Information Retrieval and Text Mining

Introduction

Nuno Escudeiro (nfe@isep.ipp.pt)
Ricardo Almeida (ral@isep.ipp.pt)

Session outline

1. Information Retrieval
2. Information Filtering
3. Information Extraction
4. Text Mining

Learning outcomes

At the end of this session, we will be able to:

- Define Information Retrieval (IR) core concepts
- Distinguish between IR, Information Filtering (IF) and Information Extraction (IE)
- Describe the hallmarks in IR history
- Define Text Mining (TM) core concepts
- Discuss the purpose, the components and applications of IR and TM

1. Information Retrieval

A general overview of the IR field: core concepts, historic perspective, technologies, applications.

Aiming to provide a high-level view of IR and TM.

Information Retrieval: what is it?

- Information retrieval (IR) is the process of obtaining/finding relevant material/information according to the users' needs from large collections normally stored on computers.
  - It deals with the representation, storage, organization of and access to information such as:
    - Documents
    - Online catalogs
    - Structured and semi structured records
    - Multimedia objects
    - Web search

Information Retrieval: what is it?

- Nowadays we frequently think first of web search, but there are many other cases like:
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval

Information Retrieval: core tasks

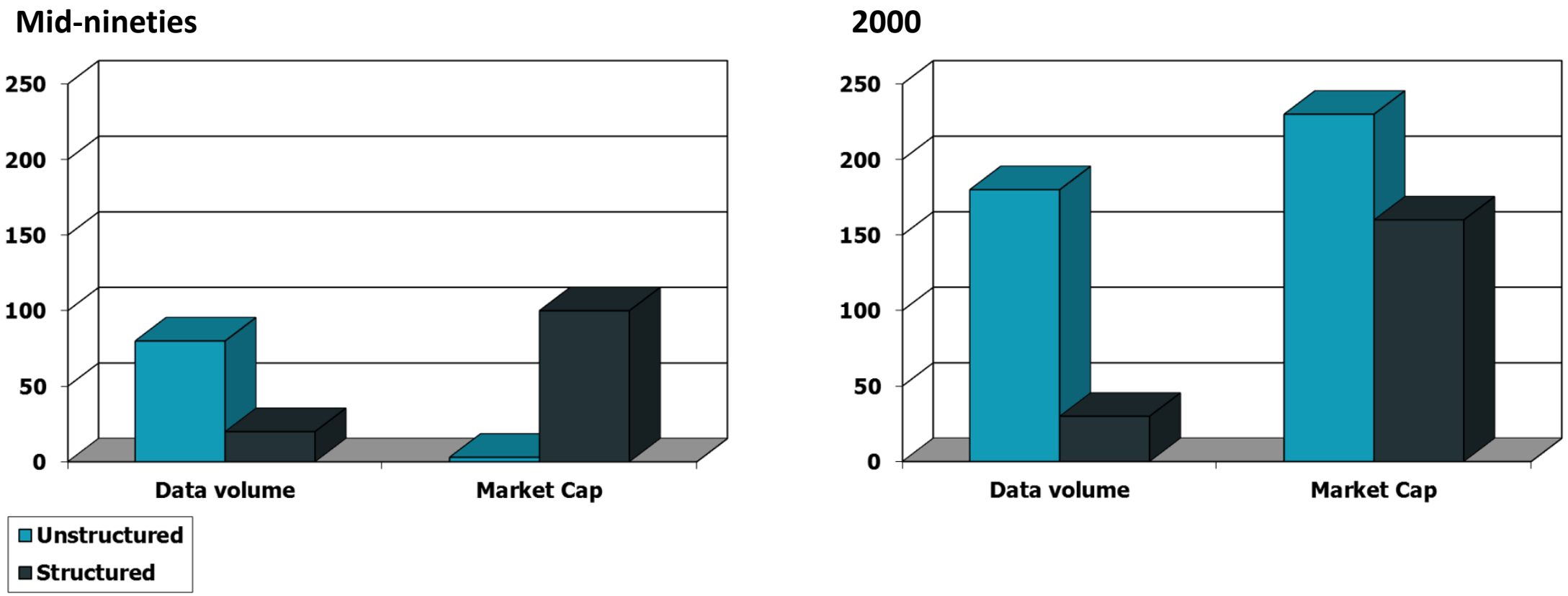
- IR normally involves:
  - **Compiling:** gathering the collection of documents that are relevant for your problem
  - **Indexing:** involves identifying and extracting important features or keywords from the documents, which are then stored in a searchable index. It is used to create an organized structure that facilitates efficient searching.
  - **Query Processing:** users submit queries – requests for information or documents relevant to their information needs.
  - **Ranking and Retrieval:** retrieved documents are ranked based on their relevance to the query.
  - **Presentation:** finally, the retrieved documents are presented to the user in a way that facilitates understanding and navigation.

Information Retrieval: timeline

- 1950-1960
  - Cranfield Experiments, conducted by Cyril W.Cleverdon – base to modern evaluation methodologies and introduce concepts like **recall and precision**
  - Gerard Salton led the SMART (System for the Mechanical Analysis and Retrieval of Text) . The project introduced the vector space model, term weighting schemes, and automatic indexing methods.
  - The Boolean model developed by George Boole, provided a formal framework for retrieving documents based on Boolean logic operators (AND, OR, NOT).
- 1970-1980
  - Salton and others proposed the VSM (**Vector Space Model**) which enabled relevance ranking based on similarity measures such as **cosine similarity** (1970).
  - Stephen E. Robertson and Karen Spärck Jones in the 1970s and 1980s, introduced a probabilistic framework for ranking documents based on the likelihood of relevance.
- 1990
  - Tim Berners-Lee created the World Wide Web early 90s. Search engines like Yahoo!, Altavista played an important role in organizing and retrieving information on the Web
  - Larry Page and Sergey Brin developed the **PageRank** algorithm
  - Tim Berners-Lee and others made the Semantic Web initiative aiming to enrich web content with machine-readable metadata, enabling more precise and intelligent information retrieval.
- 2000
  - Modern Search Engines (2000s - …) - search engines such as Google, Bing, and Baidu, which employ sophisticated algorithms and techniques for crawling, indexing, and ranking web content.
- 2010 - …
  - Deep Learning, Neural IR, Large Language Models: Advancements in deep learning and neural network models led to significant improvements in IR tasks such as: document ranking, query understanding, and relevance prediction.

Information Retrieval: timeline

Unstructured (text) vs. structured (database) data



https://slideplayer.com/slide/16362479/

Information Retrieval

- Basic assumptions for IR:
  - **Collection, corpus** : set of text documents that is assumed to be static while processing a query. Plural: **corpora**
  - **Goal:** Retrieve documents that are relevant to the user's information need and helps the user complete a task.

Information Retrieval

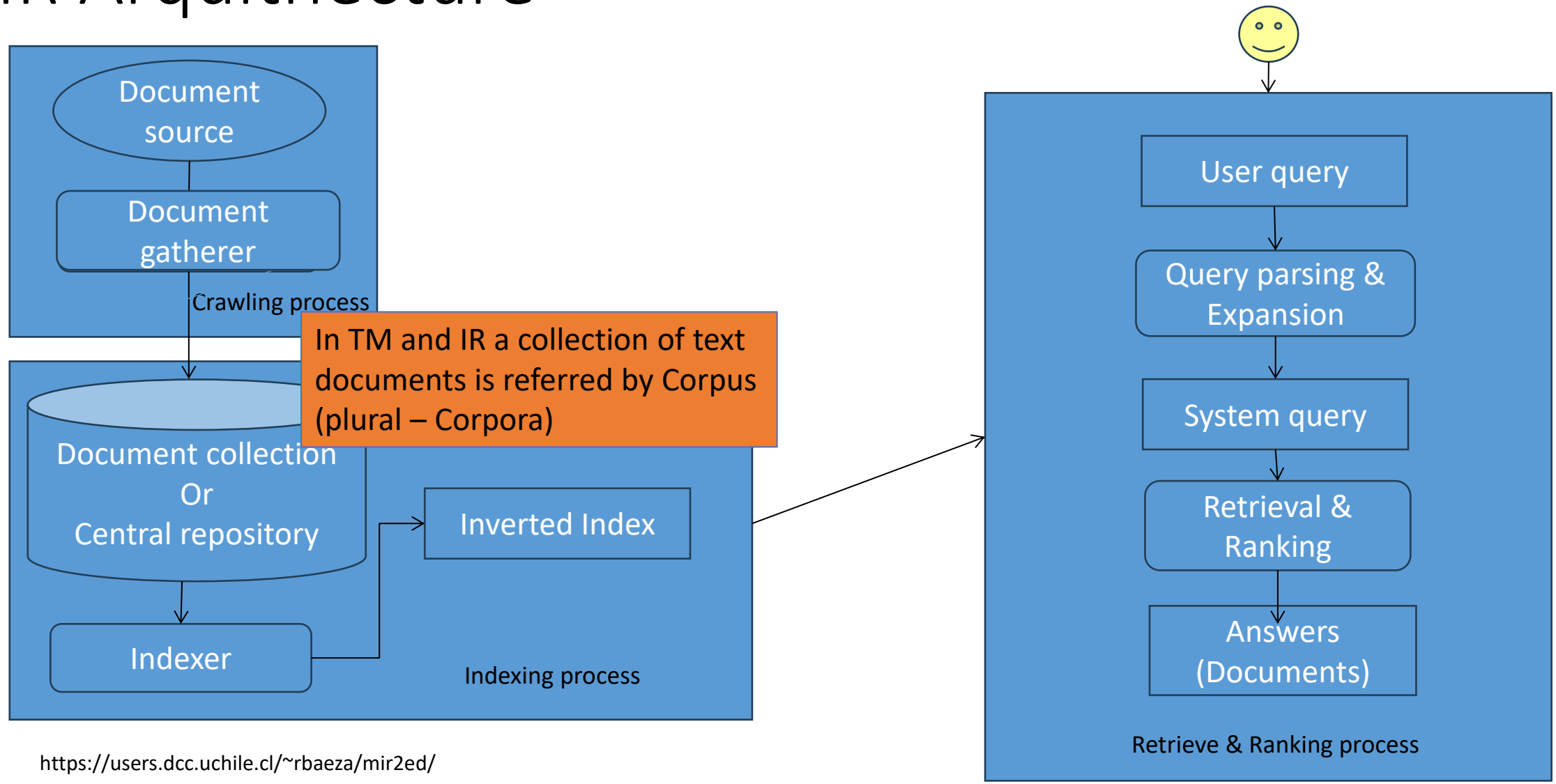
- Relevance of the retrieved documents:
  - **Precision :** number of retrieved docs, from all that were obtained, that are relevant to the user's information need

$$\textbf{Precision} = \frac{Number\ of\ Relevant\ Documents\ Retrieved}{Total\ Number\ of\ Documents\ Retrieved}$$

  - **Recall :** number of relevant docs in collection that are retrieved

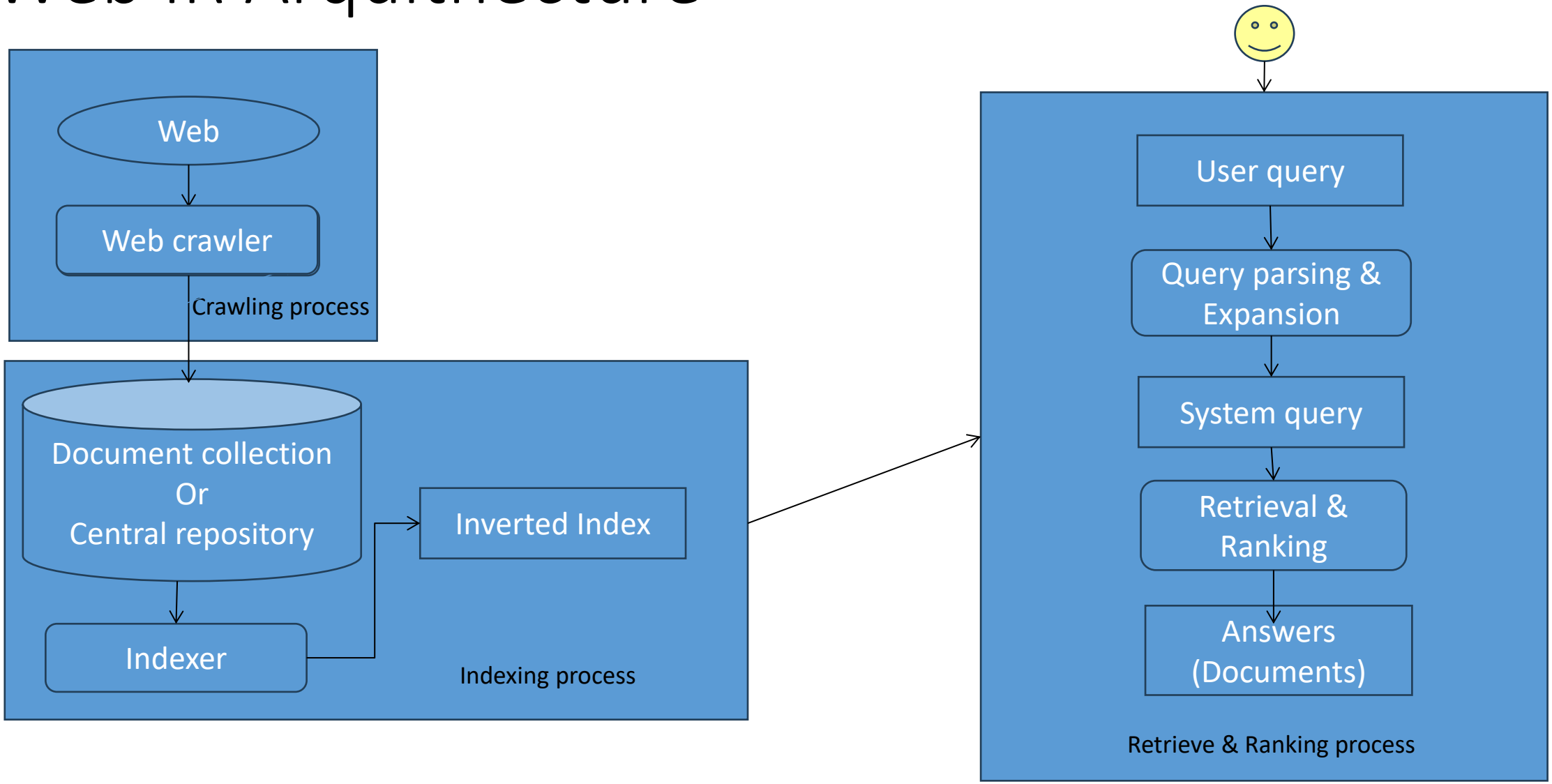
$$\textbf{Recall} = \frac{Number\ of\ Relevant\ Documents\ Retrieved}{Total\ Number\ of\ Relevant\ Documents\ in\ the\ Corpus}$$

IR Arquithecture

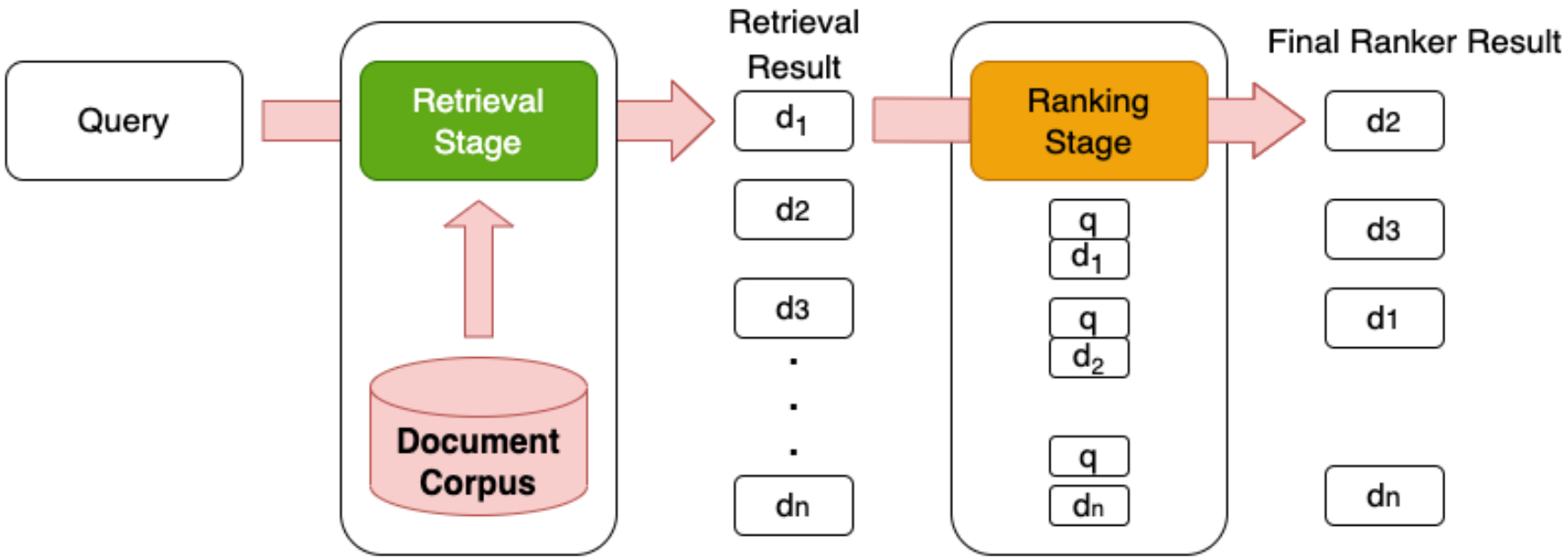


https://users.dcc.uchile.cl/~rbaeza/mir2ed/

Web IR Arquithecture



IR Arquithecture

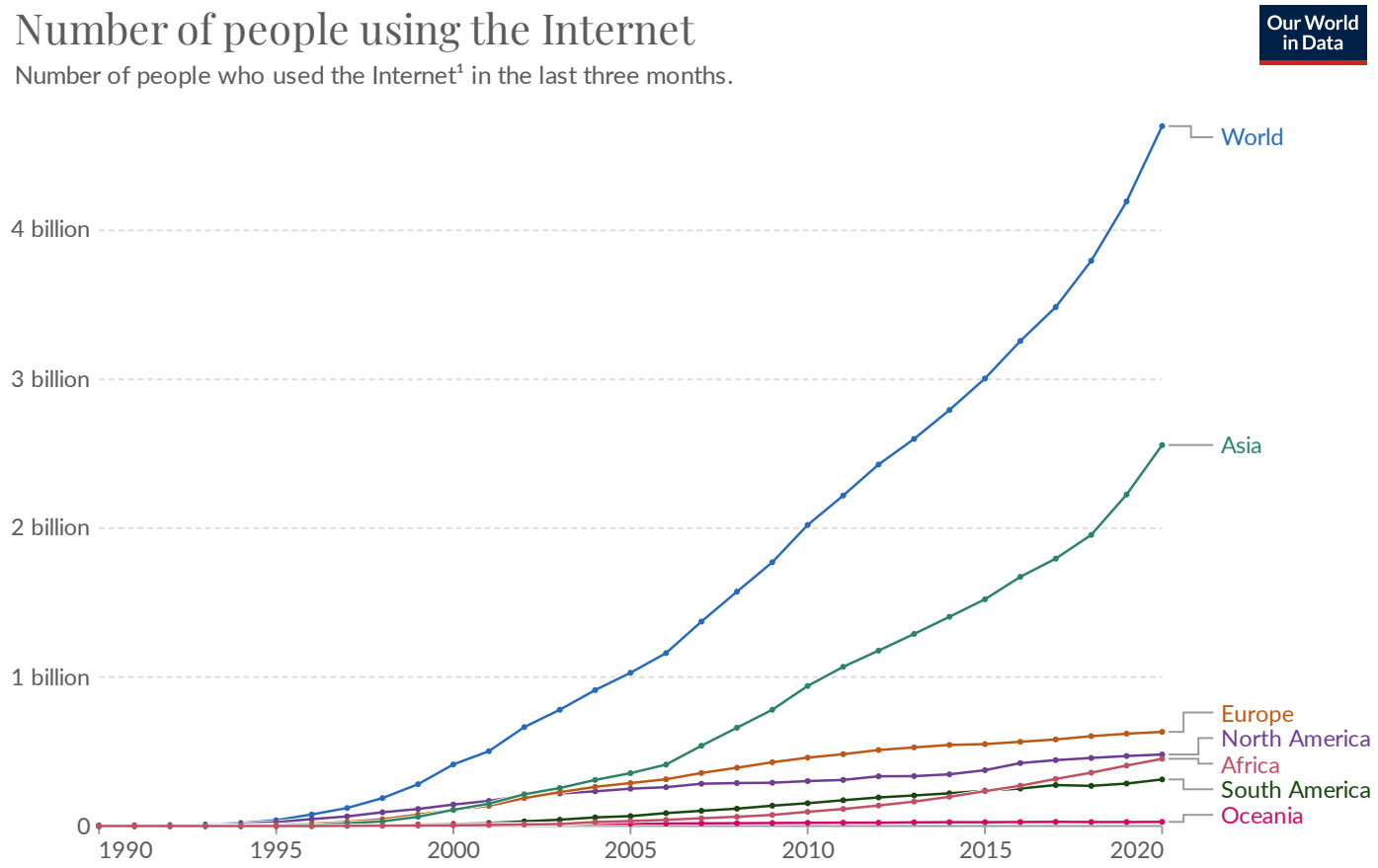


Hambarde, K.; Proença, H. Information Retrieval: Recent Advances and Beyond. Journal Not Specified **2021**, 1, 0. https://doi.org/
Academic Editor: Amelia Carolina Sparavigna

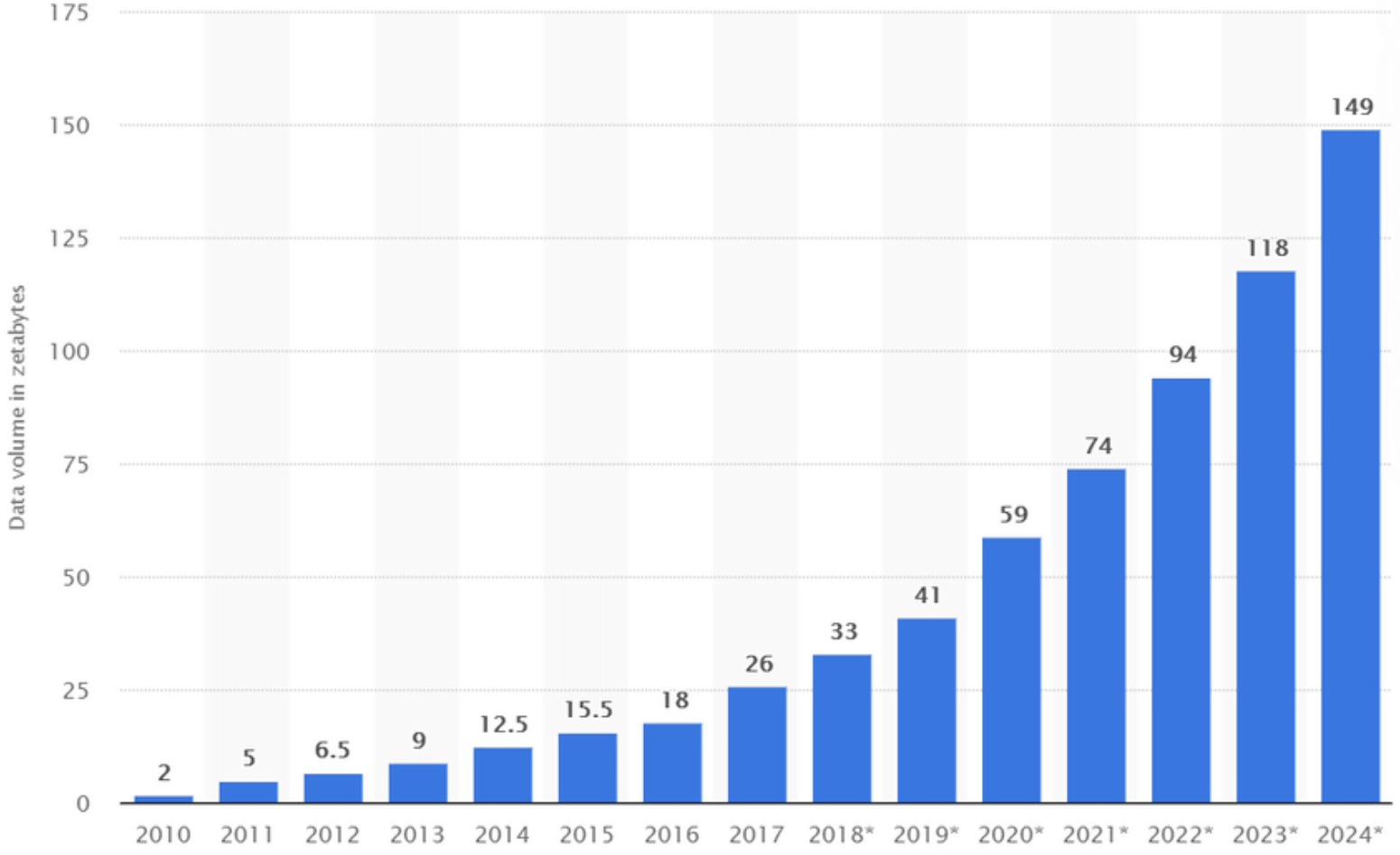
Information Retrieval

- Importance of IR:
  - Facilitate access to vast amount of information
  - Essential for decision making, research and problem solving
  - Enhances efficiency in finding relevant information on time

Information Retrieval



Information Retrieval



Zettabyte = 10^{21} bytes
Exabyte = 10^{18} bytes
Petabyte = 10^{15} bytes
Terabyte = 10^{12} bytes
Gigabyte = 10^{9} bytes
Megabyte = 10^{6} bytes
Kilobyte = 10^{3} bytes

Berisha, Blend & Mëziu, Endrit. (2021). Big Data Analytics in Cloud Computing: An overview. 10.13140/RG.2.2.26606.95048.

Statista, ”Statista,” 2020. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/.

Information Retrieval

- Challenges in IR:
  - Information Overload
  - Ambiguity in Queries
  - Relevance Judgment
  - Dynamic Information Sources

Information Retrieval

- Trends in IR:
  - Semantic Search
  - Personalized Retrieval
  - Multimodal Information Retrieval
  - Integration of AI and Machine Learning
  - Generative AI

Information Filtering

A general overview of Information Filtering: core concepts.

Aiming to provide a high-level view of Information Filtering.

Information Filtering

- Information Filtering (IF) is the process of selecting and delivering relevant information to users based on predefined criteria.
  - They are applicable for unstructured or semi-structured data (e.g. documents, e-mail messages)
  - Handle large amounts of data
  - Deal primarily with textual data
  - Based on user profiles
  - Their objective is to remove irrelevant data from incoming streams of data items

Information Filtering

- Types of information filtering:
  - Collaborative filtering
  - Content-based filtering
  - Hybrid filtering

Information Filtering

- Techniques in Information Filtering :
  - Collaborative filtering
  - Content-based filtering
  - Hybrid filtering

Information Filtering

- Applications in:
  - E-commerce
  - Social Media
  - Entertainment

Information Extraction

A general overview of Information Extraction: core concepts.

Aiming to provide a high-level view of Information Extraction.

Information Extraction

- Information Extraction (IE) is the process of identifying and extracting structured information (entities, relationships between entities and attributes) from unstructured or semi-structured data.

Information Extraction

- Techniques in IE:
  - Named Entity Recognition (NER)
  - Relation Extraction
  - Text Classification
  - Natural Language Processing (NLP)
  - etc.

Information Extraction

- Applications of IE:
  - Information Retrieval
  - Natural language processing
  - Data Integration

Text Mining

A general overview of Information Text Miiing: core concepts.

Aiming to provide a high-level view of Text Mining.

Text Mining

- Text mining or text analysis is the process of deriving useful insights, patterns, and knowledge from unstructured text data.

Text Mining

- Key topics:
  - Text Preprocessing
  - Text Representation/Modelling
  - Text Classification
  - Text Clustering

Text Mining

- Applications:
  - Sentiment Analysis
  - Information Extraction
  - Topic Modeling

Text Mining

- Document Representation Techniques:
  - Bag-of-Words Model
  - TF-IDF (Term Frequency-Inverse Document Frequency)
  - Word Embeddings: Word2Vec, GloVe, FastText, …

Text Mining

- Query Processing Methods :
  - Boolean Retrieval Model
  - Vector Space Model
  - Term Weighting and Similarity Measures

References

- https://users.dcc.uchile.cl/~rbaeza/mir2ed/

- https://www.google.pt/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiJwaWdpZyEAxWwVKQEHQhJAQ0QFnoECBoQAQ&url=https%3A%2F%2Fweb.stanford.edu%2Fclass%2Fcs276%2Fhandouts%2Flecture1-intro.ppt&usg=AOvVaw1COq6jqt5u77dUTbyVnZKJ&opi=89978449