Information Extraction and Information Retrieval

Information Retrieval

- Information Retrieval can be defined as a software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories, particularly textual information.

- Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature i.e. usually text which satisfies an information need from within large collections which is stored on computers.

- The IR system assists the users in finding the information they require but it does not explicitly return the answers to the question.

- Information retrieval also extends support to users in browsing or filtering document collection or processing a set of retrieved documents. The system searches over billions of documents stored on millions of computers.

IR Model /Components of IR

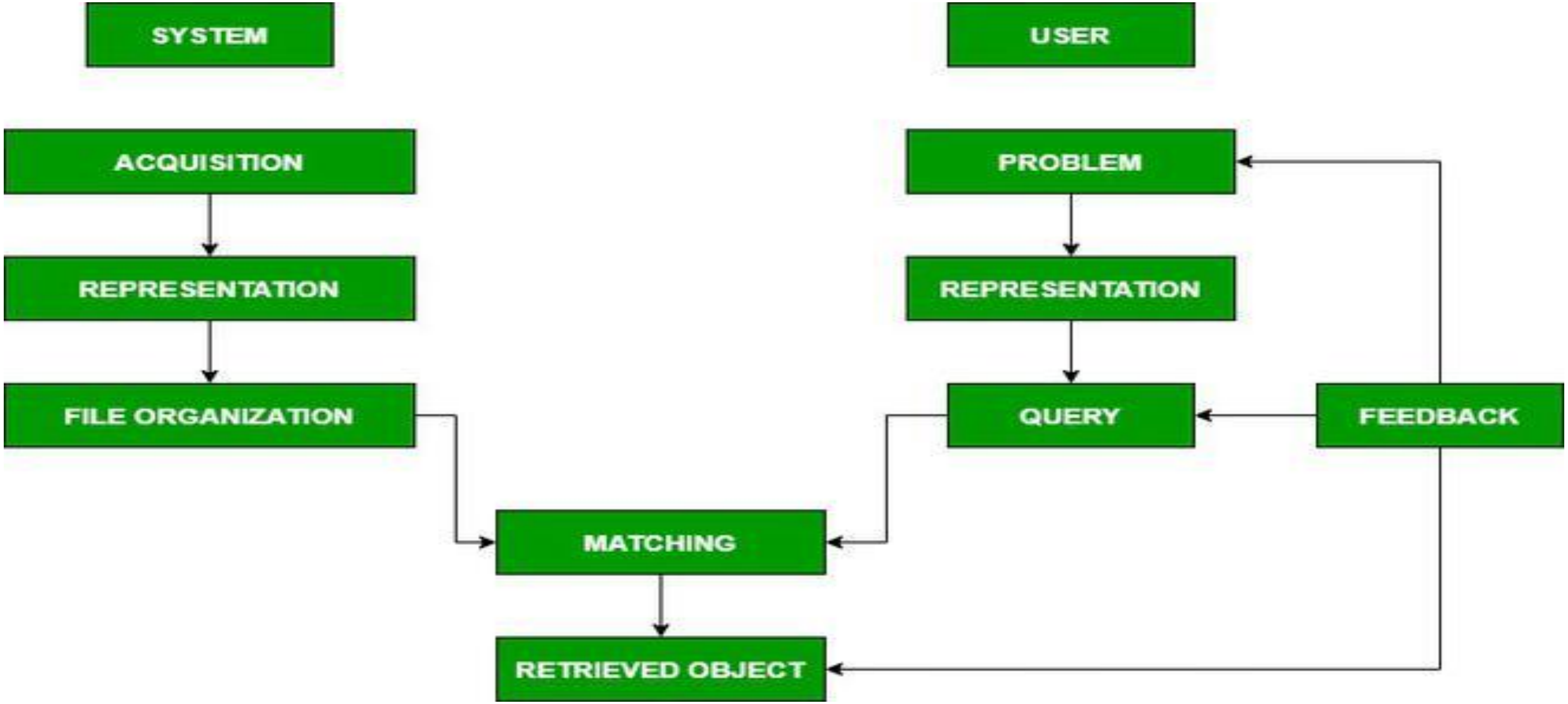
- An Information Retrieval (IR) model selects and ranks the document that is required by the user or the user has asked for in the form of a query.

- The documents and the queries are represented in a similar manner, so that document selection and ranking can be formalized by a matching function that returns a retrieval status value (RSV) for each document in the collection.

- Many of the Information Retrieval systems represent document contents by a set of descriptors, called terms, belonging to a vocabulary V.

- An IR model determines the query-document matching function according to four main approaches:

IR Model/Components of IR



IR Model/Components of IR

- **Acquisition:** In this step, the selection of documents and other objects from various web resources that consist of text-based documents takes place. The required data is collected by web crawlers and stored in the database.

- **Representation:** It consists of indexing that contains free-text terms, controlled vocabulary, manual & automatic techniques as well. example: Abstracting contains summarizing and Bibliographic description that contains author, title, sources, data, and metadata.

- **File Organization:** There are two types of file organization methods. i.e. *Sequential*: It contains documents by document data. *Inverted*: It contains term by term, list of records under each term. *Combination* of both.

- **Query:** An IR process starts when a user enters a query into the system. Queries are formal statements of information needs, for example, search strings in web search engines. In information retrieval, a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

Futures of Information Retrieval System

- **Early Developments:** As there was an increase in the need for a lot of information, it became necessary to build data structures to get faster access. The index is the data structure for faster retrieval of information. Over centuries manual categorization of hierarchies was done for indexes.

- **Information Retrieval In Libraries:** Libraries were the first to adopt IR systems for information retrieval. In first-generation, it consisted, automation of previous technologies, and the search was based on author name and title. In the second generation, it included searching by subject heading, keywords, etc. In the third generation, it consisted of graphical interfaces, electronic forms, hypertext features, etc.

- **The Web and Digital Libraries:** It is cheaper than various sources of information, it provides greater access to networks due to digital communication and it gives free access to publish on a larger medium.

Advantages of Information Retrieval

- **Efficient Access:** Information retrieval techniques make it possible for users to easily locate and retrieve vast amounts of data or information.

- **Personalization of Results:** User profiling and personalization techniques are used in information retrieval models to tailor search results to individual preferences and behaviors.

- **Scalability:** Information retrieval models are capable of handling increasing data volumes.

- **Precision:** These systems can provide highly accurate and relevant search results, reducing the likelihood of irrelevant information appearing in search results.

Disadvantages of Information Retrieval

- **Information Overload**: When a lot of information is available, users often face information overload, making it difficult to find the most useful and relevant material.

- **Lack of Context:** Information retrieval systems may fail to understand the context of a user's query, potentially leading to inaccurate results.

- **Privacy and Security Concerns:** As information retrieval systems often access sensitive user data, they can raise privacy and security concerns.

- **Maintenance Challenges:** Keeping these systems up-to-date and effective requires ongoing efforts, including regular updates, data cleaning, and algorithm adjustments.

Information Extraction

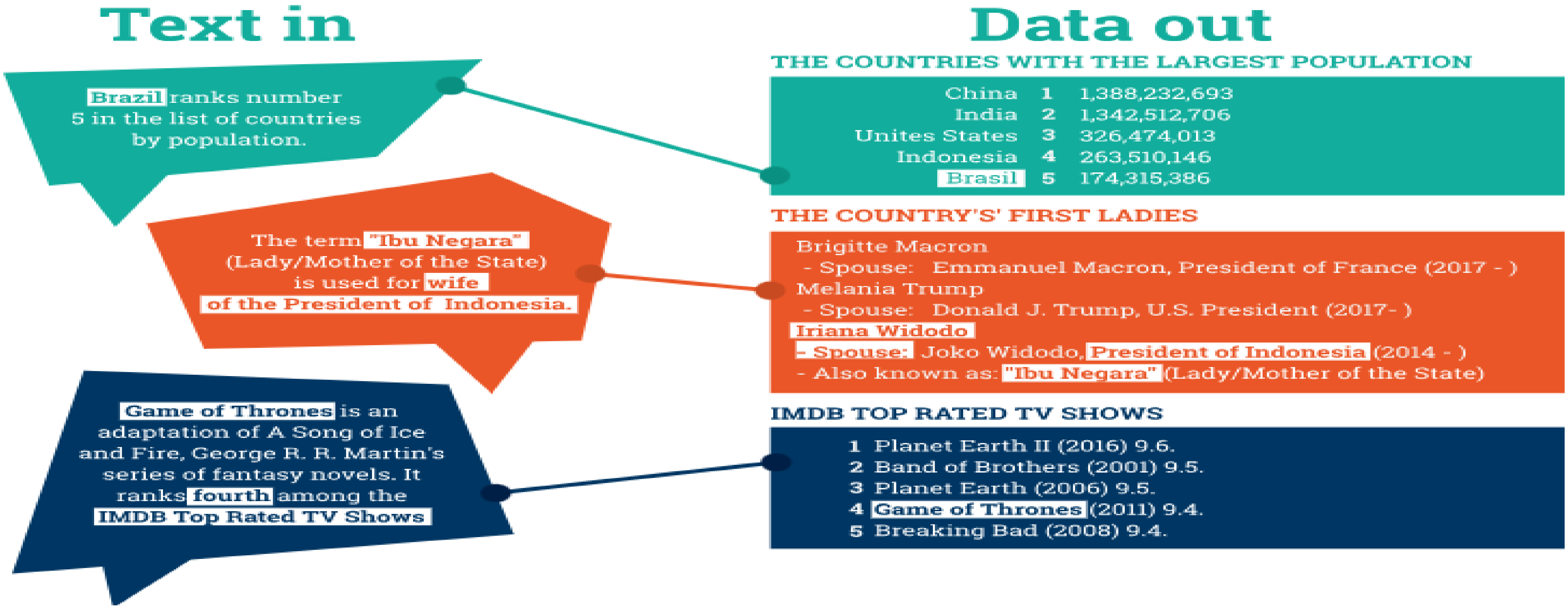
Information Extraction

- Information extraction is the process of extracting information from unstructured textual sources to enable finding entities as well as classifying and storing them in a database.

- Semantically enhanced information extraction (also known as semantic annotation) couples those entities with their semantic descriptions and connections from a knowledge graph.

- Information extraction is the process of extracting specific (pre-specified) information from textual sources.

How Does information Extraction Work



Text in

Brazil ranks number 5 in the list of countries by population.

The term "Ibu Negara" (Lady/Mother of the State) is used for wife of the President of Indonesia.

Game of Thrones is an adaptation of A Song of Ice and Fire, George R. R. Martin's series of fantasy novels. It ranks fourth among the IMDB Top Rated TV Shows

Data out

THE COUNTRIES WITH THE LARGEST POPULATION

China	1	1,388,232,693
India	2	1,342,512,706
Unites States	3	326,474,013
Indonesia	4	263,510,146
Brasil	5	174,315,386

THE COUNTRY'S' FIRST LADIES

Brigitte Macron
- Spouse: Emmanuel Macron, President of France (2017 -)

Melania Trump
- Spouse: Donald J. Trump, U.S. President (2017-)

Iriana Widodo
- Spouse: Joko Widodo, President of Indonesia (2014 -)
- Also known as: "Ibu Negara" (Lady/Mother of the State)

IMDB TOP RATED TV SHOWS

1 Planet Earth II (2016) 9.6.
2 Band of Brothers (2001) 9.5.
3 Planet Earth (2006) 9.5.
4 Game of Thrones (2011) 9.4.
5 Breaking Bad (2008) 9.4.

How Does information Extraction Work

For structured information to be extracted from unstructured texts, the following main subtasks are involved:

- **Pre-processing of the text** – this is where the text is prepared for processing with the help of computational linguistics tools such as tokenization, sentence splitting, morphological analysis, etc.
- **Finding and classifying concepts** – this is where mentions of people, things, locations, events and other pre-specified types of concepts are detected and classified.
- **Connecting the concepts** – this is the task of identifying relationships between the extracted concepts.
- **Unifying** – this subtask is about presenting the extracted data into a standard form.
- **Getting rid of the noise** – this subtask involves eliminating duplicate data.
- **Enriching your knowledge base** – this is where the extracted knowledge is ingested in your database for further use.