5.1 INFORMATION RETRIVAL

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).Generically, ―collections‖, Less-frequently used, ―corpora‖ are searched and ―documents‖ namely web pages, PDFs, PowerPoint slides, paragraphs, etc. are retrieved. Information Retrieval system consists of a software program that facilitates a user in finding the information the user needs.

The Information Retrieval System was coined by Calvin Mooers in 1952. These information retrieval systems were, truly speaking, document retrieval system, since they were designed to retrieve information. Information retrieval deals with storage, organization and access to text, as well as multimedia information resources. Information Retrieval is a process of searching some collection of documents, using the term document in its widest sense, in order to identify those documents which deal with a particular subject. Any system that is designed to facilitate this literature searching may legitimately be called an information retrieval system.

Information retrieval systems originally meant text retrieval systems, since they were dealing with textual documents, modern information retrieval systems deal with multimedia information comprising text, audio, images and video. While many features of conventional text retrieval system are equally applicable to multimedia information retrieval, the specific nature of audio, image and video information have called for the development of many new tools and techniques for information retrieval.

Modern information retrieval deals with storage, organization and access to text, as well as multimedia information resources. The concept of information retrieval presupposes that there are some documents or records containing information that have been organized in an order suitable for easy retrieval. The documents or records we are concerned with contain bibliographic information which is quite different from other kinds of information or data. We may take a simple example. If we have a database of information pertaining to an office, or a supermarket, all we have are the different kinds of records and related facts, like names of employees, their positions, salary, and so on, or in the case of a supermarket, names of different items, prices, quantity, and so on. The main objective of a bibliographic information retrieval system, however, is to retrieve the information either the actual information or the documents containing the information that fully or partially match the user_s query. The database may contain abstracts or full texts of document, like newspaper articles, handbooks, dictionaries, encyclopedias, legal documents, statistics, etc., as well as audio, images, and video information.

An information retrieval system thus has three major components- the document subsystem, the users subsystem, and the searching/retrieval subsystem. These divisions are quite

broad and each one is designed to serve one or more functions, such as:

- Analysis of documents and organization of information (creation of a document database)
- Analysis of user_s queries, preparation of a strategy to search the database
- Actual searching or matching of users queries with the database, and finally
- Retrieval of items that fully or partially match the search statement.

An IR is a 3 step Process:

- Asking a question (how to use the language to get what we want?)
- Building an answer from known data. (How to refer to a given text?)
- Assessing the answer. (Does it contain the information we are seeking.)

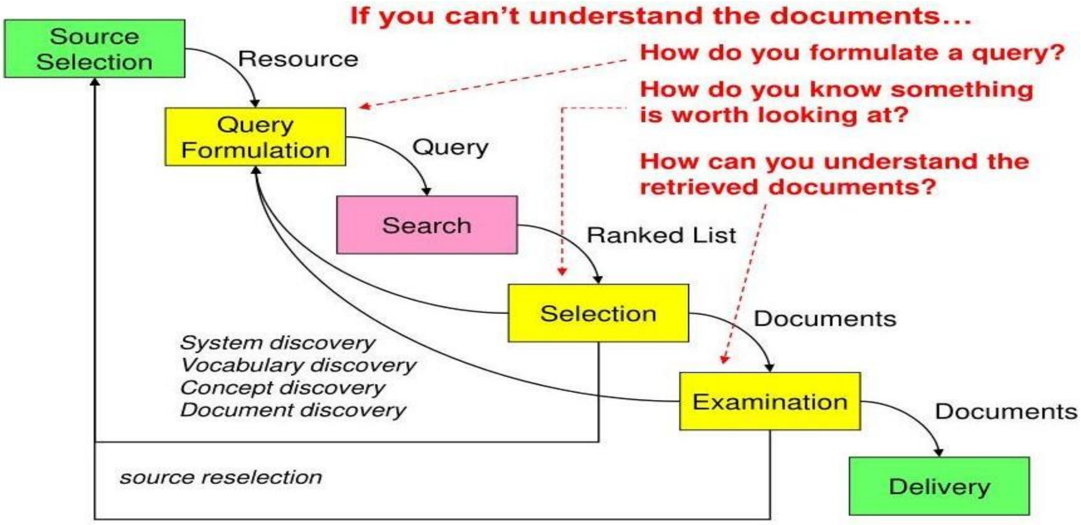
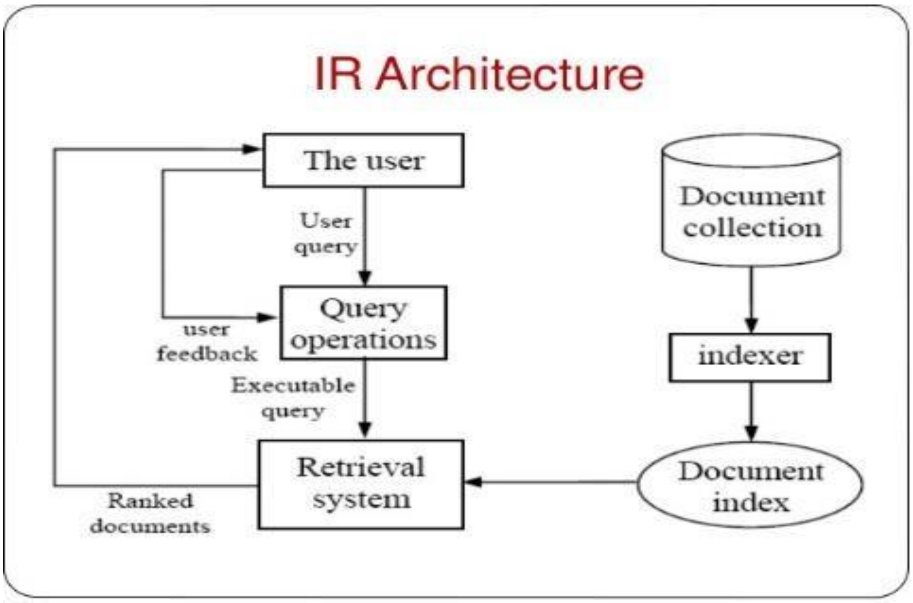


Fig: The Information Retrieval Cycle



5.1.1 IR System Components

- Text Operations forms index words (tokens).
  - ✓ Stop word removal
  - ✓ Stemming
- Indexing constructs an inverted index of word to document pointers.
- Searching retrieves documents that contain a given query token from the inverted index.
- Ranking scores all retrieved documents according to a relevance metric.
- User Interface manages interaction with the user:
  - ✓ Query input and document output.
  - ✓ Relevance feedback.
  - ✓ Visualization of results.
- Query Operations transform the query to improve retrieval:
  - ✓ Query expansion using a thesaurus.
  - ✓ Query transformation using relevance feedback.

5.1.2 Purpose of Information Retrieval System

An information retrieval system is designed to retrieve the documents or information required by the user community. It should make the right information available to the right user. Thus, an information retrieval system aims at collecting and organizing information in one or more subject areas in order to provide it to the user as soon asked for. Belkin presents the following situation which clearly reflects the purpose of information retrieval systems:

- A writer presents as set of ideas in a document using a set of concepts
- Somewhere there will be some users who require the ideas but may not be able to identify those. In other words, there will be some persons who lack the ideas put forward by the author in his/her work.
- Information retrieval system serve to match the writers ideas expressed in the document with the user requirements or demand for those.
- Thus, an information retrieval system serves as a bridge between the world of creators or generators of information and the users of that information.

Some terminology

- An IR system looks for data matching using some criteria defined by the users in their queries.
- The language used to ask a question is called the query language.
- These queries use keywords (atomic items characterizing some data).
- The basic unit of data is a document (can be a file, an article, a paragraph, etc.).
- A document corresponds to free text (may be unstructured).
- All the documents are gathered into a collection (or corpus).

Example:

1 million documents, each counting about 1000 words

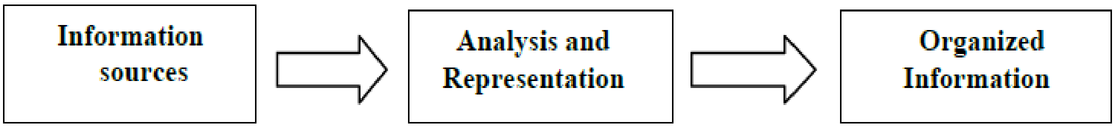
if each word is encoded using 6 bytes:

109 × 1000 × 6/1024 ≃ 6GB

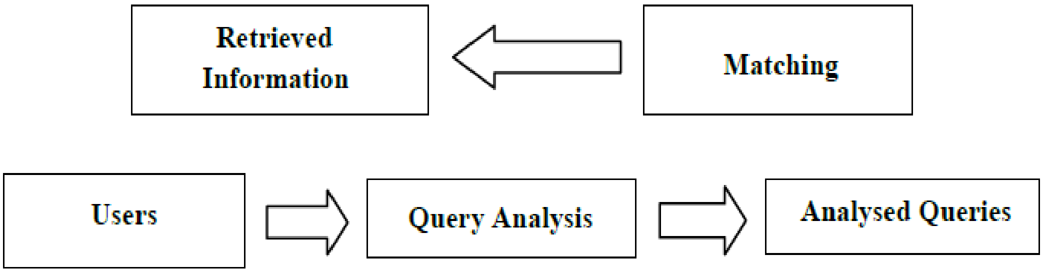
5.1.3 Components of Information Retrieval

In an information retrieval system there are the documents or sources of information on one side and on the other there are the user_s queries. These two sides are linked through a series of tasks. Lancaster mentions that an information retrieval system comprises six major subsystems: The document subsystem

- ✓ The indexing subsystem
- ✓ The vocabulary subsystem
- ✓ The searching subsystem
- ✓ The service-system interface, and
- ✓ The matching subsystem



The broad outline of an information retrieval system



Three major components of IRS

1) Document subsystem

   a) Acquisition

   b) Representation

   c) File organization

2) User sub system

   a) Problem

   b) Representation

   c) Query

3) Searching /Retrieval subsystem

   a) Matching

   b) Retrieved objects

5.1.4 Kinds of Information Retrieval Systems

Two broad categories of information retrieval system can be identified: in- house and online.

In- house information retrieval systems are set up by a particular library or information center to serve mainly the users within the organization. One particular type of in-house database is the library catalogue. Online public access catalogues (OPACs) provide facilities for library users to carry out online catalogue searches, and then to check the availability of the item required. Online IR is nothing but retrieving data from web sites, web pages and servers that may include data bases, images, text, tables, and other types.

5.3.4 Functions of information retrieval system

An information retrieval system deals with various sources of information on the one hand and user_s requirements on the other. It must:

- ✓ Analyze the contents of the sources of information as well as the user_s queries, and then
- ✓ Match these to retrieve those items that are relevant

The major functions of an information retrieval system can be listed as follows:

- ✓ To identify the information (sources) relevant to the areas of interest of the target users community
- ✓ To analyze the contents of the sources (documents)
- ✓ To represent the contents of the analyzed sources in a way that will be suitable for matching user_s queries
- ✓ To analyze user_s queries and to represent them in a form that will be suitable for matching with the database
- ✓ To match the search statement with the stored database
- ✓ To retrieve the information that is relevant, and
- ✓ To make necessary adjustments in the system based on feedback form the users.

5.3.5 Features of an information retrieval system

- ✓ An effective information retrieval system must have provisions for:
- ✓ Prompt dissemination of information
- ✓ Filtering of information
- ✓ The right amount of information at the right time
- ✓ Active switching of information
- ✓ Receiving information in an economical way
- ✓ Browsing
- ✓ Getting information in an economical way
- ✓ Current literature
- ✓ Access to other information systems
- ✓ Interpersonal communications, and
- ✓ Personalized help.

5.3.6 Indexing usually consists of the several phases

- ✓ After word segmentation, stop words are removed.
- ✓ These common words like articles or prepositions contain little meaning by themselves and are ignored in the document representation.
- ✓ Second, word forms are transformed into their basic form, the stem.
- ✓ During the stemming phase, e.g. houses would be transformed into house.
- ✓ For the document representation, different word forms are usually not necessary.
- ✓ The importance of a word for a document can be different.
- ✓ Some words better describe the content of a document than others.
- ✓ This weight is determined by the frequency of a stem within the text of a document.

In multimedia retrieval, the context is essential for the selection of a form of query and document representation. Different media representations may be matched against each other or transformations may become necessary (e.g. to match terms against pictures or spoken language utterances against documents in written text).

As information retrieval needs to deal with vague knowledge, exact processing methods are not appropriate.

- Vague retrieval models like the probabilistic model are more suitable.
- Within these models, terms are provided with weights corresponding to their importance for a document.
- These weights mirror different levels of relevance.

The result of current information retrieval systems are usually sorted lists of documents where the top results are more likely to be relevant according to the system.

- In some approaches, the user can judge the documents returned to him and tell the systems which ones are relevant for user.
- The system then resorts the result set.
- Documents which contain many of the words present in the relevant documents are ranked higher.
- This relevance feedback process is known to greatly improve the performance.
- Relevance feedback is also an interesting application for machine learning.
- Based on a human decisions, the optimization step can be modeled with several approaches, e.g. with rough sets.

5.2 INFORMATION EXTRACTION

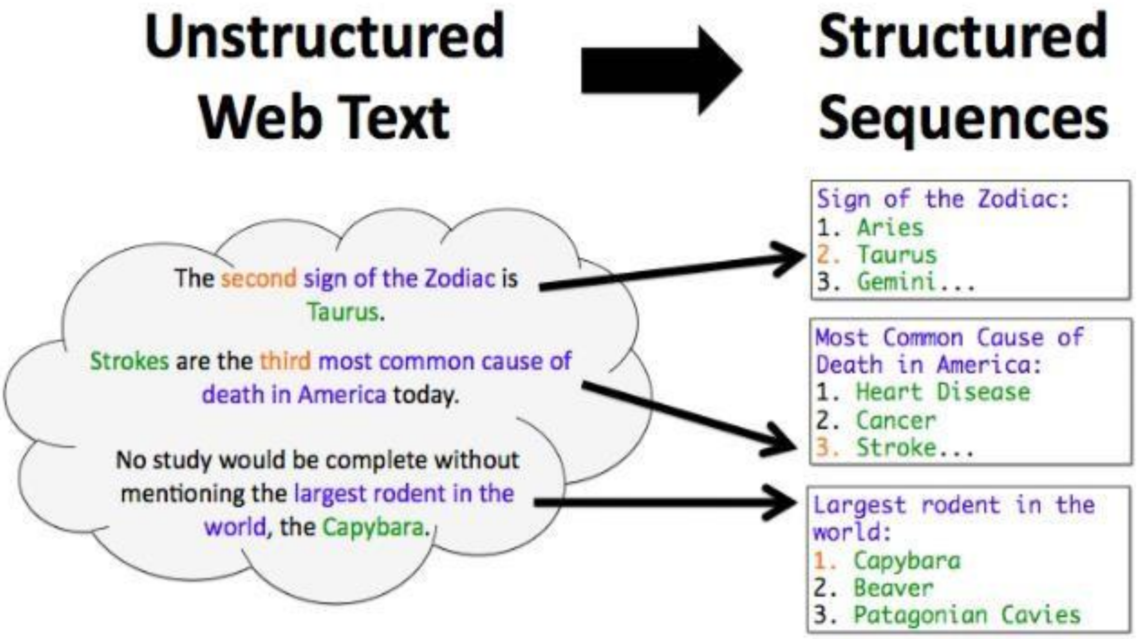
Information extraction (IE) is the automated retrieval of specific information related to a selected topic from a body or bodies of text. Information extraction is the process of extracting specific (pre-specified) information from textual sources. One of the most trivial examples is when your email extracts only the data from the message for you to add in your Calendar.

Other free-flowing textual sources from which information extraction can distill structured information are legal acts, medical records, social media interactions and streams, online news, government documents, corporate reports and more.

Information extraction tools make it possible to pull information from text documents, databases, websites or multiple sources. IE may extract info from unstructured, semi-structured or structured, machine-readable text. Usually, however, IE is used in natural language processing (NLP) to extract structured from unstructured text.

Information extraction depends on,

- Named entity recognition (NER), a sub-tool used to find targeted information to extract.
- NER recognizes entities first as one of several categories such as location (LOC), persons (PER) or organizations (ORG).
- Once the information category is recognized, an information extraction utility extracts the named entity_s related information and constructs a machine-readable document from it, which algorithms can further process to extract meaning.
- IE finds meaning by way of other subtasks including co-reference resolution, relationship extraction, language and vocabulary analysis and sometimes audio extraction.
- Current efforts in multimedia document processing in IE include automatic annotation and content recognition and extraction from images and video could be seen as IE as well.
- Because of the complexity of language, high-quality IE is a challenging task for artificial intelligence (AI) systems.


Unstructured Web Text
Structured Sequences
The second sign of the Zodiac is Taurus.
Strokes are the third most common cause of death in America today.
No study would be complete without mentioning the largest rodent in the world, the Capybara.
Sign of the Zodiac:
1. Aries
2. Taurus
3. Gemini...
Most Common Cause of Death in America:
1. Heart Disease
2. Cancer
3. Stroke...
Largest rodent in the world:
1. Capybara
2. Beaver
3. Patagonian Cavies

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

Pre-processing of text – where 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 – task of identifying relationships between 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.

5.2.1 Information Extraction Architecture

The below figure shows the architecture for a simple information extraction system. At first, the raw text of the document is split into sentences using a sentence segmenter, and each sentence is further subdivided into words using a tokenizer. Next, each sentence is tagged with part-of-speech tags, which will prove very helpful in the next step, **named entity detection**. In this step, we search for mentions of potentially interesting entities in each sentence. Finally, we use **relation detection** to search for likely relations between different entities in the text.

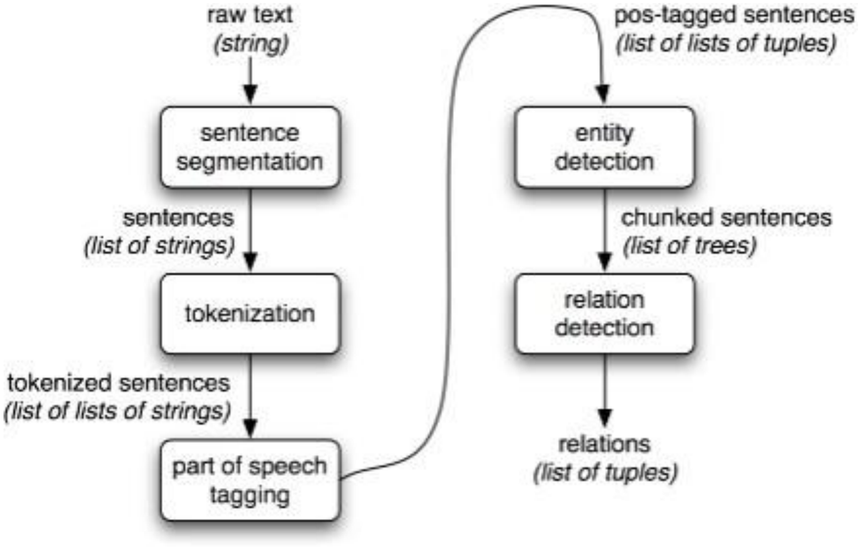


Figure : Simple Pipeline Architecture for an Information Extraction System.

This system takes the raw text of a document as its input, and generates a list of (entity, relation, entity) tuples as its output.

5.2.2 Applications of IE

- Enterprise
- News tracking
- Customer care
- Data cleaning
- Personal information management
- Scientific applications
- Web oriented applications
- Citation databases
- Opinion databases
- Community websites
- Comparison shopping
- Ad placement on webpages
- Structured web searches