Information Retrieval

Dhruv Gupta

dhruv.gupta@ntnu.no

23-August-2022

NTNU | Norwegian University of Science and Technology

1. Course Information
   - Course Team
   - Course Prerequisites
   - Course References
   - Guest Lectures
   - Assignments
   - Examination
   - Course Contents
   - Lecture Plan

2. Introduction
   - Definition
   - Search in Libraries
   - Search in Digital Libraries
   - Searching the Web

Course Team — Instructors

- I will cover the first half of the course.
- Heri will cover the second half of the course (partly via recorded lectures).

Dhruv Gupta (*dhruv.gupta@ntnu.no*)

Heri Ramampiaro (*heri@ntnu.no*)

Course Team — Teaching Assistants

- Teaching Assistant Leader: David Baumgartner and Shiva Shadrooh.
- Teaching Assistants:
  1. Anthony Vu (*thvu@stud.ntnu.no*)
  2. Agnes Hjeltnes (*agnesgh@stud.ntnu.no*)

David Baumgartner
(*david.baumgartner@ntnu.no*)

Shiva Shadrooh
(*shiva.shadrooh@ntnu.no*)

Course Team — General

- Course ID: TDT4117 Information Retrieval
- Email to Teaching Assistants: *tdt4117-assist@idi.ntnu.no*.
- Lectures: Tuesdays (10:15-12:00 HRS).
- Location: VE1 Verkstedteknisk.
- Piazza is the main platform for discussions.
  Join here: *piazza.com/ntnu.no/fall2022/tdt4117*.
- Blackboard is the main platform for course-related information.

Course Prerequisites

- The subject material is at an advanced level.
- Recommended prerequisites:
  - IT1104 Programming Advanced Course
  - TDT4100 Object-Oriented Programming
  - TDT4145 Data Modelling, Databases and Database Management Systems

Course References

- Main course textbook:
  Baeza-Yates and Ribeiro-Neto,
  “Modern Information Retrieval",
  Second Edition.
  Pearson Education Limited, 2011.

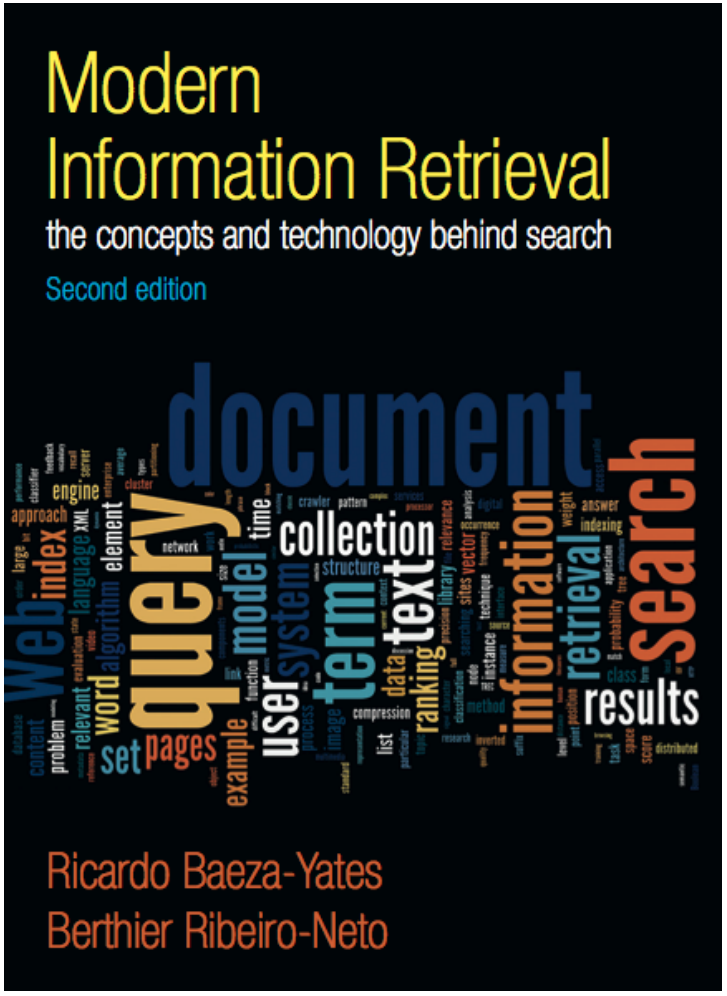


Image Credit: http://grupoweb.upf.es/mir2ed/

Course References

- Reference book: Manning et al., "Introduction to Information Retrieval", First Edition.
  Cambridge University Press, 2008.
- Book is publicly available at:
  `https://nlp.stanford.edu/IR-book/`.

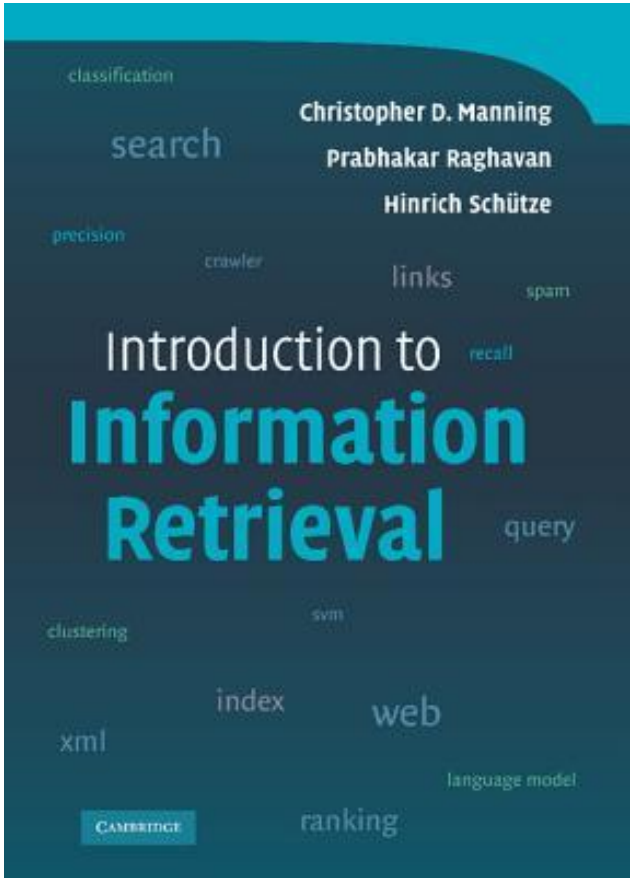


Image Credit: `https://www.goodreads.com/book/show/3278309-introduction-to-information-retrieval`

Guest Lectures

- Tentatively, there will be a guest lecture from a speaker in the Industry.
- Intention is to supplement the course knowledge from industry by daily practitioners.
- Announcement at an upcoming date.

Assignments

- 5 mandatory exercises, where 4 out of 5 must be approved.
- Submissions can be made either individually or by two-person groups.
- The exercises will be published approximately two weeks before the submission deadline.
- We expect that most of the exercises are done correctly, otherwise one an improved submission will be required.
- Plagiarism will lead to fail grade on the assignment.
- Tentative exercise plan:

Available (week)	Deadline (week)	Theme
08.09.2022 (36)	22.09.2022 (38)	Introductory Topics and Similarity Models
22.09.2022 (38)	06.10.2022 (40)	Language Models and Evaluation in IR
06.10.2022 (40)	20.10.2022 (42)	Text Operations
20.10.2022 (42)	03.11.2022 (44)	Indexing
03.11.2022 (44)	17.11.2022 (46)	Web Search

Examination

- Examination date: 07.December.2022 at 15:00.
- It will be a written exam with grades accounting for 100% of the grade.
- Old examinations: posted on Blackboard.
  There is also a solution sketch for most questions.

Course Contents[1]

- **Academic Content**
  - The course concerns automatic document storage and retrieval. In this case, the term document includes sounds and images as well as text. With this course you will learn about:
    1. File Organization,
    2. Query Operations,
    3. Document Operations,
    4. Knowledge-based Textual Information Retrieval, and
    5. Multimedia Information Retrieval.
- **Learning Objectives**
  - The students will learn and understand the principle, techniques and methods behind information retrieval.

[1] From: https://www.ntnu.edu/studies/courses/TDT4117/2022

Lecture Plan (Tentative)

Week/ Day	Topic	Lecturer	Chapter Number
34 / Tuesday	Welcome, General Information, and Introduction	Dhruv Gupta	1
35 / Tuesday	Classical Similarity Models	Dhruv Gupta	3
36 / Tuesday	Classical Similarity Models (continued), BM25, and Language Model	Dhruv Gupta	3
37 / Tuesday	Classical Similarity Models (continued), BM25, and Language Model	Dhruv Gupta	3
38 / Tuesday	Evaluation in Information Retrieval	Dhruv Gupta	4
39 / Tuesday	User Relevance Feedback and Query Expansion	Dhruv Gupta	5
-	Text Operation	Heri Ramampiaro (recorded video)	6
-	Indexing and Searching	Heri Ramampiaro (recorded video)	9
-	Guest Lectures (Tentative)	Tentative	-
-	Web Search and Search Engines	Heri Ramampiaro (recorded video)	11
-	Introduction to Multimedia Retrieval and Image Retrieval	Heri Ramampiaro (recorded video)	-
-	Audio Retrieval	Heri Ramampiaro (recorded video)	-
-	Audio Retrieval	Heri Ramampiaro (recorded video)	-

1. Course Information
   - Course Team
   - Course Prerequisites
   - Course References
   - Guest Lectures
   - Assignments
   - Examination
   - Course Contents
   - Lecture Plan

2. Introduction
   - Definition
   - Search in Libraries
   - Search in Digital Libraries
   - Searching the Web

1. Course Information
   - Course Team
   - Course Prerequisites
   - Course References
   - Guest Lectures
   - Assignments
   - Examination
   - Course Contents
   - Lecture Plan

2. Introduction
   - Definition
   - Search in Libraries
   - Search in Digital Libraries
   - Searching the Web

Information Retrieval — Definition

"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." [2]

[2] Manning et al., "Introduction to Information "Retrieval", First Edition. Cambridge University Press, 2008.

1. Course Information
   - Course Team
   - Course Prerequisites
   - Course References
   - Guest Lectures
   - Assignments
   - Examination
   - Course Contents
   - Lecture Plan

2. Introduction
   - Definition
   - Search in Libraries
   - Search in Digital Libraries
   - Searching the Web

Search in Libraries

- Information Retrieval ≡ Search.
- Search before digitization: libraries.

https://commons.wikimedia.org/wiki/File:Bookshelf_at_Yale.jpg

Search in Libraries

- Search before digitization: libraries.
- Search procedure using catalogs.
- Catalogs contained manually curated metadata about the book.

Sample Catalog Record
Author: Kesey, Ken.
Title: One flew over the cuckoo's nest, a novel.
Published: New York, Viking Press [1962]
LC Call No.: PZ4.K42On
Subjects: Psychiatric hospital patients-- United States--fiction
Control No.: 62008602

https://commons.wikimedia.org/wiki/File:Sample_Catalog_Record.png

https://en.wikipedia.org/wiki/Library_catalog#/media/File:Schlagwortkatalog.jpg

Search in Libraries

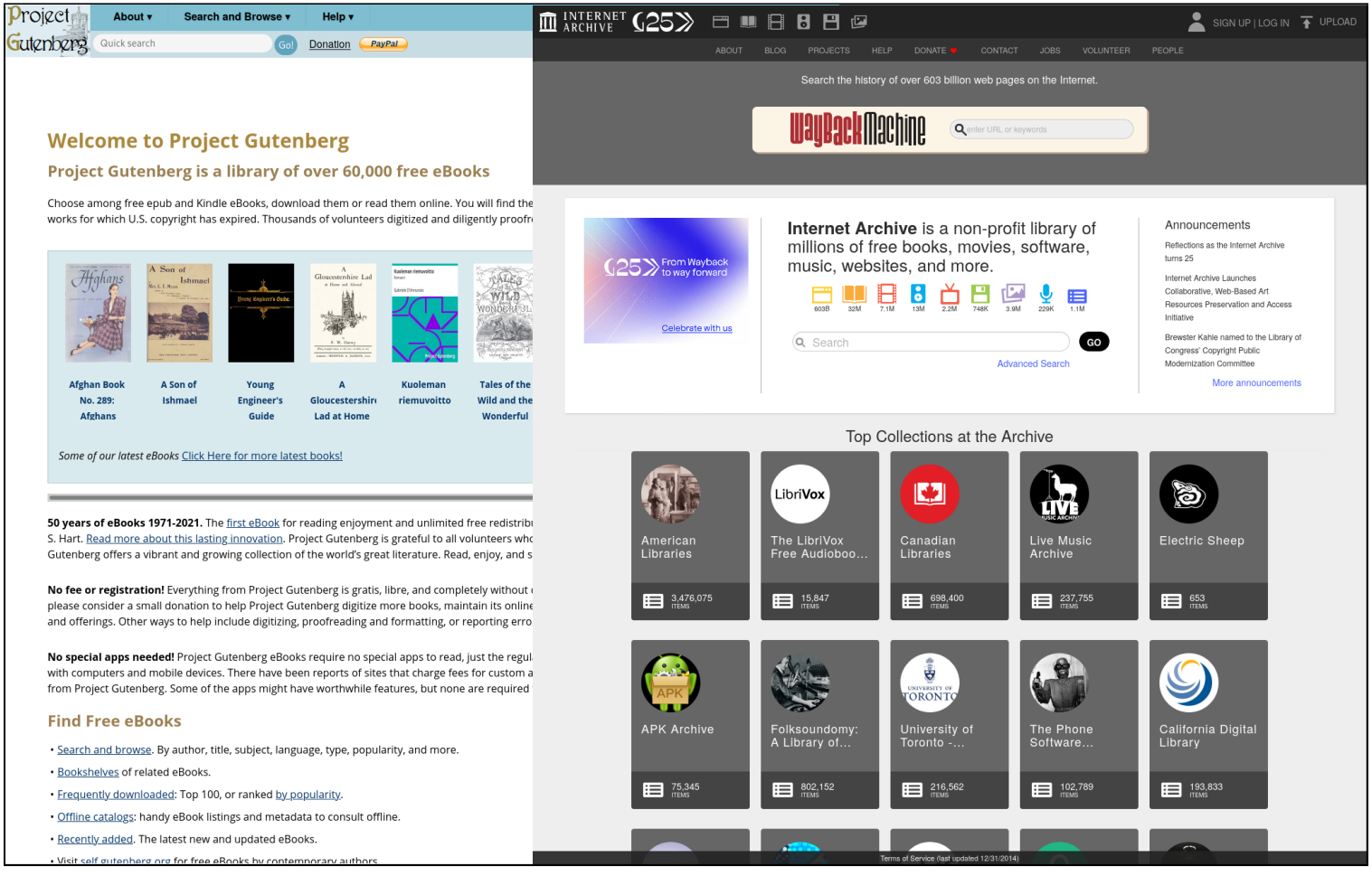
- Search before digitization: libraries.
- Search procedure using manually created catalogs.
- Location of book encoded using Dewey Decimal Classification (DDC).

Decoding Dewey Decimal Call Numbers		
Main Class	800	Literature
Division	810	American literature in English
Section	813	American fiction in English
	813.54	. . . further narrowing of topic
	813.54 M37	Cutter Number identifying author's name
	813.54 M37 2007	Edition date

https://www.ipl.org/div/farq/deweyFARQ.html

Digitizing Libraries

- Converting physical objects of knowledge to digital counterparts.
- Famous initiatives: Project Gutenberg (https://gutenberg.org/) and Internet Archive (https://archive.org/).



Search in Digitial Libraries

- Since, most catalogs are in structured format we can leverage a database solution.
- Naïve approach: map existing manual catalogs to digital database schema.

ID	Author	Book Name	ISBN	Subject
10014	Ricardo Baeza-Yates	Modern Information Retrieval	9780321416919	Computer Science

Search in Digitial Libraries

- Since, most catalogs are in structured format we can leverage a database solution.
- Naïve approach: map existing manual catalogs to digital database schema.

ID	Author	Book Name	ISBN	Subject
10014	Ricardo Baeza-Yates	Modern Information Retrieval	9780321416919	Computer Science

```
SELECT *
FROM   DIGITAL_LIBRARY
WHERE  BOOK_NAME LIKE '%Information Retreival%'
```

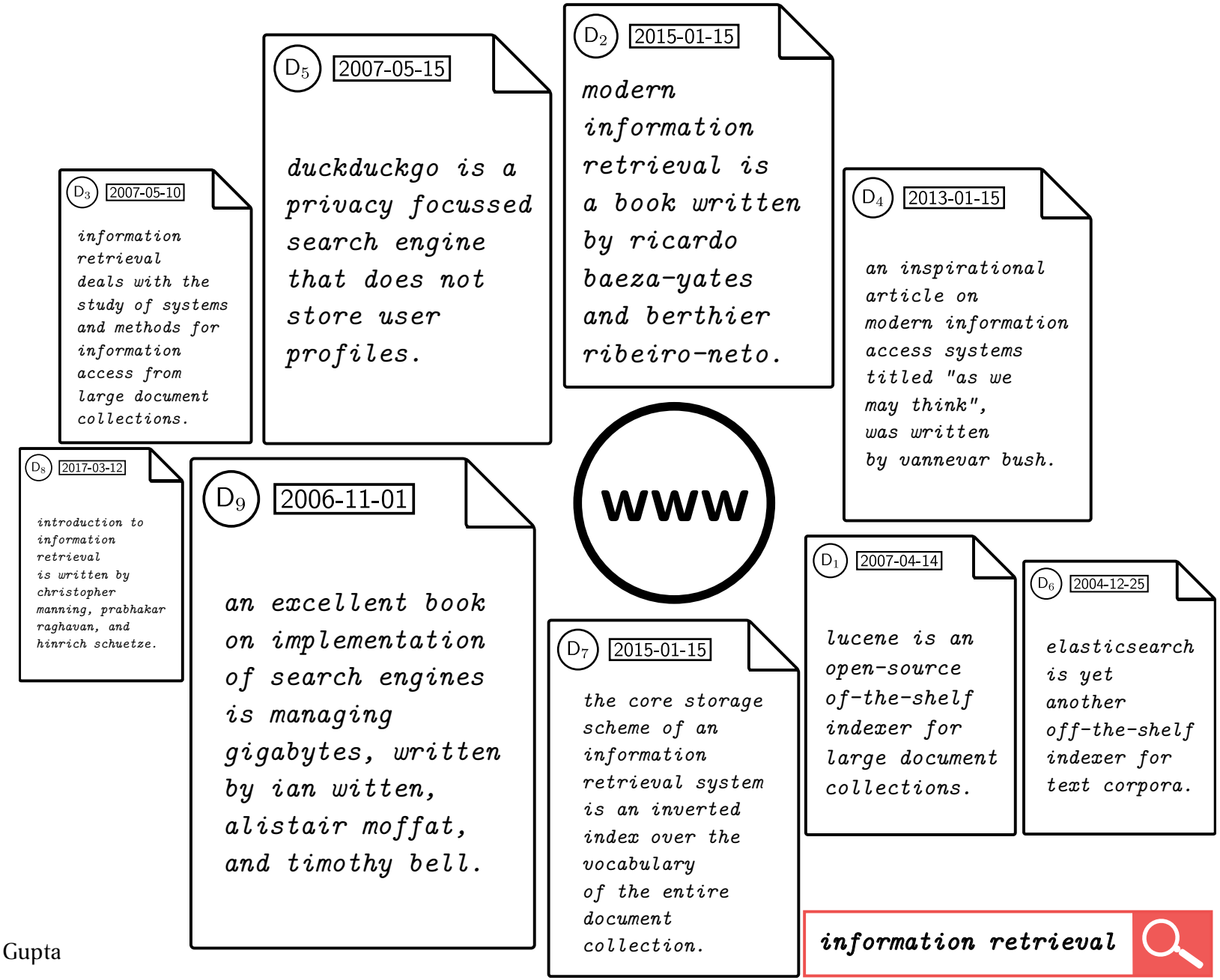
Search in Digitial Libraries

- Since, most catalogs are in structured format we can leverage a database solution.
- Naïve approach: map existing manual catalogs to digital database schema.
- Drawbacks:
  - Slow
  - Unstructured text within documents is in-accessible
  - Presentation of results
  - Determination of which is the best book

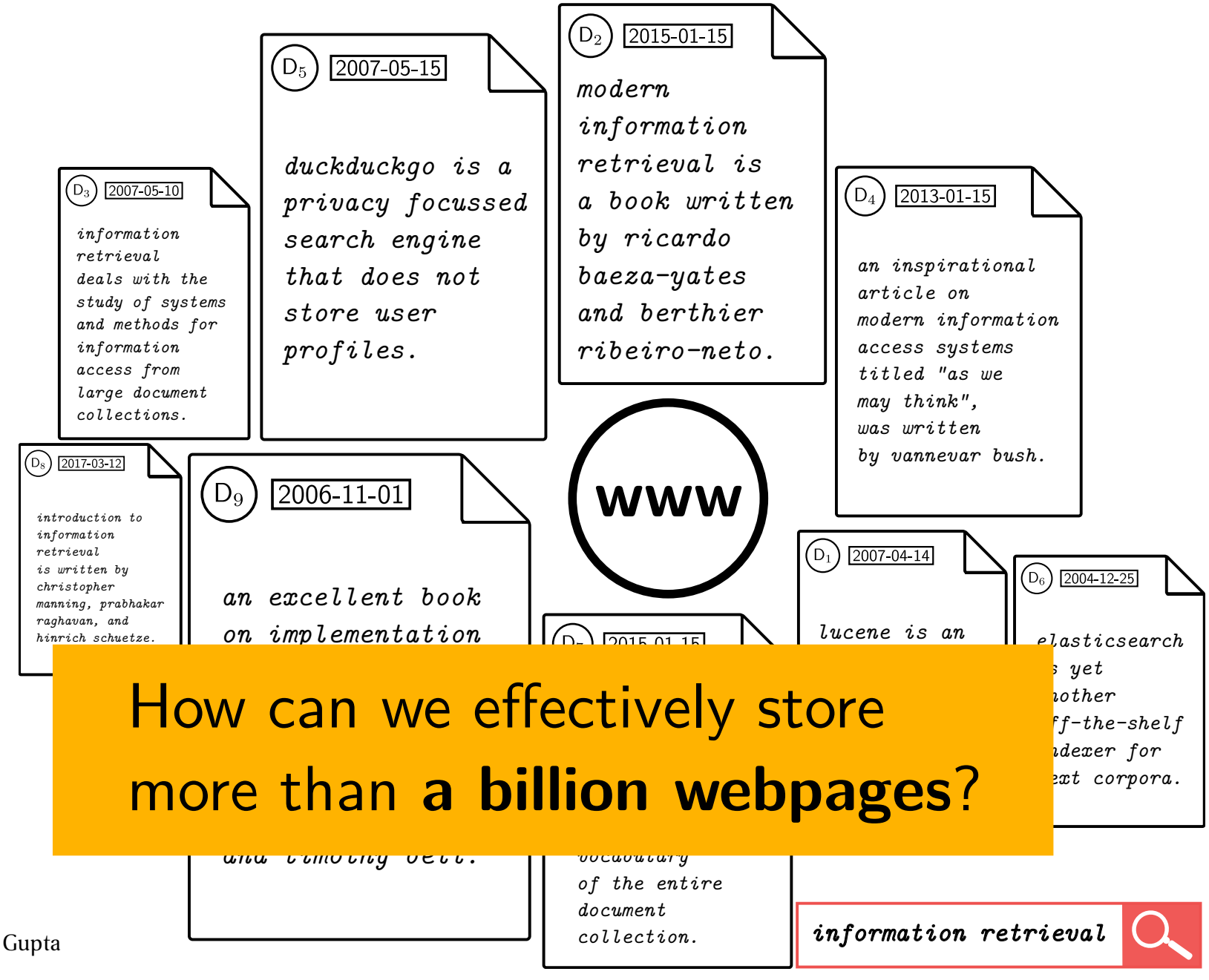
Search in Digitial Libraries

- Since, most catalogs are in structured format we can leverage a database solution.
- Naïve approach: map existing manual catalogs to digital database schema.
- Drawbacks:
  - Slow
  - Unstructured text within documents is in-accessible
  - Presentation of results
  - Determination of which is the best book
- Can not cast "Information Retrieval" as "Data Retrieval" task!

Searching the Web — Document Collection

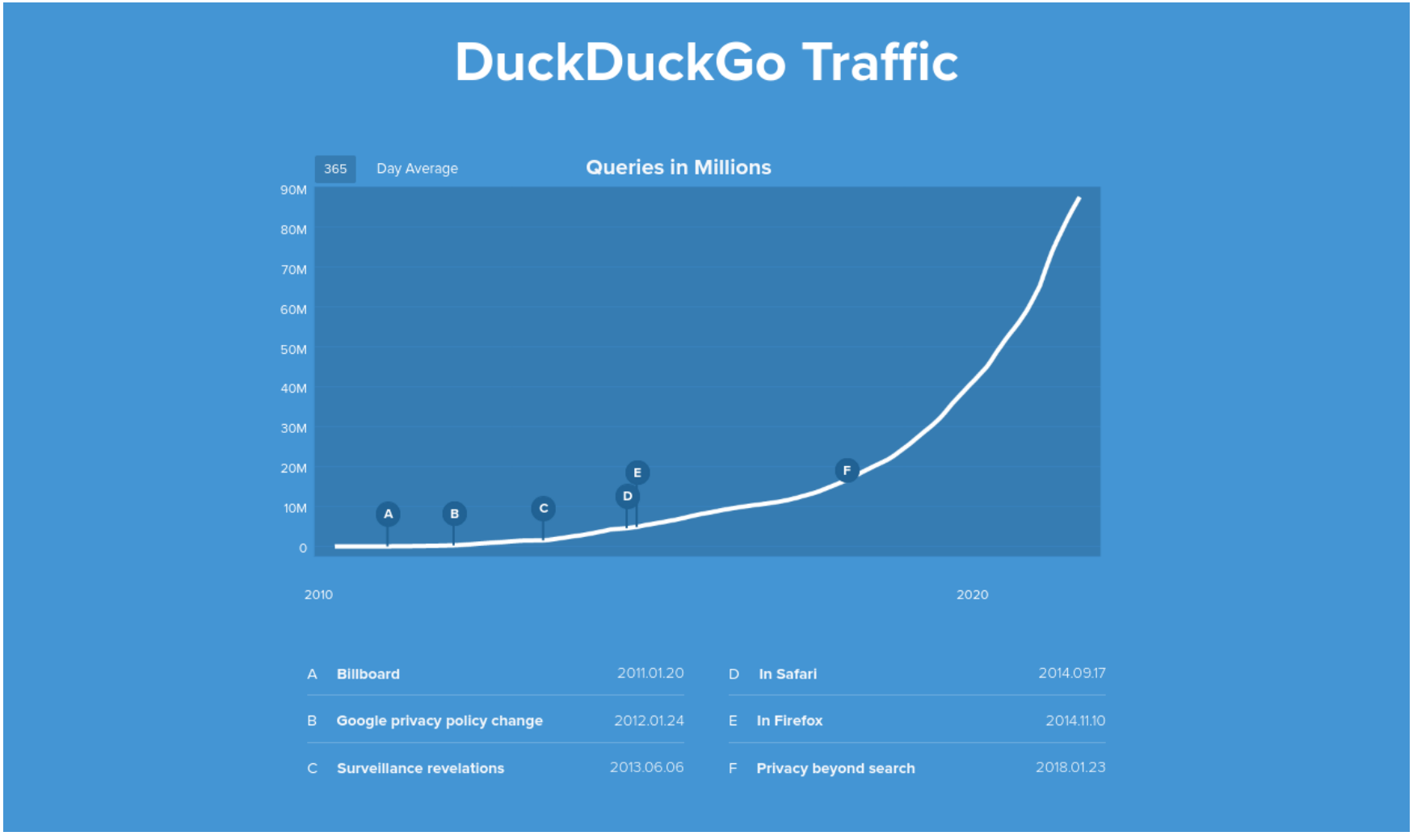


Searching the Web — Challenge of Size



How can we effectively store more than **a billion webpages**?

Searching the Web — Fault Tolerance and Latency



https://duckduckgo.com/traffic

Searching the Web — Users, Queries, and Interface

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics. Wikipedia entry of the topic web search.

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing."

- Wikipedia

`https://en.wikipedia.org/wiki/Information_retrieval`

Searching the Web — Queries and Information Needs

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection

Web Search - Wikipedia

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

Wikipedia is a public portal fo

Wikipedia entry of the topic w

Modern Information Re

http://grupoweb.upf.es/mir2

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Users express their **information needs** using natural language expressions.

Searching the Web — Queries

Information need can be expressed as:
keywords, phrases, and many more!

Searching the Web — Query Operators

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection

Web Search - Wikipedia

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

Wikipedia is a public portal fo

Wikipedia entry of the topic w

Boolean Operators further enable **conjunctive, disjunctive and negation queries**!

Modern Information Re

http://grupoweb.upf.es/mir2

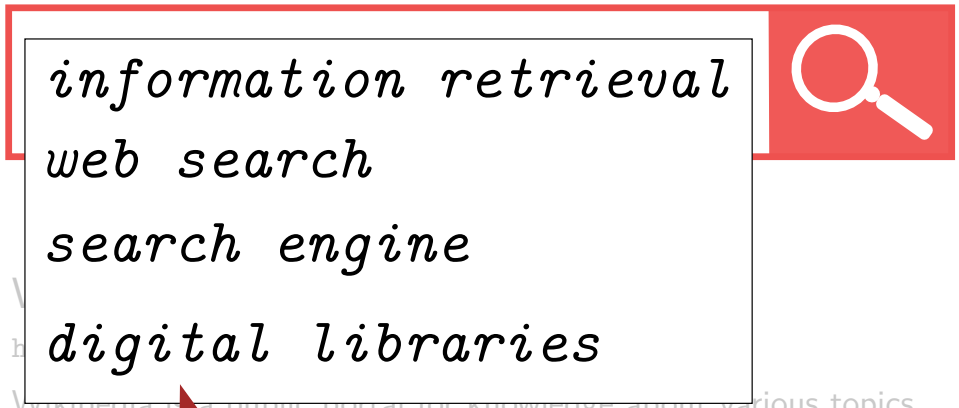
Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Searching the Web — Query Reformulations



Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing."

- Wikipedia

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

Wikipedia is a public portal for knowledge about various topics.
Wikipedia entry of the topic web search.

Modern Information Retrieval

http://grupoweb.upf.es/mir2ed/

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Query reformulations can be identified using **pseudo-relevance feedback**!

Searching the Web — Full-Text Search

information retrieval

Search results are identified by **full-text search** through document contents.

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing."

- Wikipedia

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

Searching the Web — Synonym Discovery

information retrieval

Search results are also identified by **synonym discovery!**

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Searching the Web — Relevance

information retrieval

Search results are **ranked by relevance** to the end-user!

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Searching the Web — Personalization

information retrieval

The relevance can further be "improved" based on **previous searches by the user**!

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

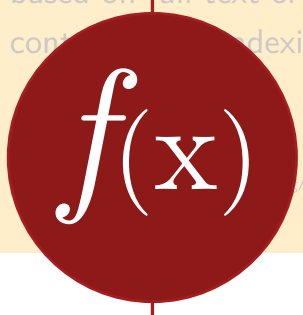
Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Searching the Web — Machine Learning

information retrieval

The relevance can further be "improved" based on previous searches by the user and **machine learning**!



Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Searching the Web — User Interfaces

User Interface:
To address the increasingly complicated information needs of users, the search interface is consantly evolving.

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing."

- Wikipedia

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

Modern Information Retrieval

http://grupoweb.upf.es/mir2ed/

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

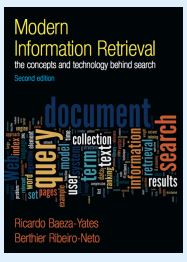
Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

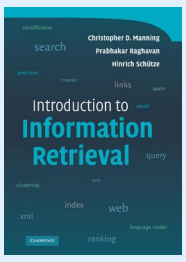
Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Searching the Web — User Interfaces

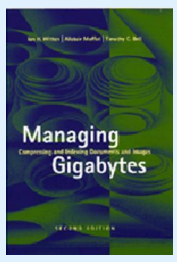
information retrieval



BUY 70 EUR
www.akademika.no



BUY 50 EUR
www.akademika.no



BUY 50 EUR
www.akademika.no

Modern Information Retrieval

http://grupoweb.upf.es/mir2ed/

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing."

- Wikipedia

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

Shopping User Interface

Alternative user interface allows users to buy related items to their queries instantly !

Image Credits:
http://grupoweb.upf.es/mir2ed/
https://www.goodreads.com/book/show/3278309-introduction-to-information-retrieval
https://www.goodreads.com/book/show/942576.Managing_Gigabytes

Searching the Web — Summary

- Information retrieval: making knowledge accessible to the user.
- Objectives:
  - Fast (millisecond response time) and fault tolerant.
  - Full-text search.
  - Ordering of documents according to relevance to the user.
  - User-centric presentation of results.