$: this trades off (i) how much money the advertiser is willing to pay against (ii) how relevant the ad is
- **rank**: rank in auction
- **paid**: price paid by the advertiser. (Can be less than the bid)

Second Price Auction

Second price auction: The highest bidder pays the price bid by the second-highest bidder (plus 1 cent).

- Used by Google AdWords
- Earlier used by Facebook, which has now switched to Vickrey–Clarke–Groves auction

More reading:

<https://www.linkedin.com/pulse/why-do-second-price-auctions-work-chetan-prabhu>

Search ads: A win-win-win?



- The **search engine** company gets revenue every time somebody clicks on an ad.
- The **user** only clicks on an ad if they are interested in the ad.
 - Search engines punish misleading and non relevant ads.
 - As a result, users are often satisfied with what they find after clicking on an ad.
- The **advertiser** finds new customers in a cost-effective way

Exercise



- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- The advertiser pays for all this. How can the advertiser be cheated?
- Any way this could be bad for the user?
- Any way this could be bad for the search engine?

Not a win-win-win: Keyword arbitrage



- Buy a keyword on Google.
- Then redirect traffic to a third party that is paying much more than you are paying Google.
 - E.g., redirect to a page full of ads
- This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.
- The search engines need time to catch up with them.

Not a win-win-win: Violation of trademarks



- Example: geico
- During part of 2005:
- The search term “geico” on Google was bought by competitors.
- [Geico lost this case in the United States.](#)
- [Louis Vuitton lost similar case in Europe.](#)
- It's potentially misleading to users to trigger an ad off of a trademark if the user can't buy the product on the site.

Web Spam

The goal of spamming on the web



- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.
- One way of doing this: get your page ranked highly in search results.
- Exercise: How can I get my page ranked highly?

Spam technique: Keyword stuffing / Hidden text



- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks etc.
- Used to be very effective, most search engines now catch these.

Basic keyword stuffing example

Here at ABC123, product marketing is everything. We simply love product marketing. Good product marketing can change your company. The right product marketing can close more deals. Connect with us if you need help with product marketing.



Image source: <https://learn.g2.com/keyword-stuffing>

Spam technique: Doorway and lander pages



- Doorway page: optimized for a single keyword, redirects to the real target page.
- Lander page: optimized for a single keyword or a misspelled domain name, designed to attract surfers who will then click on ads. Eg. gogul.com

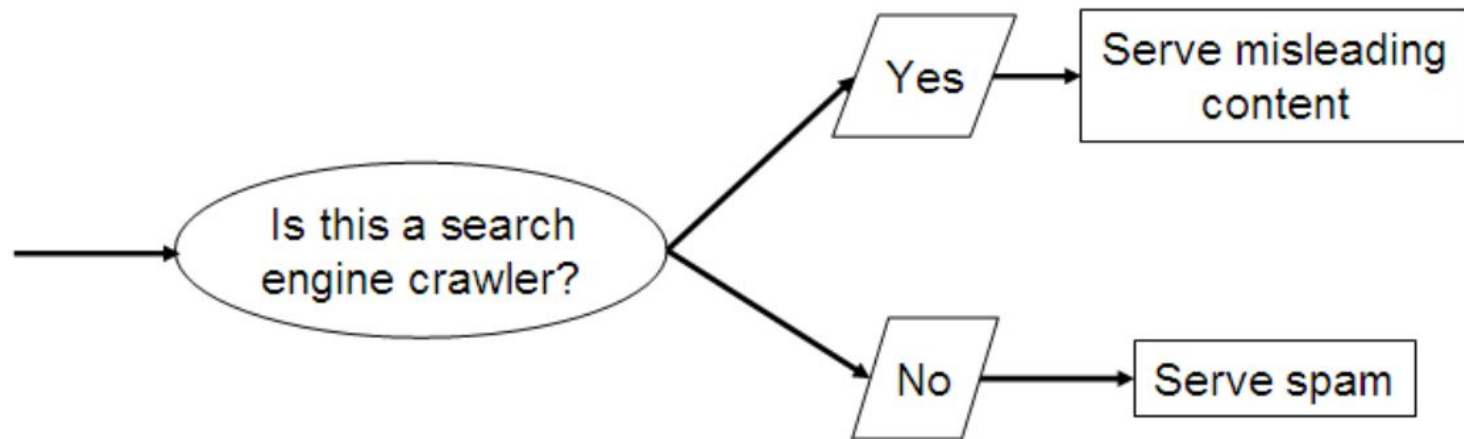


Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it.
- For example, publish the answer to a tax question with the spelling variations of “tax deferred”.

Spam technique: Cloaking

- Serve fake content to search engine spider
- So do we just penalize this always?
- No: legitimate uses (e.g., different content to US vs. European users)



Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
 - Newly registered domains (domain flooding)
 - A set of pages that all point to each other to boost each other's PageRank (mutual admiration society)
 - Pay somebody to put your link on their highly ranked page
 - Leave comments that include the link on blogs

SEO: Search engine optimization



- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
 - For example, Google bombs like [idiot](#), [Who is the failure?](#)
- And there are many legitimate ways of achieving this:
 - Restructure your content in a way that makes it easy to index
 - Talk with influential bloggers and have them link to your site
 - Add more interesting and original content

The war against spam

- Quality indicators
 - Links, statistically analyzed (PageRank etc)
 - Usage (users visiting a page)
 - No adult content (e.g., no pictures with flesh-tone)
 - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning
- Editorial intervention
 - Blacklists
 - Top queries audited
 - Complaints addressed
 - Suspect patterns detected

Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).
- There is often a fine line between spam and legitimate SEO.
- Scientific study of fighting spam on the web: [adversarial information retrieval](#)

Takeway



- Characteristics of Web
- Search Advertisements
- Spam Detection

Next:

- Duplicate detection
- Web IR (Size of web)
- Web Crawlers

Duplicates on Web

- By some estimate approx 40% pages on web are duplicates.
- Many of these are legitimate copies.
 - To provide redundancy and access reliability
 - For example:
 - <https://encyclopedia.thefreedictionary.com/Computer+science>
 - https://en.wikipedia.org/wiki/Computer_science
- Search engines tries to avoid multiple copies of same content, to keep down storage and processing overheads.

How to Detect Duplicates?

- Simplest approach: fingerprinting
 - Hashing techniques such as md5sum, etc.
- Fails to detect near duplicates.
 - The contents of one web page are identical to those of another except for a few characters – say, a notation showing the date and time at which the page was last modified.

Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want “**syntactic**” (as opposed to **semantic**) similarity.
 - True semantic similarity (similarity in content) is too difficult to compute.
- We do not consider documents near-duplicates if they have the same content, but express it with different words.
- Use similarity threshold θ to make the call “is/isn’t a near-duplicate”.
- E.g., two documents are near-duplicates if $\text{similarity} > \theta = 80\%$.

Represent each document as set of shingles



- A shingle is simply a **word n-gram**.
- Shingles are used as features to **measure syntactic similarity** of documents.
- For example, for $n = 3$, “**a rose is a rose is a rose**” would be represented as this set of shingles:
 - $\{ (a, \text{rose}, \text{is}), (\text{rose}, \text{is}, a), (\text{is}, a, \text{rose}) \}$
- We define the similarity of two documents as the **Jaccard coefficient of their shingle sets**.

Recall: Jaccard Coefficient

- A commonly used measure of overlap of two sets
- Let A and B be two sets Jaccard coefficient:

$$\text{JACCARD}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

($A \neq \phi$ or $B \neq \phi$)

- $\text{JACCARD}(A, A) = 1$
- $\text{JACCARD}(A, B) = 0$ if $A \cap B = \emptyset$
- A and B don't have to be the same size.
- Always assigns a number between 0 and 1.



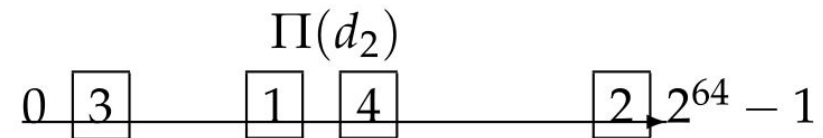
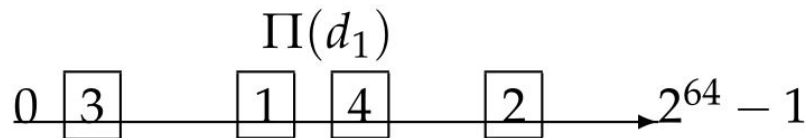
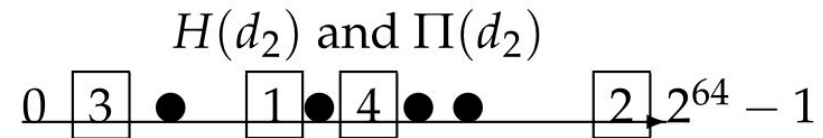
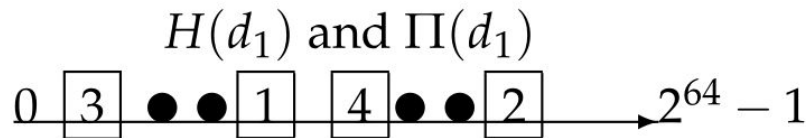
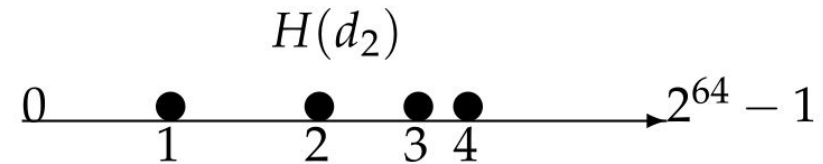
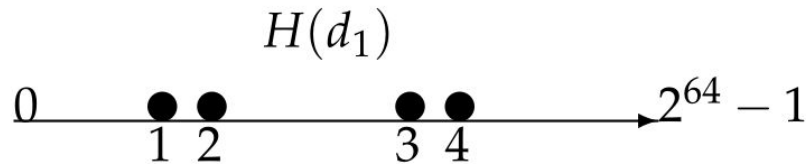
Shingles + Set Intersection

- Computing **exact** set intersection of shingles between all pairs of documents is expensive.
 - Billion of documents and each document can have thousands of shingles.
- We can map shingles to $1..2^m$ (e.g., $m = 64$) by fingerprinting.
- From now on: s_k refers to the shingle's fingerprint in $1..2^m$.
- Approximate using a cleverly chosen subset of shingles from each shingles set (a **sketch**)
- **Estimate** ($\text{size_of_intersection} / \text{size_of_union}$) based on a short sketch.

Sketch of a Document

- To increase efficiency, we will use a **sketch**, a cleverly chosen **subset** of the shingles of a document.
- The size of a sketch is, say, $n = 200$. . .
- . . . and is defined by a set of permutations $\pi_1 \dots \pi_{200}$.
- Each π_i is a random permutation on $1..2^m$
- The sketch of d is defined as:
 - $\langle \min_{s \in d} \pi_1(s), \min_{s \in d} \pi_2(s), \dots, \min_{s \in d} \pi_{200}(s) \rangle$
 - (a vector of 200 numbers).

Permutation and minimum: Example



Document 1

Document 2

Computing Jaccard for sketches (1)



- Sketches: Each document is now a vector of $n = 200$ numbers.
- Much easier to deal with than the very high-dimensional space of shingles
- But how do we compute Jaccard?

Computing Jaccard for sketches (2)



- How do we compute Jaccard?
- Let U be the union of the set of shingles of d_1 and d_2 and I the intersection.
- There are $|U|!$ permutations on U .
- For $s' \in I$, for how many permutations π do we have $\arg \min_{s \in d_1} \pi(s) = s' = \arg \min_{s \in d_2} \pi(s)$?
- Answer: $(|U| - 1)!$
- There is a set of $(|U| - 1)!$ different permutations for each s in I . $\Rightarrow |I|(|U| - 1)!$ permutations make $\arg \min_{s \in d_1} \pi(s) = \arg \min_{s \in d_2} \pi(s)$ true
- Thus, the proportion of permutations that make $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ true is:

$$\frac{|I|(|U| - 1)!}{|U|!} = \frac{|I|}{|U|} = J(d_1, d_2)$$

Estimating Jaccard

- Thus, the proportion of successful permutations is the Jaccard coefficient.
 - Permutation π is successful iff $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
- Picking a permutation at random and outputting 1 (successful) or 0 (unsuccessful) is a Bernoulli trial.
- Estimator of probability of success: proportion of successes in n Bernoulli trials. ($n = 200$)
- Our sketch is based on a random selection of permutations.
- Thus, to compute Jaccard, count the number k of successful permutations for $\langle d_1, d_2 \rangle$ and divide by $n = 200$.
- $k/n = k/200$ estimates $J(d_1, d_2)$. □

Implementation

- We use hash functions as an efficient type of permutation:
 $h_i : \{1..2^m\} \rightarrow \{1..2^m\}$
- Scan all shingles s_k in union of two sets in arbitrary order.
- For each hash function h_i and documents d_1, d_2, \dots : keep slot for minimum value found so far.
- If $h_i(s_k)$ is lower than minimum found so far: update slot

Example

	d_1	d_2
s_1	1	0
s_2	0	1
s_3	1	1
s_4	1	0
s_5	0	1

$h(x) = x \bmod 5$
 $g(x) = (2x + 1) \bmod 5$
 $\min(h(d_1)) = 1 \neq 0 =$
 $\min(h(d_2)) \quad \min(g(d_1)) =$
 $2 \neq 0 = \min(g(d_2))$
 $\hat{J}(d_1, d_2) = \frac{0+0}{2} = 0$

	d_1 slot		d_2 slot	
h	∞		∞	
g	∞		∞	
$h(1) = 1$	1	1	–	∞
$g(1) = 3$	3	3	–	∞
$h(2) = 2$	–	1	2	2
$g(2) = 0$	–	3	0	0
$h(3) = 3$	3	1	3	2
$g(3) = 2$	2	2	2	0
$h(4) = 4$	4	1	–	2
$g(4) = 4$	4	2	–	0
$h(5) = 0$	–	1	0	0
$g(5) = 1$	–	2	1	0

final sketches

Exercise



	d_1	d_2	d_3	
s_1	0	1	1	
s_2	1	0	1	$h(x) = 5x + 5 \pmod{4}$
s_3	0	1	0	$g(x) = (3x + 1) \pmod{4}$
s_4	1	0	0	

Estimate $\hat{J}(d_1, d_2)$,

$$\hat{J}(d_1, d_3), \hat{J}(d_2, d_3)$$

Solution: Slide 48 and 49. <https://www.cis.uni-muenchen.de/~hs/teach/14s/ir/pdf/19web.flat.pdf>

Shingling: Summary

- Input: N documents
- Choose n -gram size for shingling, e.g., $n = 5$
- Pick 200 random permutations, represented as hash functions
- Compute N sketches: $200 \times N$ matrix shown on previous slide, one row per permutation, one column per document
- Compute $N(N-1)/2$ pairwise similarities
- Transitive closure of documents with similarity $> \theta$
- Index only one document from each equivalence class

Efficient near-duplicate detection



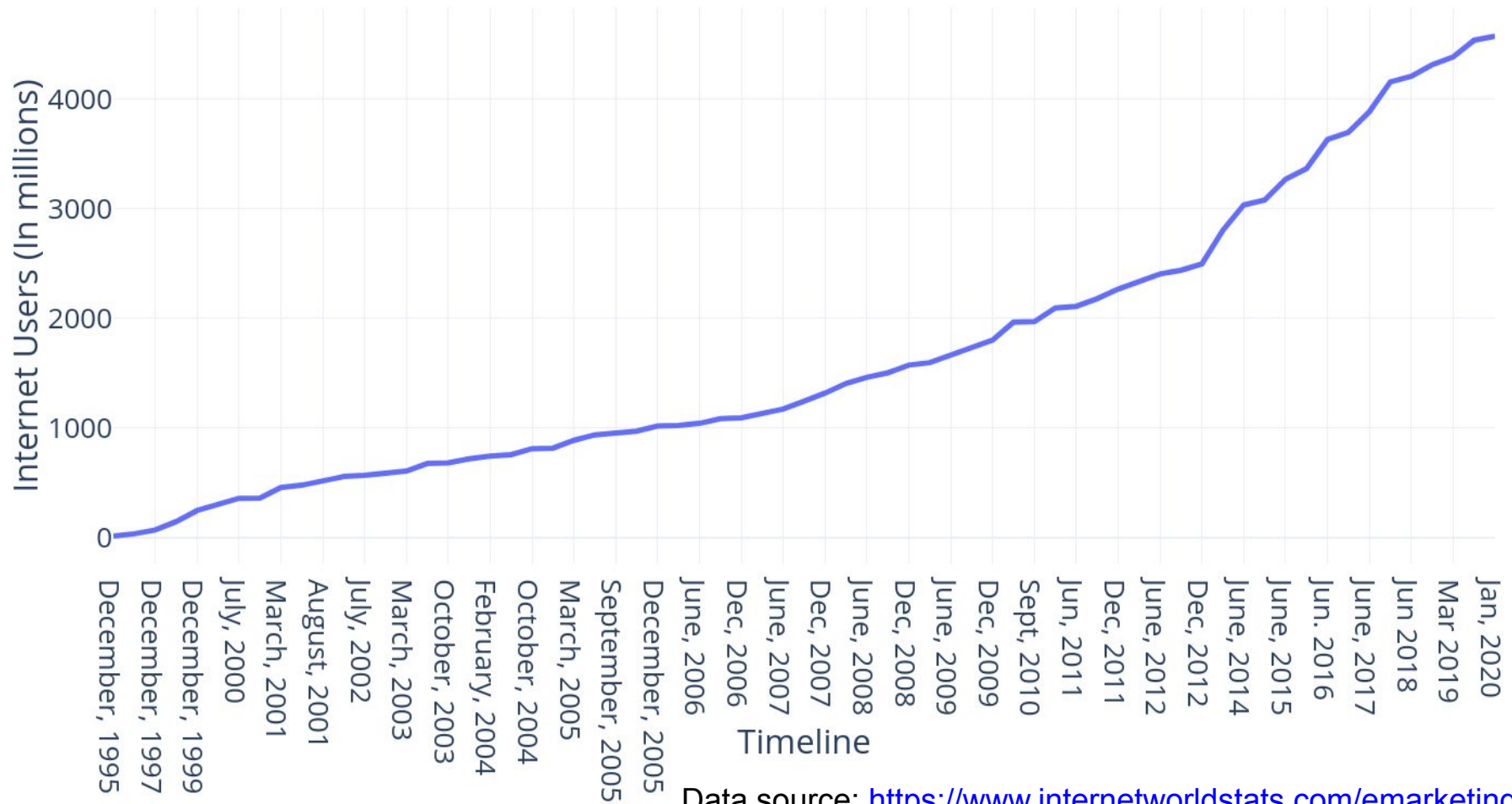
- Now we have an extremely efficient method for estimating a Jaccard coefficient for a single pair of two documents.
- But we still have to estimate $O(N^2)$ coefficients where N is the number of web pages.
- Still intractable
- Other solutions: locality sensitive hashing (LSH) and sorting
 - [Locality-sensitive hashing](#)
 - [Detecting Near-Duplicates for Web Crawling, WWW, 2007](#)
 - [Finding Near-Duplicate Web Pages: A Large-Scale Evaluation of Algorithms, SIGIR, 2016](#)

Size of Web: Hard to answer

Growth of Web: Internet Users



Growth of Web: Internet users

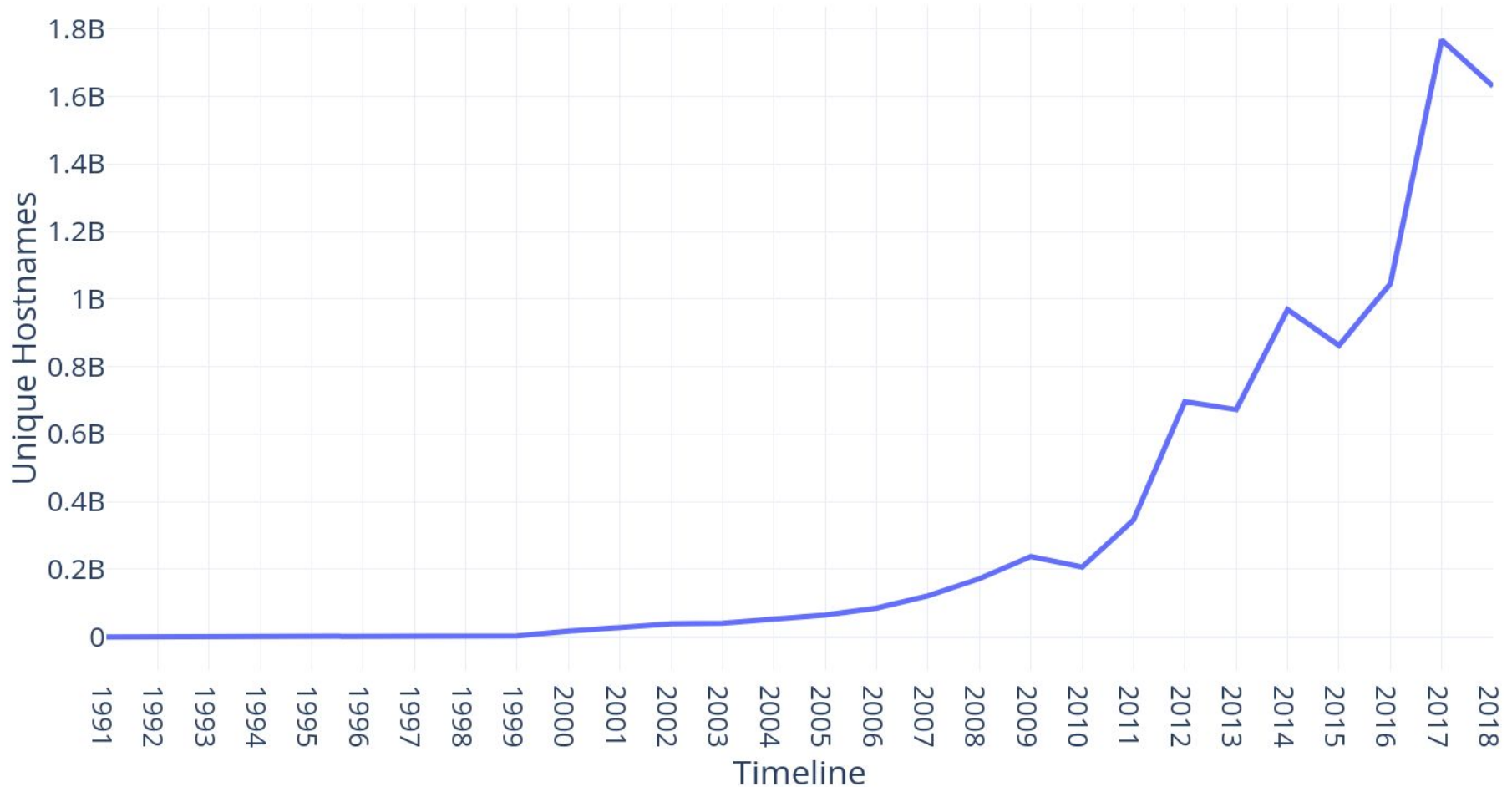


Data source: <https://www.internetworldstats.com/emarketing.htm>

Growth of Web: Unique Hostnames



Growth of Web: Hostnames



Data source: <https://www.internetlivestats.com/total-number-of-websites/#sources>

Size of the web: Issues

- What is size? Number of web servers? Number of pages? Terabytes of data available?
- Some servers are seldom connected.
 - Example: Your laptop running a web server
 - Is it part of the web?
- The “dynamic” web is infinite.
 - Any sum of two numbers is its own dynamic page on Google. (Example: “2+4”)
 - Valid 404 pages generated by `imdb.com/any_string`

“Search engine index contains N pages”: Issues



- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
- There used to be (and still are?) billions of pages that are only indexed by anchor text.

Size of the web: Who cares?



- Media
- Users
 - They may switch to the search engine that has the best coverage of the web.
 - Users (sometimes) care about recall.
- Search engine designers (how many pages do I need to be able to handle?)
- Crawler designers (which policy will crawl close to N pages?)

Size of Index

Simple method for determining a lower bound



- OR-query of frequent words in a number of languages.
- But page counts of Google search results are only rough estimates.

Exercise:

- Estimate the size of Google and Bing search index.

Better method for determining a lower bound



- Paper: [Estimating search engine index size variability: a 9-year longitudinal study](#)
- Demo: <https://www.worldwidewebsize.com/>

Key idea:

- Estimate index size by document frequency extrapolation on an external corpus.

Paper sections to read:

Estimating the size of a search engine through extrapolation

Estimating word frequencies for corpus size extrapolation

Selecting a representative corpus: DMOZ

Taking the arithmetic mean



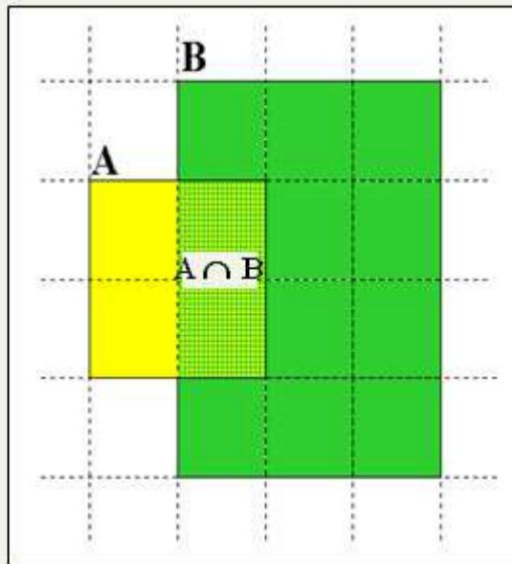
Relative sizes of Two Indexes

- There are significant differences between indexes of different search engines.
- Different engines have different preferences.
 - max url depth, max count/host, anti-spam rules, priority rules etc.
- Different engines index different things under the same URL.
 - anchor text, frames, meta-keywords, etc.

Capture-recapture method



- Assumptions:
 - There is finite set of Web from which each search engine chooses a subset.
 - The search engine chooses an independent, uniformly chosen subset.



Sample URLs randomly from A

Check if contained in B

and vice versa

$$A \cap B = (1/2) * \text{Size A}$$

$$A \cap B = (1/6) * \text{Size B}$$

$$(1/2) * \text{Size A} = (1/6) * \text{Size B}$$

$$\therefore \text{Size A} / \text{Size B} =$$

$$(1/6) / (1/2) = 1/3$$

Each test involves: (i) Sampling (ii) Checking

Sampling methods

- Random searches
- Random IP addresses
- Random walks
- Random queries

Reading: Page 435, 436 and 437

Thank You!