



THE UNIVERSITY
of EDINBURGH

Text Technologies for Data Science

INFR11145

Web Search (2)

Instructor:
Tuğrulcan “Tj” Elmas


30-Oct-2024

1

Lecture Objectives

- Learn about:
 - Basics of Web search
 - Brief History of web search
 - SEOs
 - Web Crawling (intro)

Tuğrulcan Elmas, TTDS 2024/2025



THE UNIVERSITY
of EDINBURGH

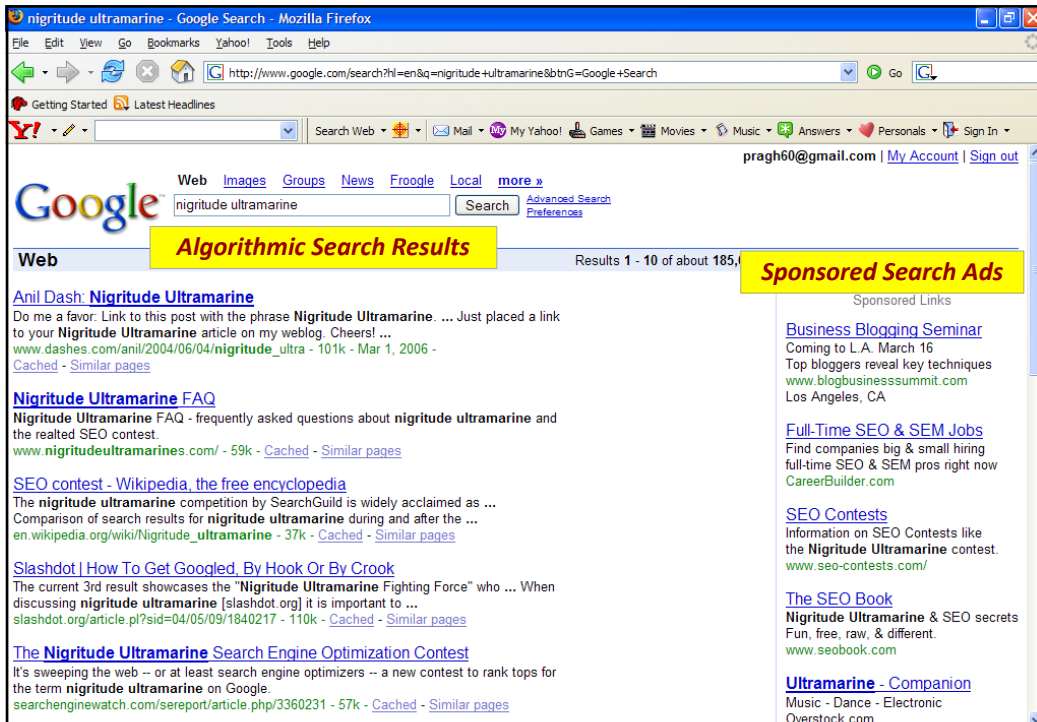
2

Brief History

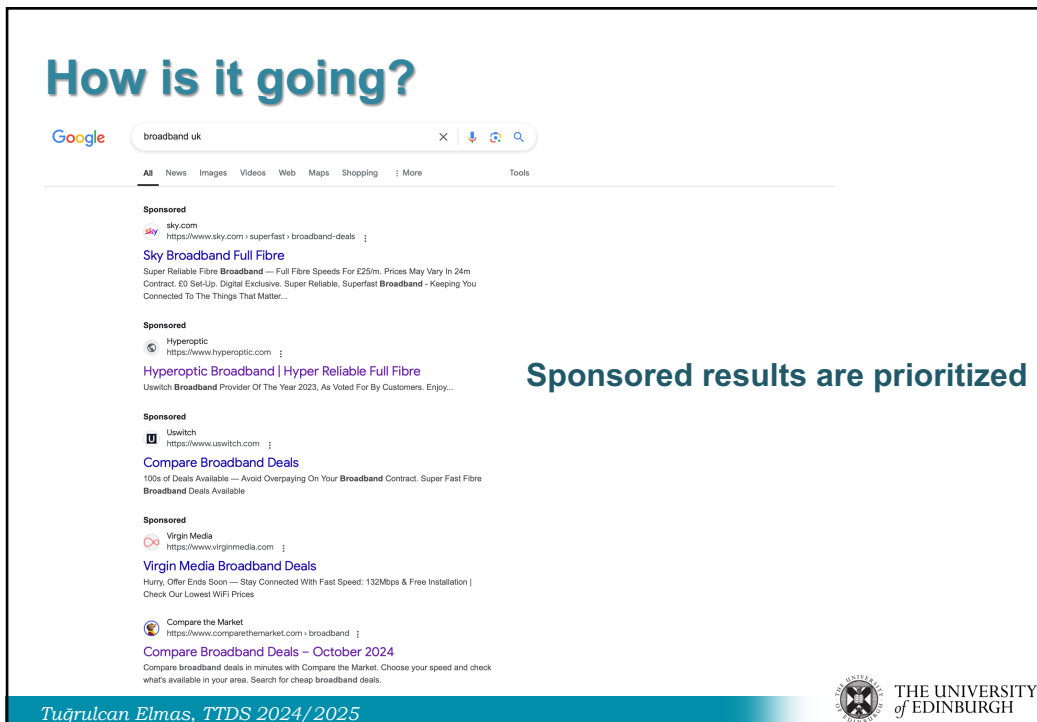
- Early keyword-based engines (1995-1997)
 - Altavista, Excite, Infoseek, Lycos, AOL
 - Traditional IR techniques
 - Scalability is an issue
- Paid search ranking: Goto (morphed into Overture.com → Yahoo!)
 - Your search ranking depended on how much you paid
 - Auction for keywords
 - Called “sponsored search”
 - CPM (Cost Per Thousand Impressions)
 - CPC (Cost Per Click)

Brief (non-technical) History

- 1998+: Link-based ranking pioneered by Google
 - Blew away all early engines
 - Great user experience in search of a business model
 - Meanwhile Goto/Overture’s annual revenues: ~ \$1 billion
- Result: Google added paid search “ads” to the side, independent of search results
 - Yahoo followed, acquiring Overture (for paid placement) and Inktomi (for search)
- 2005+: Google gains search share, dominating in Europe and very strong in North America
 - 2009: Microsoft & Yahoo Search Alliance
 - Bing’s search technology & Yahoo focusing on ads



5




6

How is it going?


Google

All Shopping Images Videos News Product sites Maps More Tools

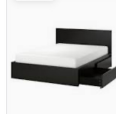
Sponsored :




DUSK Ascot Ottoman Storage Bed
£319.00
 DUSK
 +£19.99 delivery
 ★★★★★ (551)
 Bed · Double · Ottoman · Velvet
 By Crowd Sho...




Collect today
IKEA NEIDEN Standard Double Bed
£89.00
 IKEA
 ★★★★★ (2k+)
 Bed · Double · Wood
 By Swoop




SALE
Collect today
IKEA MALM Standard Double Bed
£195.00 ~~£239~~
 IKEA
 ★★★★★ (3k+)
 Bed · Double · Wood
 By Swoop




DUSK Ascot Ottoman Storage Bed
£319.00
 DUSK
 +£19.99 delivery
 ★★★★★ (551)
 Bed · Double · Ottoman · Linen
 By Crowd Sho...




Double Bed Frame Aurora
£66.99
 Wayfair.co.uk
 +£4.99 delivery
 ★★★★★ (164)
 Bed · Double
 By Productcaster



BEDZONLINE Double Bed
£99.95
 Amazon.co.uk
 Free delivery
 Double · Mattress · Orthopaedic ...
 By Kelkoo



Small Single Divan Bed with Storage
£99.00
 Beds.co.uk
 Free delivery
 Free returns
 Bed · Mattress · Headboard ...
 By Feedoptimise



DUSK Ascot Ottoman Storage Bed
£319.00
 DUSK
 +£19.99 delivery
 ★★★★★ (551)
 Bed · Double · Ottoman
 By Crowd Sho...

Shopping posts (only sponsored)

THE UNIVERSITY of EDINBURGH

Tuğrulcan Elmas, TTDS 2024/2025

7

How is it going?

Google

All Images Videos News Web Maps Books More Tools

Did you mean: what is **nigritude** ultramarine

AI Overview Learn more

The nigritude ultramarine competition was an SEO contest created by DarkBlue.com and run by SearchGuild. It is considered the first major SEO contest in the English-speaking world. The contest began on May 7, 2004, and Anil Dash won it two months later. The challenge was for webmasters to rank number one on Google for the search phrase "seraphim proudleduck" within three months.

SEO contest - Wikipedia

The first recorded SEO contest was Schnitzelmitkartoffelsalat by German webmasters, started on November 15, 2002, in th...
 W Wikipedia

[Show more](#)

In the English-language world, the nigritude ultramarine competition created by DarkBlue.com and run by SearchGuild is widely acclaimed as the mother of all SEO contests.

Wikipedia
https://en.wikipedia.org/wiki/SEO_contest

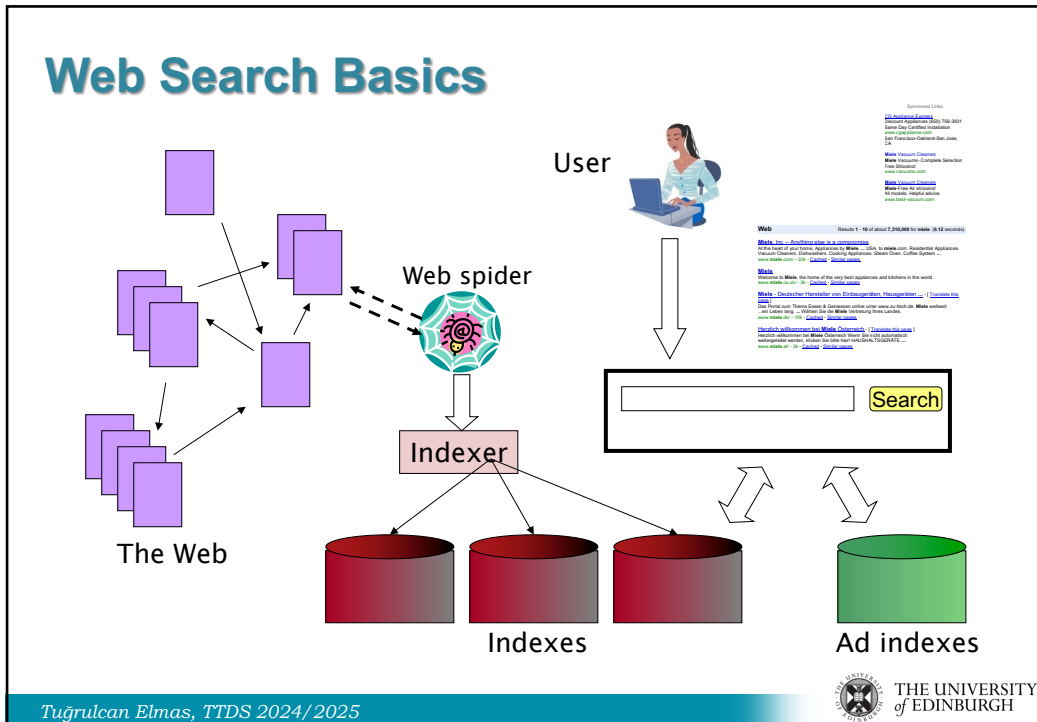
SEO contest - Wikipedia

About featured snippets · Feedback

THE UNIVERSITY of EDINBURGH

Tuğrulcan Elmas, TTDS 2024/2025

8



9

User Need on Web Search

- **Informational** – want to learn about something (~40% / 65%)
Information Retrieval
- **Navigational** – want to go to that page (~25% / 15%)
"Facebook"
- **Transactional** – want to do something (web-mediated) (~35% / 20%)
 - Access a service Seattle weather
 - Downloads Mars surface images
 - Shop Canon S410
- **Gray areas**
 - Exploratory search "see what's there"

Tuğrulcan Elmas, TTDS 2024/2025

THE UNIVERSITY of EDINBURGH

10

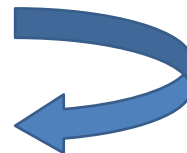
Search Engine Optimization (SEO)

- The Trouble with Paid Search Ads:
It costs money. It is explicit. What's the alternative?
- **Search Engine Optimization (SEO):**
 - “Tuning” your web page to rank highly in the algorithmic search results for selected keywords
- Performed by companies, webmasters and consultants (“Search engine optimizers”) for their clients
- Some perfectly legitimate, some very shady

SEO: Keyword Stuffing

- First generation engines relied heavily on *tf/idf*
 - The top-ranked pages for the query **William Shakespeare** were the ones containing the most **William's** and **Resort's**
- SEOs responded with dense repetitions of chosen terms
 - e.g., **William Shakespeare William Shakespeare...**
 - Misleading meta-tags, excessive repetition
 - Same color as the background of the web page (e.g., **white**)
 - Repeated terms got indexed by crawlers
 - But not visible to humans on browsers

***Pure word density cannot be trusted
as an IR signal***

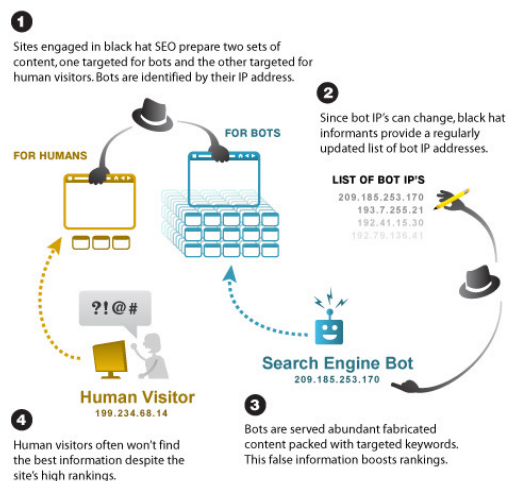


SEO keyword stuffing examples

- XYZ Hotel in ABC city
 - Accommodation, hotel, room, flat, travel, sights, attractions, vacation, holiday, in ABC ABC ABC
- XYZ for family advices
 - Family, couples, parents, spouse, wife, husband, fights, relationship, cheating, communication, kids, children
- XYZ Umbrellas
 - Raining, rainy, wet, weather, day

SEO: Black Hat Cloaking

Black Hat Cloaking Explained



Duplicate Detection

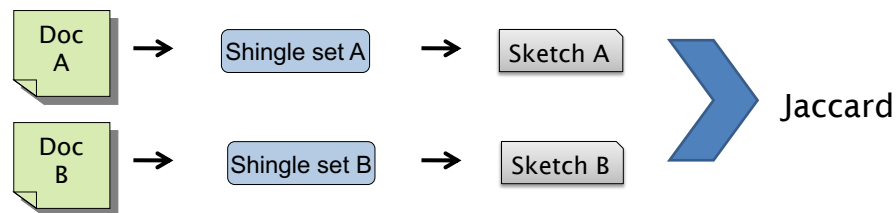
- The web is full of duplicated content
 - Can be detected by document fingerprints
- Strict duplicate detection = exact match
 - Not as common
- But many, many cases of **near duplicates**
 - e.g., Wikipedia articles
- *Near-Duplication*: Approximate match
 - Use similarity threshold to detect near-duplicates
 - e.g., Similarity > 80% => Documents are “near duplicates”
 - Not transitive though sometimes used transitively
 - $A \approx B \ \& \ B \approx C \rightarrow$ doesn't have to mean $A \approx C$

Duplicate Detection: MinHash

- Features of similarity:
 - Segments of a document (natural or artificial breakpoints)
 - **Shingles** (word n-grams)
 - *a rose is a rose is a rose* →
 - a_rose_is_a
 - rose_is_a_rose
 - is_a_rose_is
 - ~~a_rose_is_a~~
- Similarity between two docs (= sets of shingles)
 - $\frac{\text{size of intersection}}{\text{size of union}}$ (Jaccard Coefficient)

Shingles + Set Intersection

- Computing exact set intersection of shingles between all pairs of documents is expensive/intractable
- Approximate using a cleverly chosen subset of shingles from each (a sketch)
- Estimate Jaccard Coefficient based on a short sketch

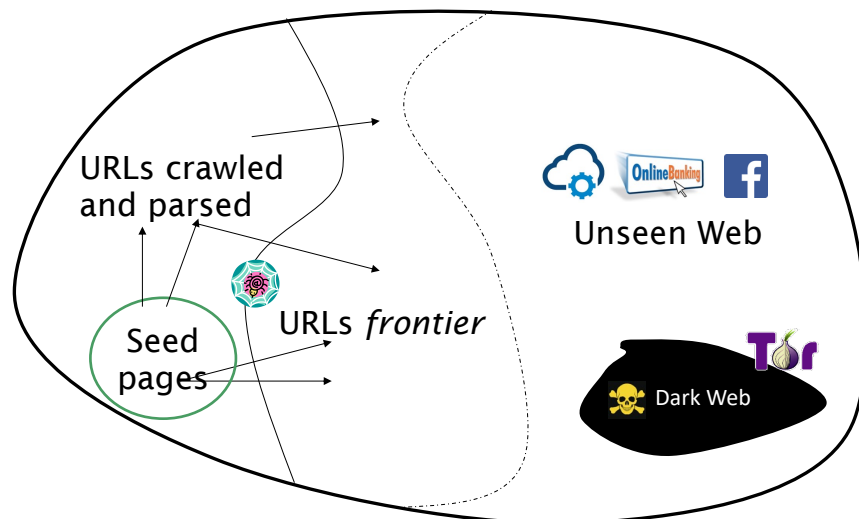


Tuğrulcan Elmas, TTDS 2024/2025



17

Web Crawling



Tuğrulcan Elmas, TTDS 2024/2025



18

Basic Crawler Operation

- Begin with known “seed” URLs
- Fetch and parse them ←
- Extract URLs they point to
- Place the extracted URLs on a queue
- Fetch one URL from the queue
- Repeat

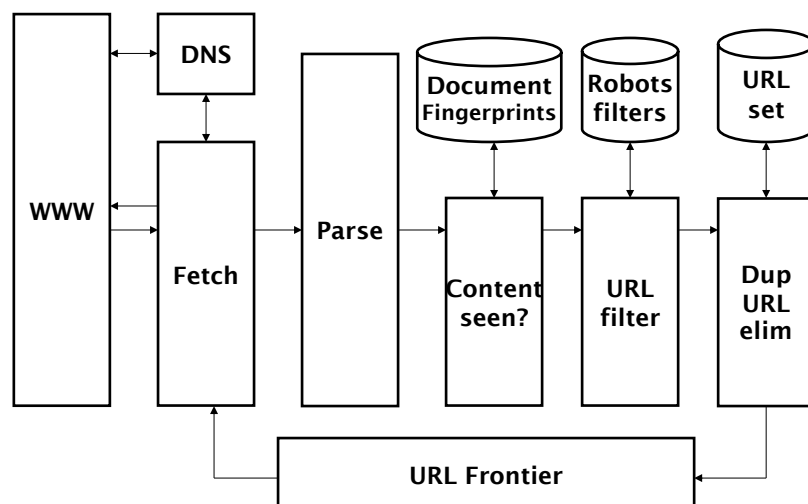
What Any Crawler Must Do

- Be Polite: Respect implicit and explicit politeness considerations
 - Only crawl allowed pages
 - respect `robots.txt`
 - Avoid hitting any site too often
- Be Robust: Be immune to spider traps and other malicious behaviour from web servers
 - Black hat cloaking, link farms etc.

What Any Crawler Should Do

- Be capable of distributed operation
 - run on multiple distributed machines
- Be scalable: increase the crawl rate by adding more machines
- Performance/efficiency: permit full use of available processing and network resources
- Fetch pages of “higher quality” first
- Freshness/Continuous operation: Continue fetching fresh copies of a previously fetched page
- Extensible: Adapt to new data formats, protocols

Basic Crawler Architecture



Processing Steps in Crawling

1. Pick a URL from the frontier
2. Fetch the document at the URL
3. Parse the document
 1. Extract links from it to other docs (URLs)
4. Check if document has content already seen
 1. If not, add to indexes
5. For each extracted URL
 1. Ensure it passes certain URL filter tests
 2. Check if it is already in the frontier (duplicate URL elimination)

URL Frontier

- Can include multiple pages from the same host
- Must avoid trying to fetch them all at the same time
- Must try to keep all crawling threads busy

Explicit and Implicit Politeness

- Explicit politeness: specifications from webmasters on what portions of site can be crawled e.g., robots.txt
- Implicit politeness: even with no specification, avoid hitting any site too often

```

User-agent: *
Disallow: /yoursite/temp/

User-agent: searchengine
Disallow:

User-agent: GPTBot
Disallow: /
  
```

robots.txt

- No robot should visit any URL starting with "/yoursite/temp/", except the robot called "searchengine"

URL Frontier: 2 Main Considerations

- Politeness: do not hit a web server too frequently
- Freshness: crawl some pages more often than others
 - Pages whose content changes often (e.g. News sites)
- These goals may conflict each other.
 - e.g., simple priority queue fails – many links out of a page go to its own site, creating a burst of accesses to that site.
- Even if we restrict only one thread to fetch from a host, can hit it repeatedly
- Common heuristic: insert time gap between successive requests to a host that is >> time taken in most recent fetch from that host

Summary

- History of Web search
- Basics of web search
- Usage of web search
- SEO
- Web crawling

Resources

- Text book 1: Intro to IR, Chapter 19
- Text Book 2: IR in Practice: Chapter 3
- YouTube Videos (nice to watch)
 - How Search Works. Google
<https://www.youtube.com/watch?v=BNHR6IQJGZs>
 - The Evolution of Search. Google
<https://www.youtube.com/watch?v=mTBShtwCnD4>
 - What Is The Deep Web?. Mashable
<https://www.youtube.com/watch?v=UOK7aRmUtw>
 - Most popular websites (search engines) over time
<https://www.youtube.com/watch?v=MirrGCbslp4>
 - This is How Much YouTube Pays Me
<https://www.youtube.com/watch?v=l3MeCEwVxB0>