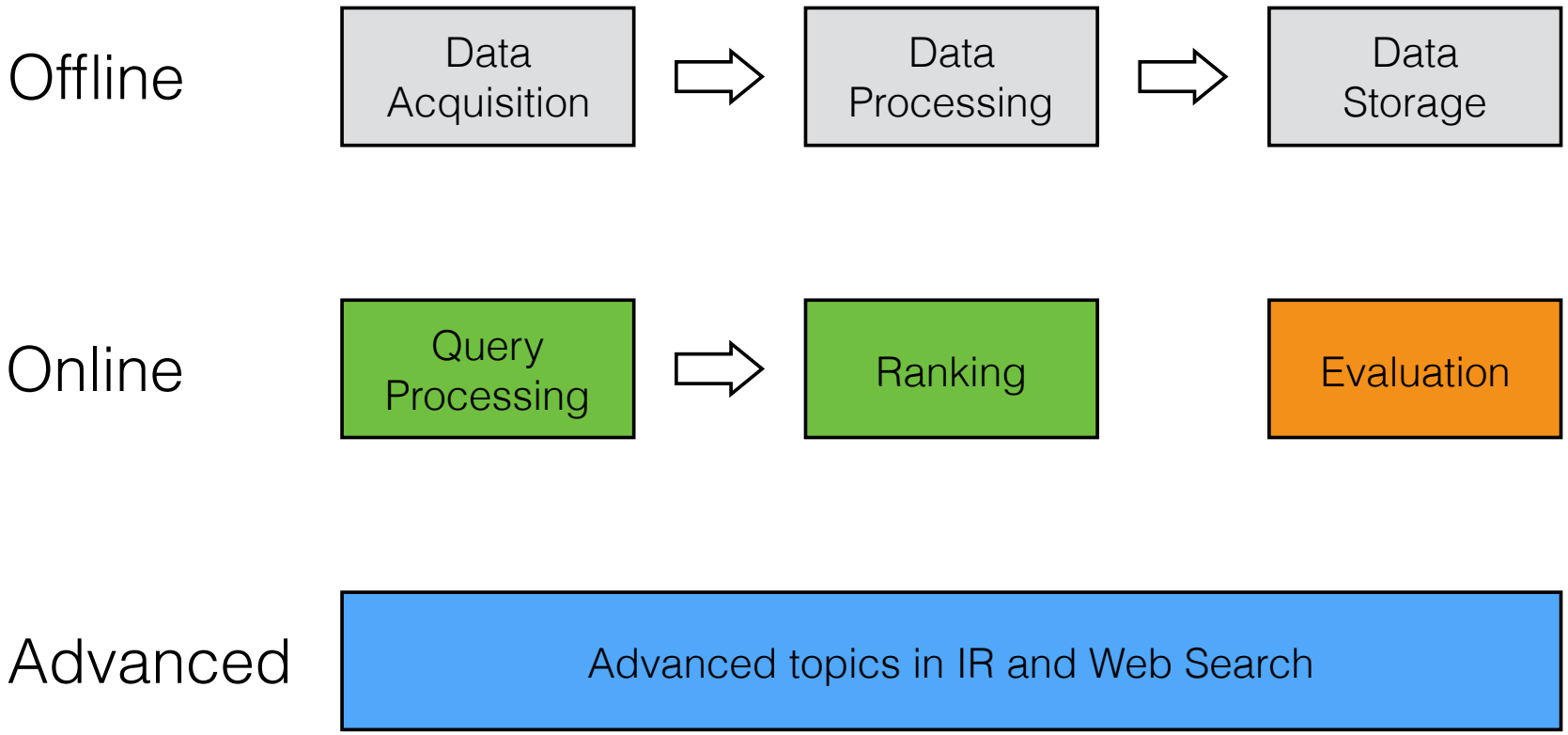
Information Retrieval

Data Acquisition

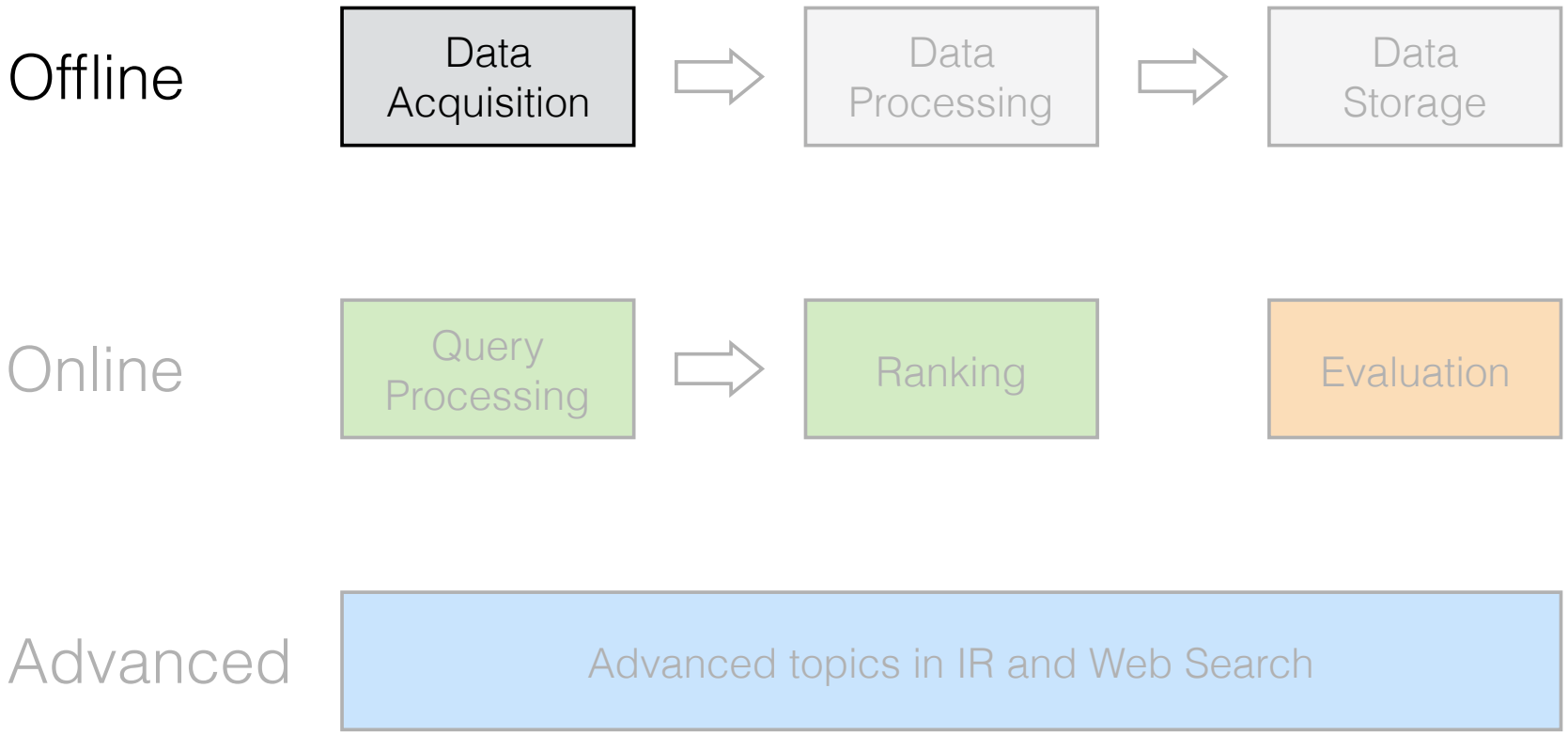
Ilya Markov
i.markov@uva.nl

University of Amsterdam

Course overview



This lecture

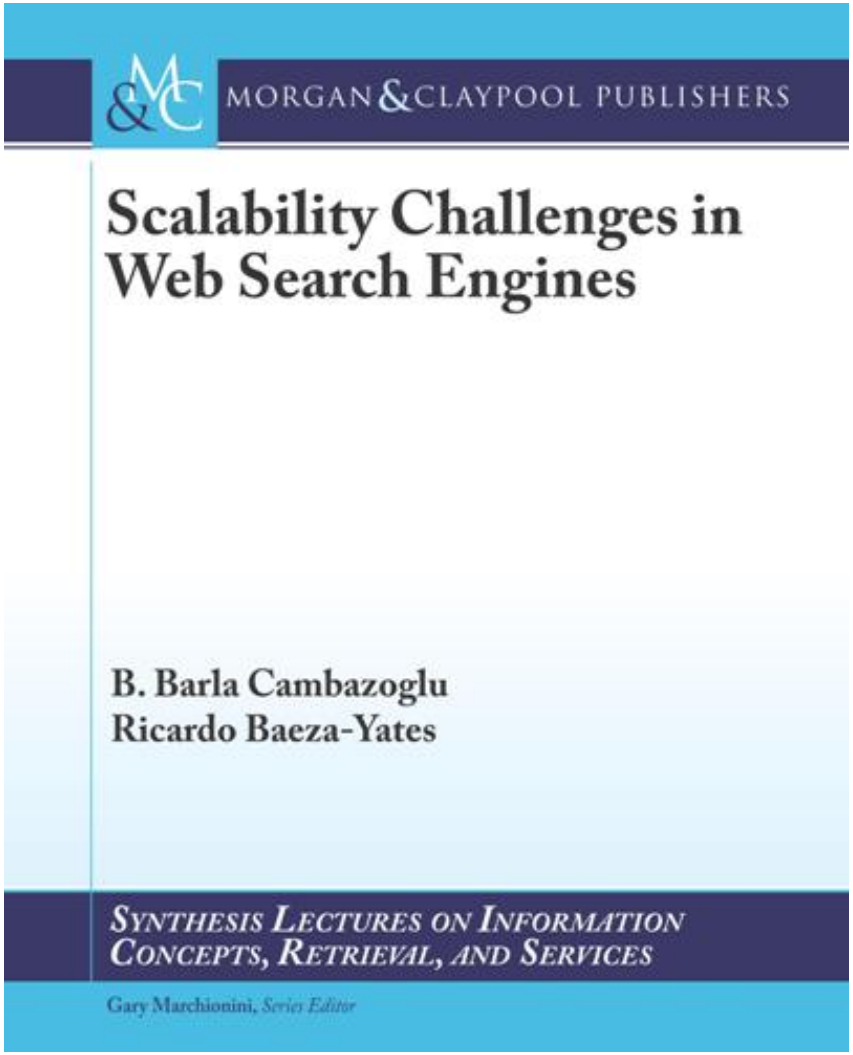


Data acquisition methods

- Downloading
- Feeds
- API
- Metadata harvesting
- **Crawling**

Outline

Scalability challenges in web search engines

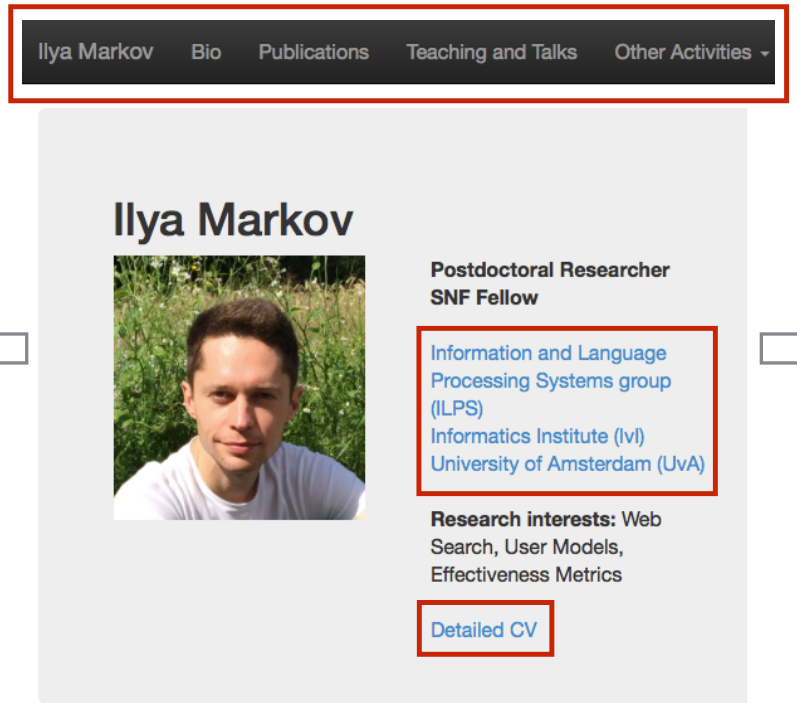


Outline

Outline

Crawling

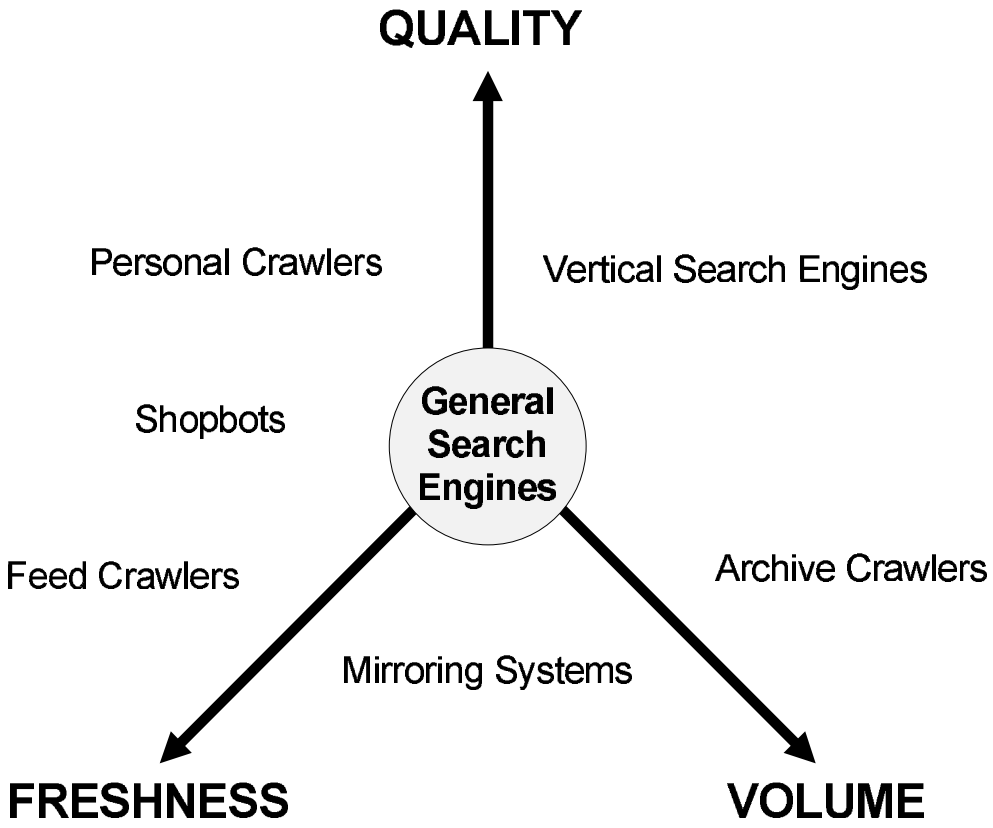
https://staff.fnwi.uva.nl/i.markov/



Content Storage

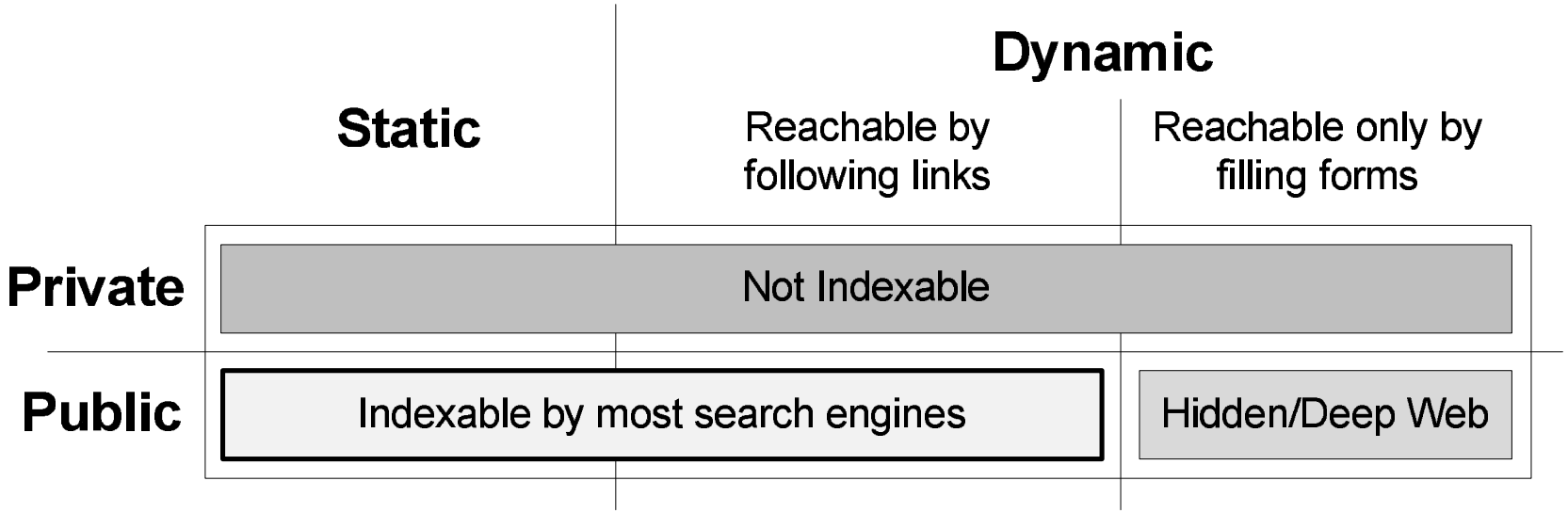
index.html
bio.html
publications.html
teaching.html
ilps.science.uva.nl
ivi.uva.nl
www.uva.nl
cv/imarkov_cv/pdf

Taxonomy of crawlers



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval"

Taxonomy of pages



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval"

Outline

How to be polite?

A Web crawler must...

1. identify itself
2. obey the robots exclusion protocol
3. keep a low bandwidth usage in a given web site

Robot identification

- Fill the `user-agent` field in the HTTP request
- Include the word "crawler", "robot", "bot", etc.

Robot exclusion protocol

- Server-wise exclusion
  `robots.txt`
- Page-wise exclusion
  `<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">`
- Cache exclusion
  `<META NAME="ROBOTS" CONTENT="NOARCHIVE">`

Bandwidth usage

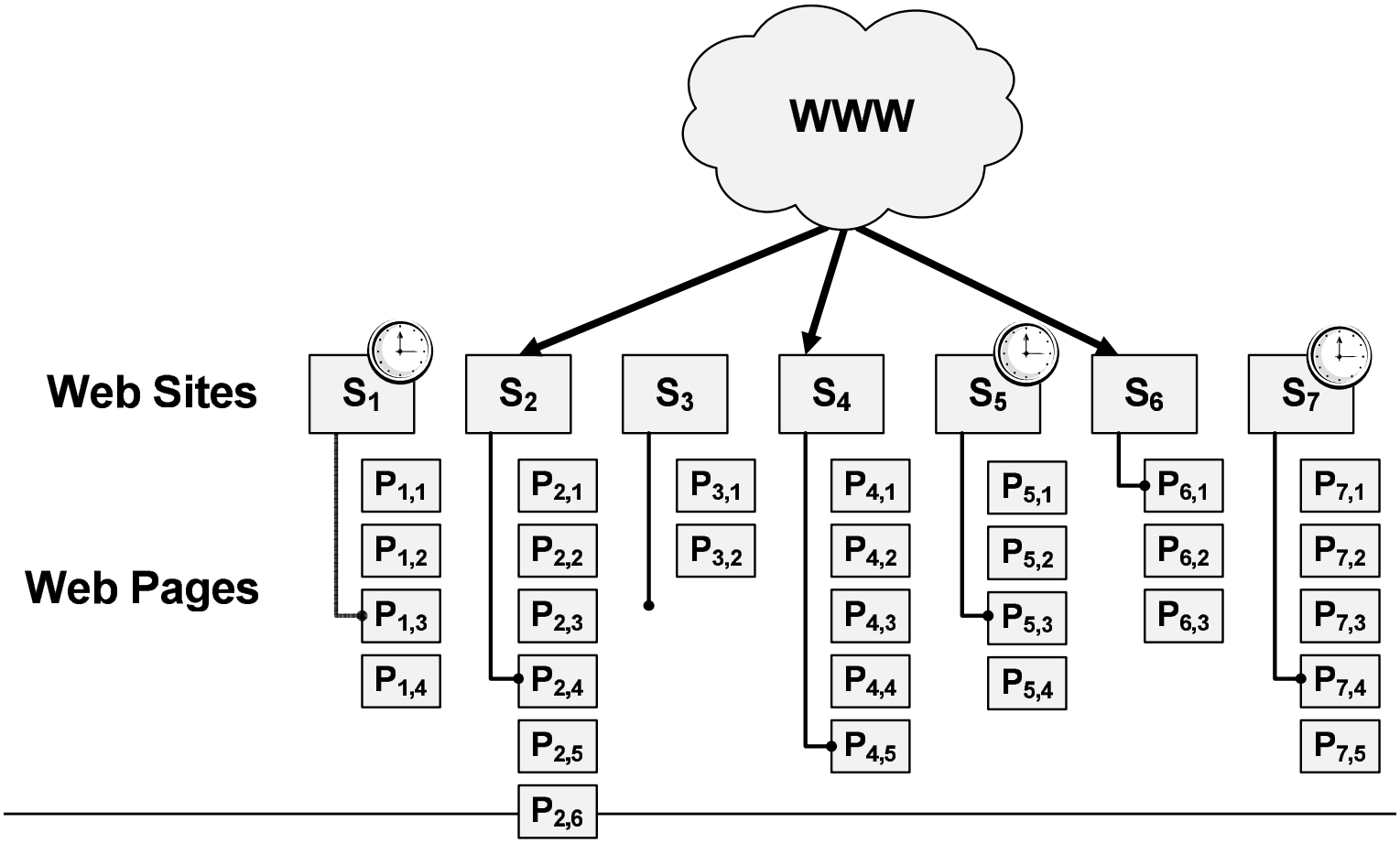
- Empirical thresholds
- Adaptive politeness policy, e.g., $10 \times t$
- `crawl-delay:45` (in seconds in `robots.txt`)

Simple crawling thread implementation

```
procedure CrawlerThread(frontier)
    while not frontier.done() do
        website ← frontier.nextSite()
        url ← website.nextURL()
        if website.permitsCrawl(url) then
            text ← retrieveURL(url)
            storeDocument(url, text)
            for each url in parse(text) do
                frontier.addURL(url)
            end for
        end if
        frontier.releaseSite(website)
    end while
end procedure
```

Croft et al., "Search Engines, Information Retrieval in Practice"

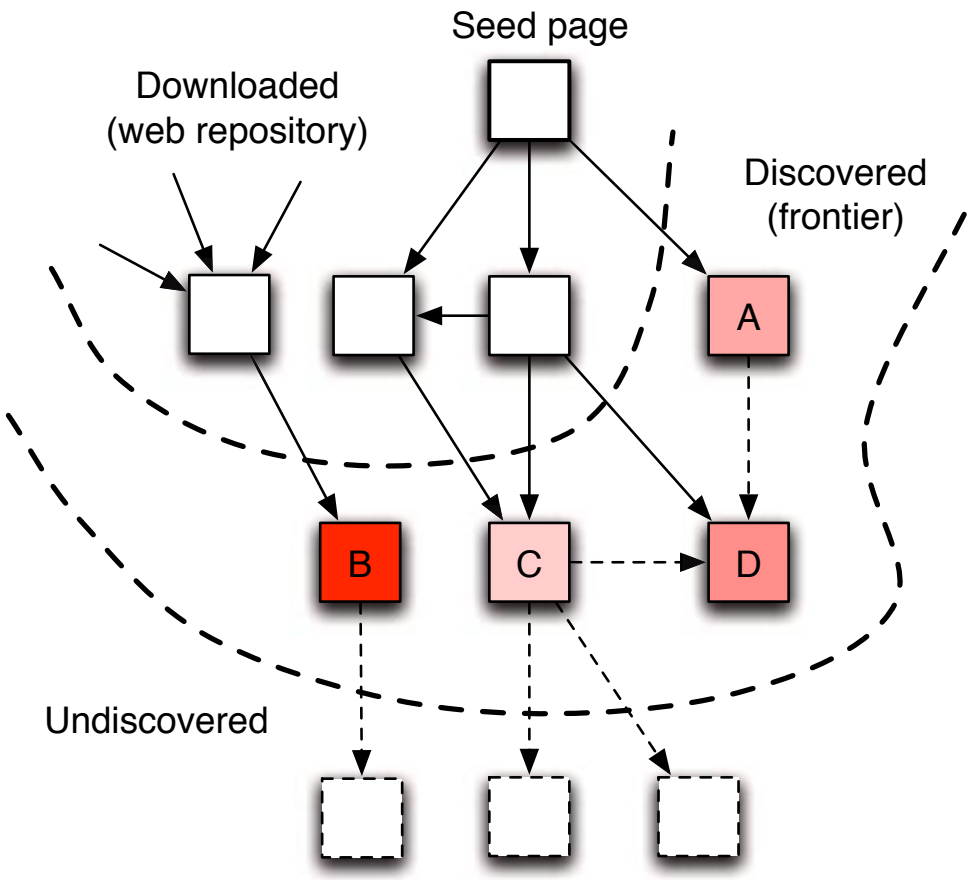
Simple crawling thread implementation



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval"

Outline

Extending the Web repository



- Random ordering
- Breadth-first
- In-degree
- Potential impact on search quality

B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

Extending the Web repository

What data structures could be used to implement the frontier?

- Random ordering
- Breadth-first
- In-degree
- Potential impact on search quality

Outline

Refreshing the Web repository

1. Identify changes in web content
2. Measure these changes
3. Predict changes
4. Select pages to update

Identifying changes

Client request:
```
HEAD /csinfo/people.html HTTP/1.1
Host: www.cs.umass.edu
```

Server response:
```
HTTP/1.1 200 OK
Date: Thu, 03 Apr 2008 05:17:54 GMT
Server: Apache/2.0.52 (CentOS)
Last-Modified: Fri, 04 Jan 2008 15:28:39 GMT
ETag: "239c33-2576-2a2837c0"
Accept-Ranges: bytes
Content-Length: 9590
Connection: close
Content-Type: text/html; charset=ISO-8859-1
```

Croft et al., "Search Engines, Information Retrieval in Practice"

Measuring changes

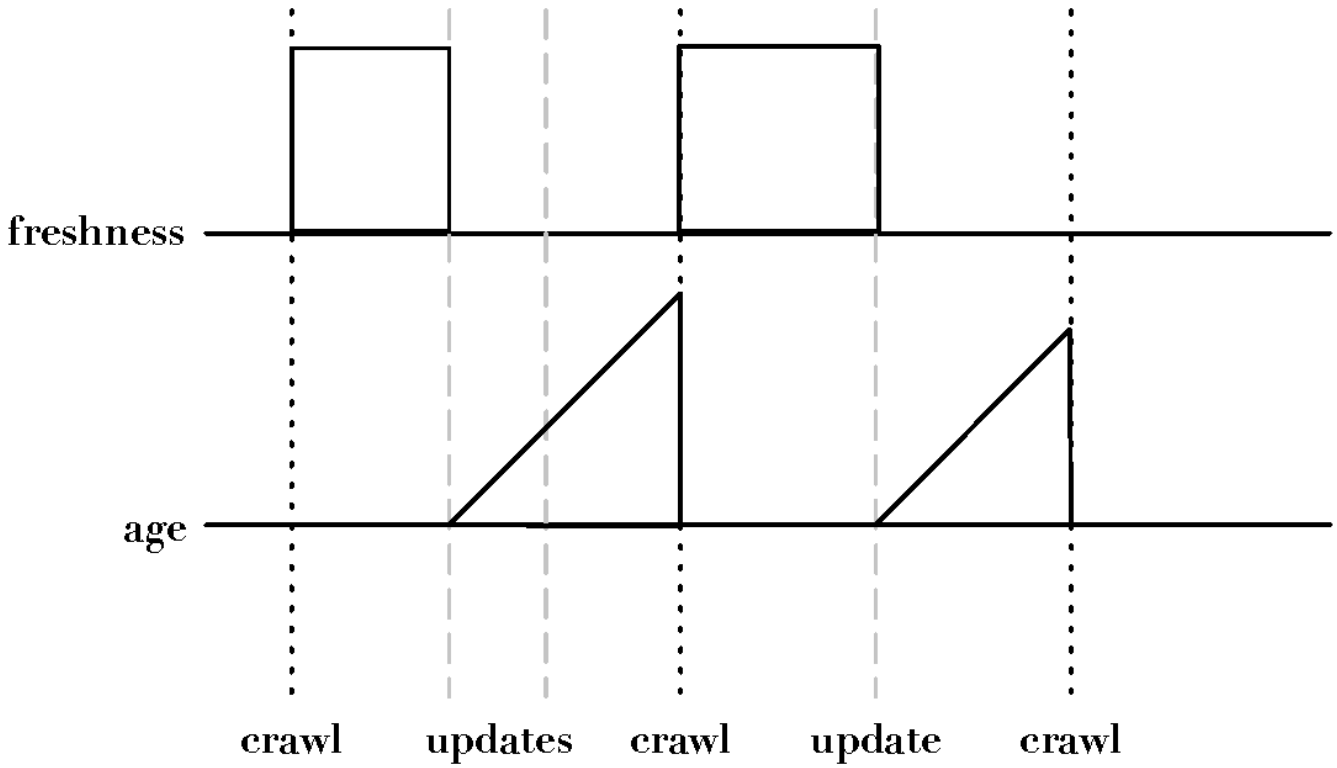
- Freshness

$$F_p(t) = \begin{cases} 1 \text{ if } p \text{ has not changed since last crawl} \\ 0 \text{ otherwise} \end{cases}$$

- Age

$$A_p(t) = \int_0^t P(p \text{ changed at time } x)(t - x)dx$$

Measuring changes



Croft et al., "Search Engines, Information Retrieval in Practice"

Predicting changes

- Age

$$A_p(t) = \int_0^t P(p \text{ changed at time } x)(t - x)dx$$

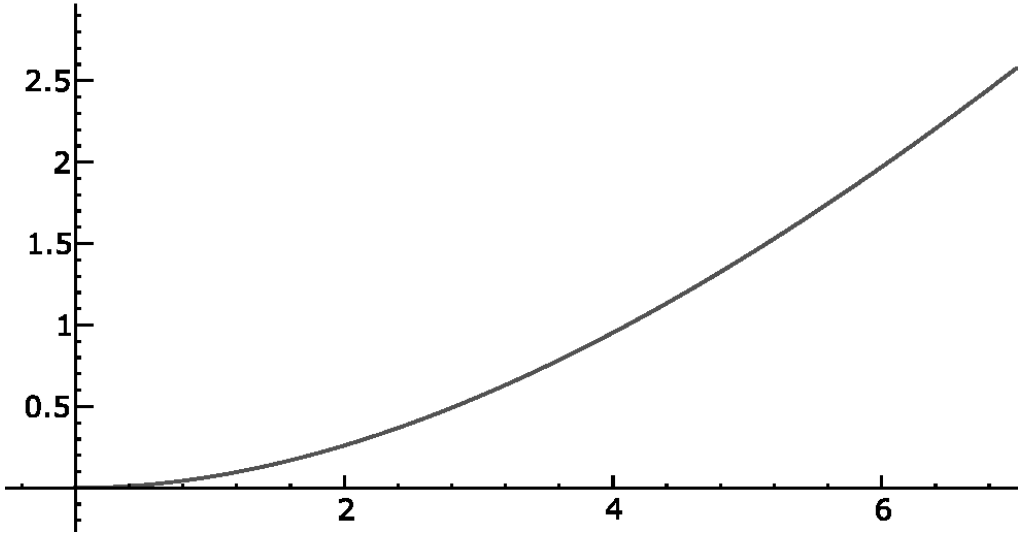
- Time between two changes follows an exponential distribution

$$A_p(t) = \int_0^t \lambda e^{-\lambda x}(t - x)dx$$

- λ – average number of changes

Example

- One update a week, $\lambda = 1/7$
- In the end of the week, the expected age is approx. 2.6



Croft et al., "Search Engines, Information Retrieval in Practice"

Selecting pages to update

- Optimizing freshness or age?
- Optimizing freshness
  - Pages with low change frequency do not need to be refreshed often
  - Pages with high change frequency are never fresh
  - Refresh pages with medium change frequency
- Optimizing age
  - The older a page gets, the more it costs not to crawl it

Selecting pages to update

	A	B	C	D
PageRank	0.0003	0.0007	0.0002	0.0001
Average daily click count	47	332	2	1974
Last download time	2 hours ago	1 day ago	8 days ago	6 hours ago
Estimated update frequency	daily	never	minutely	yearly

What data structures could be used to implement refreshing?

B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

Practical implementation

Maintain several queues

- A queue for news sites that is refreshed several times a day
- A daily or weekly queue for popular or relevant sites
- A large queue for the rest of the Web

Outline

Outline

2. Practical considerations
   - Storage and data structures
   - Distributed crawling
   - Factors affecting crawling performance
   - Deep web

Storing documents

- Support random and bulk access
- Flat file system
- A page is accessed through an identifier, e.g., a hash value of its URL

Storing document info

- Location on disk, size, history of HTTP status codes, download times, etc.
- Catalog-like data structures, e.g., B+ tree

Distributed storage

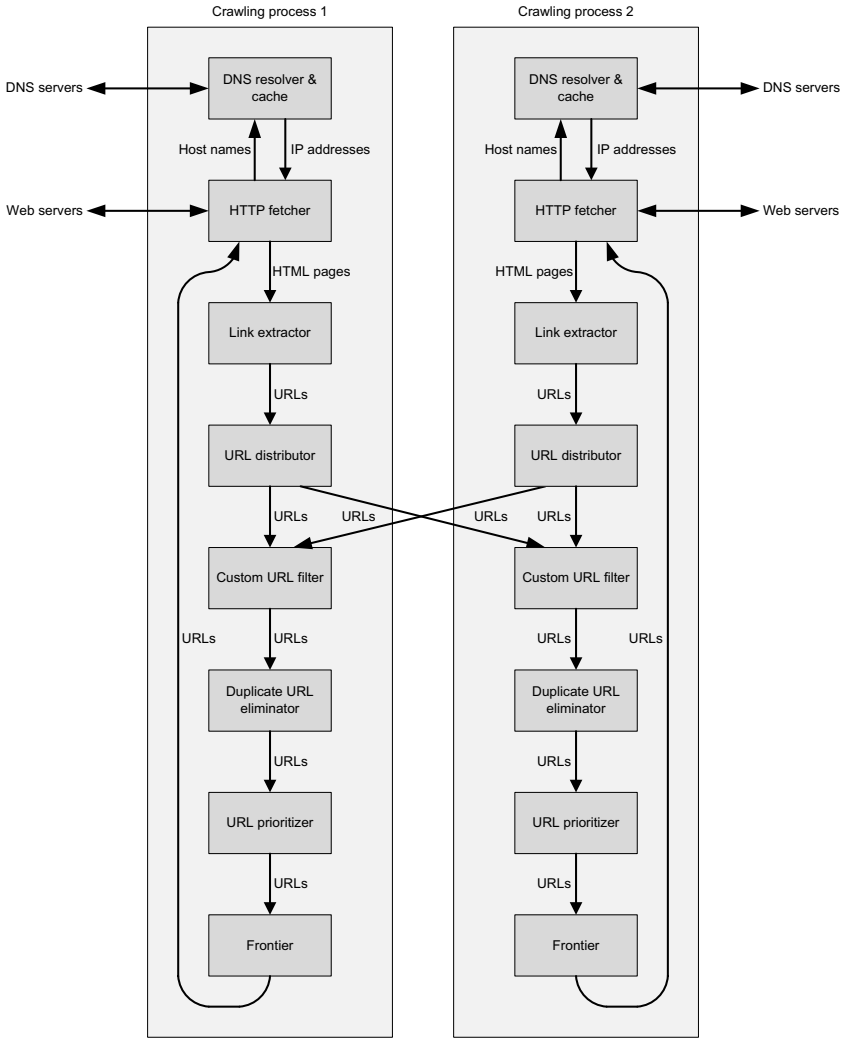
- Pages are distributed over storage nodes
- Uniformly or based on hash value ranges
- The mapping between pages and storage nodes is maintained in an index
- Distribution based on hash value ranges $\rightarrow$ smaller index
- Uniform distribution $\rightarrow$ simplified redistribution

Outline

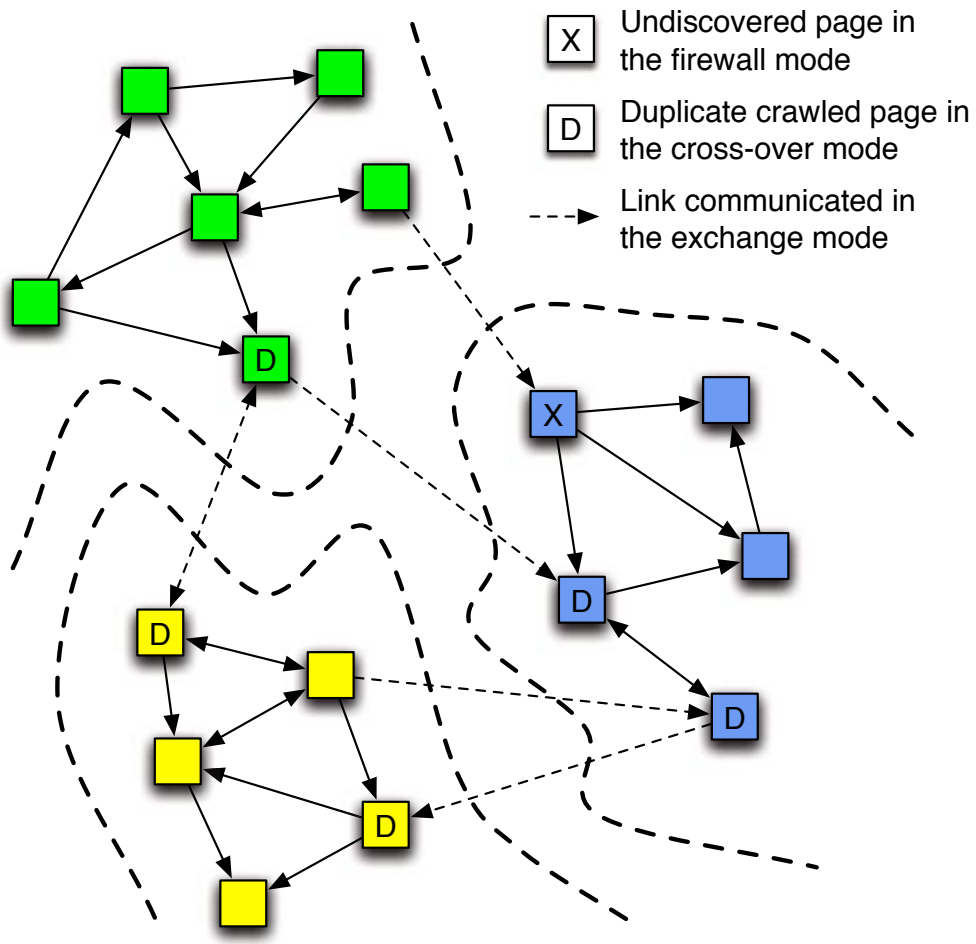
2. Practical considerations
   - Storage and data structures
   - Distributed crawling
   - Factors affecting crawling performance
   - Deep web

Distributed crawling

- Synchronize URLs between threads
- Minimize communication overhead

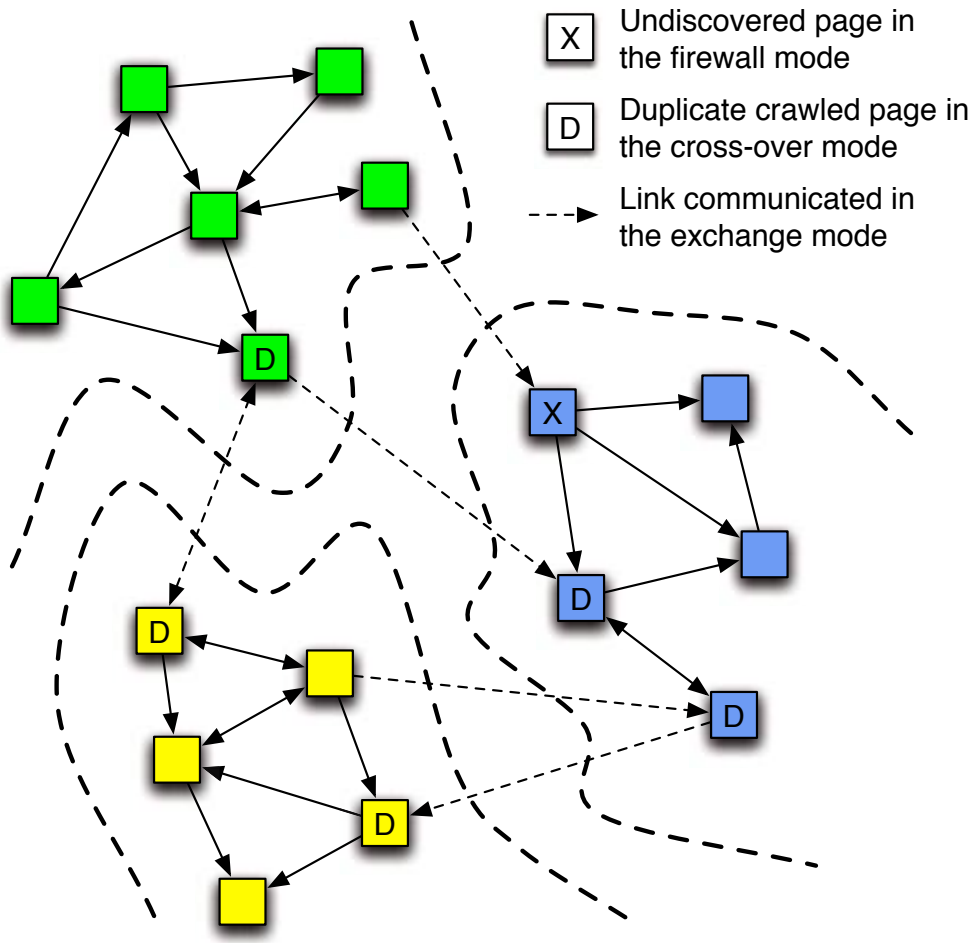


Firewall and cross-over modes



B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

Exchange mode



1. Partition pages between crawling nodes based on domains
2. Exchange only non-local links
3. Send URLs in small batches

B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

Outline

2. Practical considerations
   - Storage and data structures
   - Distributed crawling
   - Factors affecting crawling performance
   - Deep web

Factors affecting crawling performance

- Delay attack
- Spider trap
- Link farm
- Website mirroring
- Soft 404 error

Outline

Deep web

- No incoming links
- Password-protected
- Dynamic content
- Web forms

Deep web example

Google

hard drive wdc

All Images Videos News Shopping More Search tools

About 471,000 results (0.69 seconds)

Hard Drives, Network Drives, Media Players | Western Digital (WD)
www.wdc.com/
WD is a leading provider of best-selling Internal & External **Hard Drives** and award-winning Media Player & Network Drives. Discover the power of data ...

Results from wdc.com

Internal Hard Drive Overview
NAS - Mobile - WD VelociRaptor - CE/AV - Internal Drive Kits - ...

WD Support
How to format a WD external hard drive in ... Contact Support ...

WD Support / Downloads ...
Software & Downloads. WD Software Product Firmware ...

Internal Desktop Hard Drives ...
Our internal hard drives deliver the performance, capacity, speed ...

External Portable Hard Drives
Check out the best selling WD External Portable Hard Drives ...

External Storage
External Storage. For PC. Desktop Drives ... Network Attached ...

Deep web example

hard drive site:wdc.com

All Shopping Images News Videos More Search tools

About 15,000 results (0.42 seconds)

Hard Drives, Network Drives, Media Players | Western Digital (WD)
www.wdc.com/ Western Digital
WD is a leading provider of best-selling Internal & External **Hard Drives** and award-winning Media Player & Network Drives. Discover the power of data ...

Internal Hard Drive Overview - Western Digital
www.wdc.com/en/products/internal/ Western Digital
Our internal **hard drives** deliver the performance, capacity, speed and reliability that you expect from the best selling **hard drive** manufacturer in the world.

External Portable Hard Drives | Western Digital (WD)
www.wdc.com/en/products/external/portable/ Western Digital
Check out the best selling WD External Portable **Hard Drives**. Discover which type of data storage is right for you with an easy side-by-side comparison.

Internal Mobile Hard Drives Overview - Western Digital
www.wdc.com/en/products/internal/mobile/ Western Digital
Our internal **hard drives** deliver the performance, capacity, speed and reliability that you expect from the best selling **hard drive** and storage manufacturer in the ...

WD Elements - Portable Hard Drives
www.wdc.com/EN/PRODUCTS/products.aspx?id=470 Western Digital
WD Elements Portable **Hard Drives**. ... Passport Ultra Metal Edition · My Passport Ultra · WD Elements Portable · Compare All. For Mac. Desktop Drives for Mac.

WD a Western Digital brand

Blog |

External Storage Internal Hard Drives Personal Cloud Network Attached Storage

WD Home /

About 2,590 results (0.40 seconds)

Internal **Hard Drive** Overview
www.wdc.com/en/products/internal/
Our internal **hard drives** deliver the performance, capacity, speed and reliability that you expec world.

External Portable **Hard Drives** | Western Digital (WD)
www.wdc.com/en/products/external/portable/
Check out the best selling WD External Portable **Hard Drives**. Discover which type of data stor

Internal Mobile **Hard Drives** Overview
www.wdc.com/en/products/internal/mobile/
Our internal **hard drives** deliver the performance, capacity, speed and reliability that you expec manufacturer in the ...

External **Hard Drives** | Western Digital (WD)
www.wdc.com/en/products/external/desktop/
Check out the best selling WD External **Hard Drives**. Discover which type of data storage is rig

WD Gold - Datacenter **Hard Drives** | Western Digital
www.wdc.com/en/products/products.aspx?id=1670
WD Gold **hard drives** employ advanced technology to deliver among the best in reliability, cap

Outline

Duplicates

- Mirroring
- Different URLs, same content
- URL modifiers, e.g., `/dir/page.html&jsessid=09A89732`

Duplicate detection: Hashing

1. Compute a hash value of a document
2. Directly compare documents with the same hash value

Near duplicate detection: Shingling

1. Extract all unique (consecutive) sequences of n words from a document (n-grams/shingles)
2. Compute hash values for the extracted shingles
3. For two documents compute the Jaccard coefficient

$$J(d_1, d_2) = \frac{|H(d_1) \cap H(d_2)|}{|H(d_1) \cup H(d_2)|}$$

4. Documents are near-duplicates if $J(d_1, d_2) > \mathit{threshold}$
5. In practice, hashes are sorted and only the top-k hashes are considered to calculate $J(d_1, d_2)$

Near duplicate detection: Shingling

p1: A B C D E F → 79 189 44 14 99 → H1 = {14, 44, 79}

p2: A B C X D Y F → 79 189 84 68 6 87 → H2 = {6, 68, 79}

→ J(H1,H2) = 1/5

B. Cambazoglu and R. Baeza-Yates, "Scalability Challenges in Web Search Engines"

Outline

Spam

- Cloacking, redirection spam
- Link spam
- Content spam
- Link spam

Cloacking, redirection spam

Serving crawlers and users with different content

Link spam

- Link farms
  - Buy a large number of domains
  - Create a large number of sites
  - Link them to each other
- Blog spam, comment spam, wiki spam
- Hidden links

Content spam

- Keyword stuffing
  - Raise the keyword count, variety, and density
- Hidden or invisible text
  - Unrelated text using same color as the background
  - Using tiny font size
  - Hiding within HTML code
- Doorway pages
  - Designed to rank highly within search results
  - Redirect users to spam content
- Scraper sites
  - "Scrape" other sources of content
  - Create "content" for a website

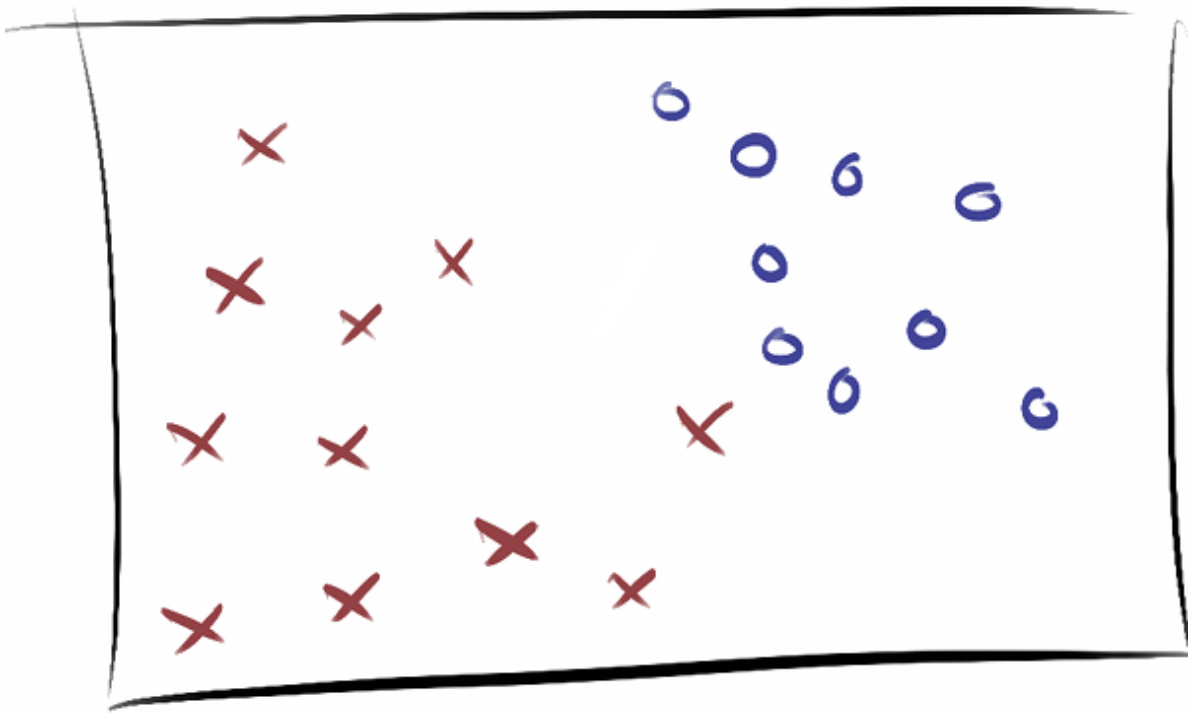
https://en.wikipedia.org/wiki/Spamdexing

Click spam

- Reasons
  - Make competitors pay for clicks
  - Compromise publishers
- Implementation
  - Many computers in different geographic locations
  - Use Trojan code on personal computers
  - Hit inflation attack

https://en.wikipedia.org/wiki/Click_fraud

Identifying spam



Picture taken from http://smerity.com/media/talks/ml_for_your_robotic_army/template.html

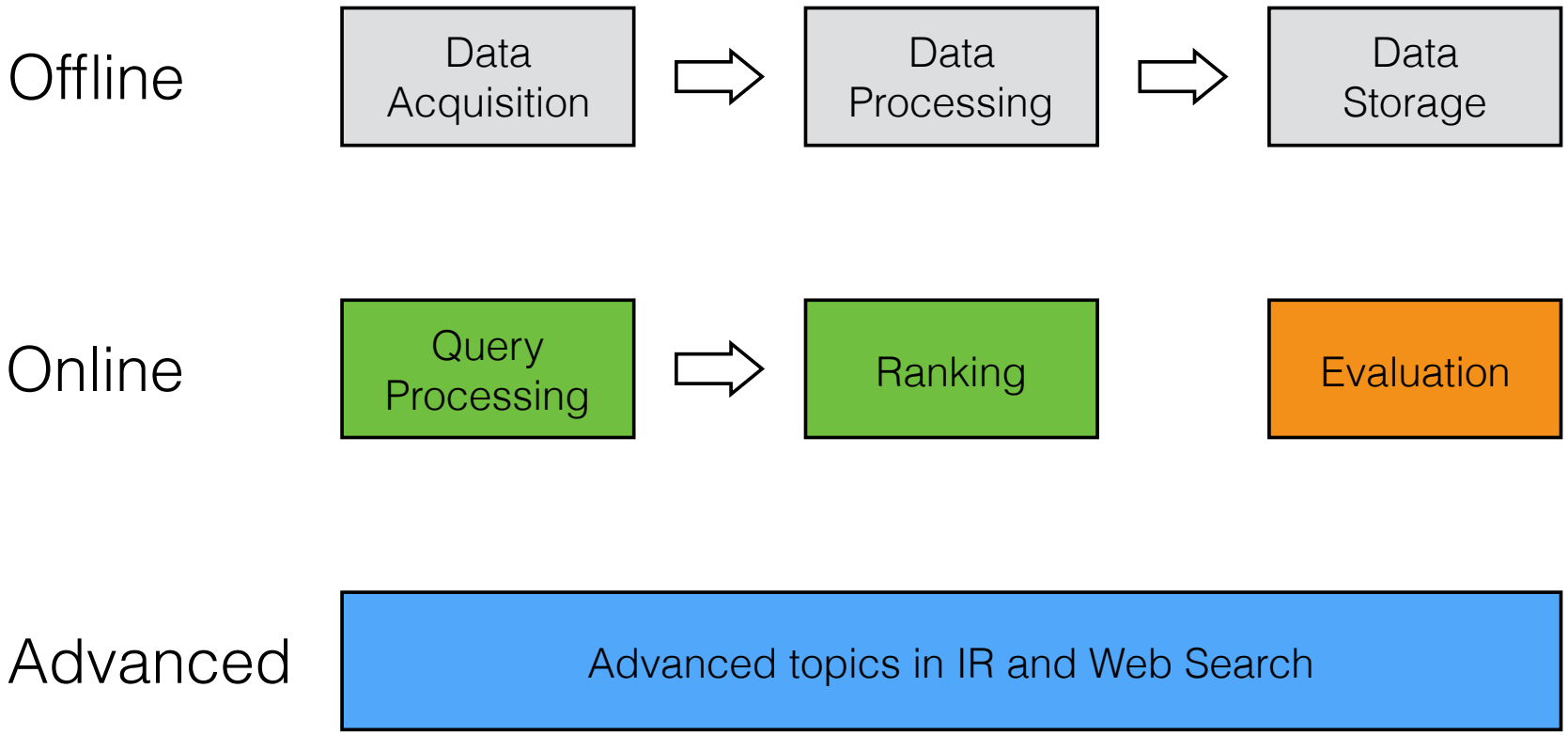
Outline

Data acquisition summary

Materials

- Croft et al., Chapters 3.1–3.4, 9.1.5
- Baeza-Yates and Ribeiro-Neto, Chapter 12
- B. Barla Cambazoglu and Ricardo Baeza-Yates
  Scalability Challenges in Web Search Engines
  Morgan & Claypool Publishers, 2017

Course overview



Next lecture