Information Retrieval

By

Dr Syed Khaldoon Khurshid

Focus of the week

- **Introduction of the course**
- **Basic Concepts**
- **Introduction to Information Retrieval**
- **Information Retrieval in the Library**
- **Information Retrieval Systems Operations**
- **The Retrieval Process**
- **Case Study Project:**

Design a Simple Document Search Engine

Things to know before the commencement of the course:

- **Reference Book:** Modern Information Retrieval by Ricardo Baeza Yates, Berthier Ribeiro Neto (Author), Pearson Education Limited (Publisher), 2nd Edition Onwards

- **Office Days**: Wednesday and Thursday

- **Evaluation Components weightings:**

  Mid-term: 20%

  Assignments: 50%

  Final paper: 30%

Evaluation Components marks:

Mid-term: 30

Assignments: 20

Final paper: 40

Course Description:

This course provides an in-depth exploration of the concepts in information retrieval, focusing on the integration of real-world case studies or assignments. Students will learn how information retrieval systems work, delve into various retrieval models, and analyze practical applications through case studies or assignments. The course covers both theoretical foundations and hands-on implementation, equipping students with the skills to design and evaluate effective information retrieval solutions.

Course Objectives:

1. Understand the fundamentals of information retrieval, including retrieval models, indexing, and ranking algorithms.
2. Explore advanced topics such as user modeling, query expansion, and result diversification.
3. Analyze real-world case studies to identify challenges and solutions in information retrieval.
4. Develop skills in implementing and evaluating information retrieval systems.
5. Foster critical thinking and problem-solving abilities through assignment-based learning.

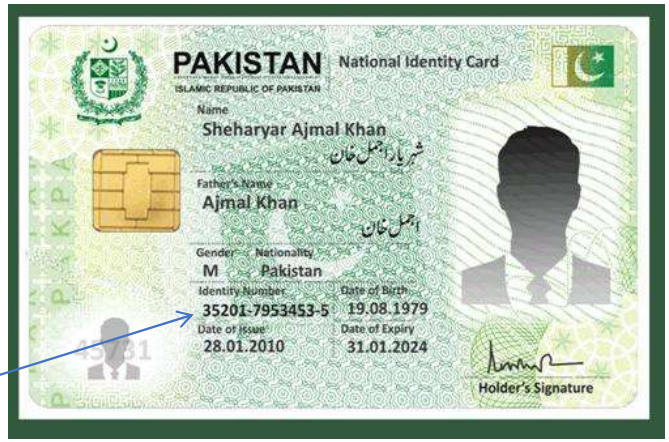
Basic Concepts:

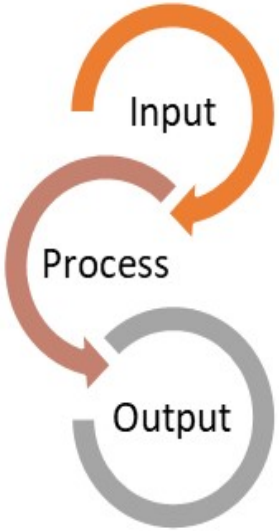
- Data → Raw fact
- Example of Data
  3520179534535

- Information → Processed Data
- Example of Information
  CNIC: 35201-7953453-5

- Main Functions related to Data or Information:
  Store, Process, Representation and Access

  Input → Process→ Outcome





Example: Bio-data details hidden in the CNIC

Understanding in the terms of Computers:

- Data Structures → Data in Memory
- Database Management Systems → Data in Hard Disk
- Data Science → Data in Machine Learning/AI
- **Information Retrieval** → **Searching Data/Information**

MOTIVATION

INFORMATION RERIEVAL(IR)

Representation, storage, organization of, and access to information items.

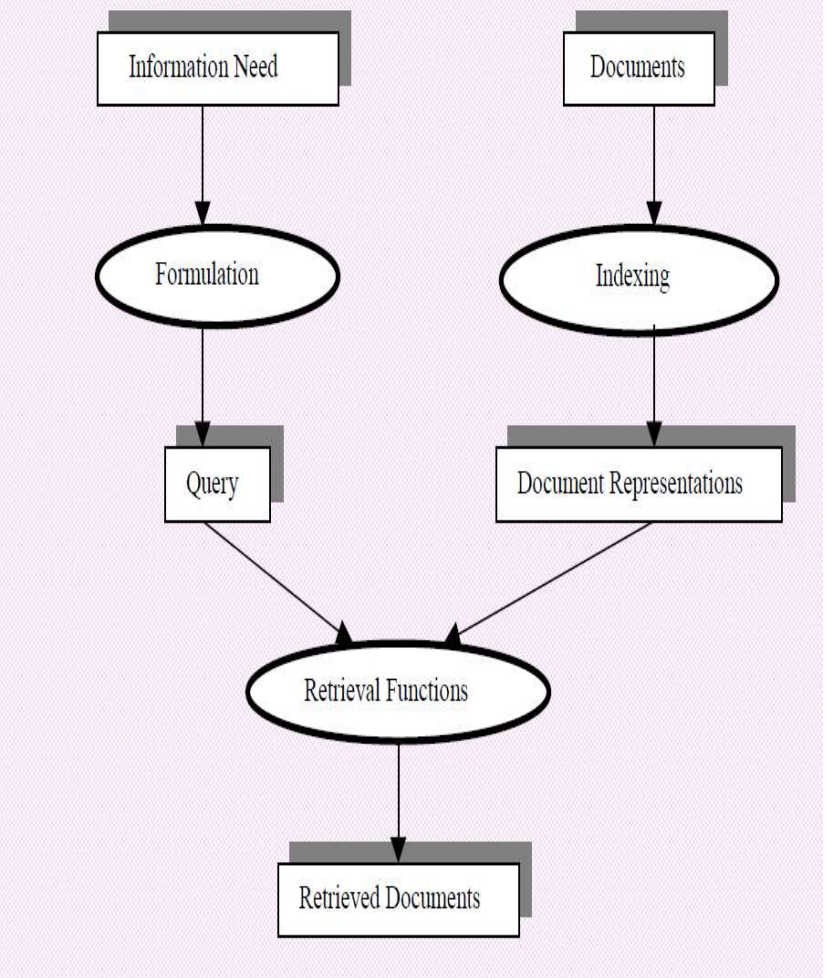
- Focus is on the *user information need*.
- **Information item:** Usually text, but possibly also image, audio, video, etc.
- Text items are often referred to as documents, and may be of different scope (book, article, paragraph, etc.).

Information is communicated or received knowledge concerning a particular fact or circumstance. Retrieval refers to searching through stored information to find information relevant to the task at hand.

MOTIVATION

- **User information need:**
  - *Find all docs containing information on Computer Systems which: (1) are designed by IBM and (2) designed after 2020.*

  - User information need cannot be used directly to request information using the current interfaces of Web search engines.
  - Emphasis is on the retrieval of information **not data**.



Information Retrieval Systems

IRS is with two basic aspects:

(i) How to **store** information, and

(ii) How to **retrieve** information.

Information Retrieval text based:

- **An information retrieval system is designed to analyze process and store sources of information and retrieve those that match a particular users requirements.** Modern information retrieval systems can either retrieve bibliographic items or the exact text that matches a user's search criteria from a stored database of documents.
- **IRS originally meant text retrieval systems as they were dealing with textual documents**.

Information Retrieval in the Library

- Libraries were among the first institutions to adopt systems for retrieving information.

- In the **first generation**, such systems consisted basically of an automation of previous technologies (such as card catalogs).

- In the **second generation**, increased search functionality was added which allowed searching by subject headings, by keywords, and some more complex query facilities.

- In the **third generation**, which is currently being deployed, the focus is on improved graphical interfaces, electronic forms, hypertext features, and open system architectures

Information Retrieval (IR) in Library

- The principal function of any library is to make available to the users, the **information** they need. In order to fulfill this function, the information which is stored in the library must be retrieved from the library database.
- **Information retrieval (IR)** is the activity of obtaining information resources relevant to an information need from a collection of information resources.

Importance of Information Retrieval in Library

- Libraries contain information in various **physical forms**. While for many users, the book is still a major vehicle for communication of information for others, the periodical or the technical report have taken its place; and for yet others, films or gramophone records are significant.
- It is clear that the same work can appear in various physical forms. The intellectual content will be the same in each case, but obviously it is not practical to try to arrange the different physical forms together.
- The library catalogue, however, is only one of the tools which serve as the key to library documents. A library containing a large number of periodicals will not attempt to list all articles of every issue if receives.
- Instead, we rely on **indexes, abstracts and similar bibliographic tools** which present the contents of periodicals as well. This enables us to obtain access to any particular item through number of approaches.

Modern Information Retrieval Systems

- Modern information retrieval systems deal not only with textual information but also with multimedia information comprising text, audio, images and video. Thus, modern information retrieval systems deal with storage, organization and access to text, as well as multimedia information resources.

The Web and Digital Libraries

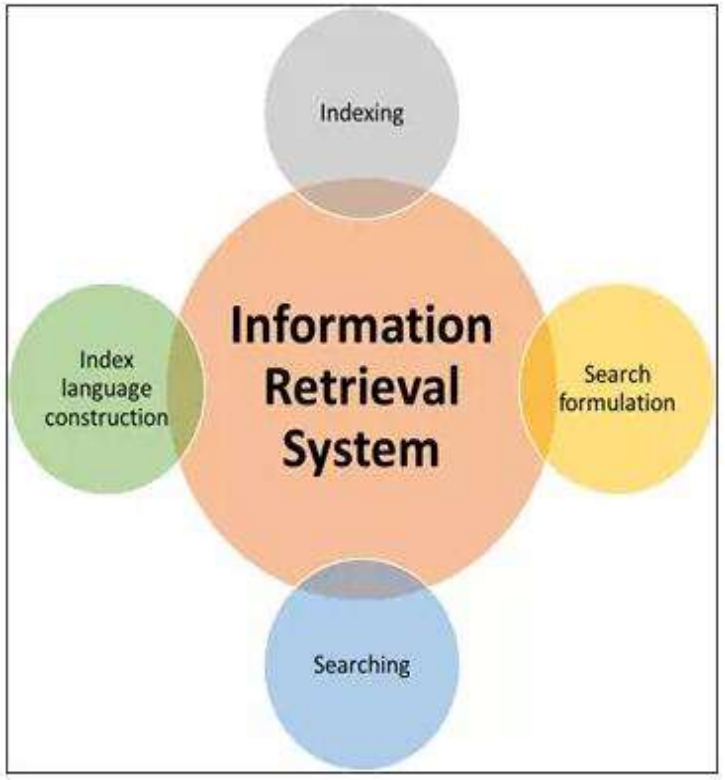
If we consider the search engines on the Web today, we conclude that **they continue to use indexes** which are very similar to those used by librarians a century ago. What has changed then?

Three dramatic and fundamental changes have occurred.

- It became **a lot cheaper to have access to various sources of information**
- The advances in **all kinds of digital communication provided greater access to networks**.
- The **freedom to post** whatever information someone judges useful has greatly contributed to the popularity of the Web.

Information Retrieval Systems Operations:

- **An IR system is a set of rules and procedures, for performing some or all of the following operations:**
- **Indexing** (or constructing of representations of documents)
- **Search formulation** (or constructing of representations of information needs)
- **Searching** (or matching representations of documents against representations of needs)
- **Index language construction** (or generation of rules of representation) So information retrieval is collectively defined as a "science of search" or a process, method and procedure used to select or recall, recorded and/or indexed information from files of data.



INFORMATION VS DATA RETRIEVAL

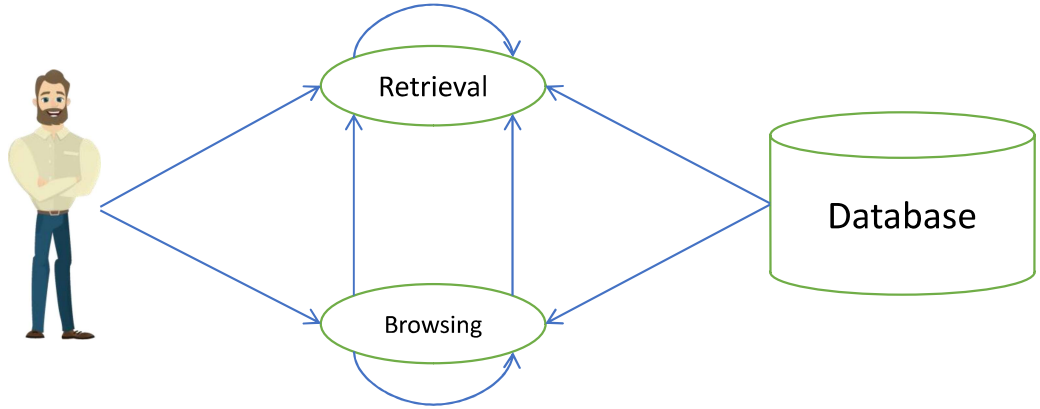
1. Information Retrieval	2. Data Retrieval
1) The software that deals with the organization, storage, retrieval, and evaluation of information from document particularly textual information.	1) Data retrieval deals with obtaining data from a database management system such as RDBMS.
2) Retrieves information about a subject.	2) Determines the keywords in the user query and retrieves the data.
3) Small errors are likely to go unnoticed.	3) A single error object means total failure.
4) Not always well structured and is semantically ambiguous.	4) Has a well-defined structure and semantics.
5) Does not provide a solution to the user of the database system.	5) Provides solutions to the user of the database system.
6) The results obtained are approximate matches.	6) The results obtained are exact matches.
7) Results are ordered by relevance.	7) Results are unordered by relevance.

Information Retrieval was the main focus:

- **PAST 30-40 YEARS**
  - Information Retrieval has grown well beyond
  - Research in IR includes modeling, document classification, user interface, languages etc.
- **BEGINNING OF 1990s**
  - Introduction of World Wide Web.
  - Its success is based on the conception of a standard user interface which is always the Same.
  - Any user can create his own Web documents Without any Restriction.
- **What about Now ??? Data Science Emergence**

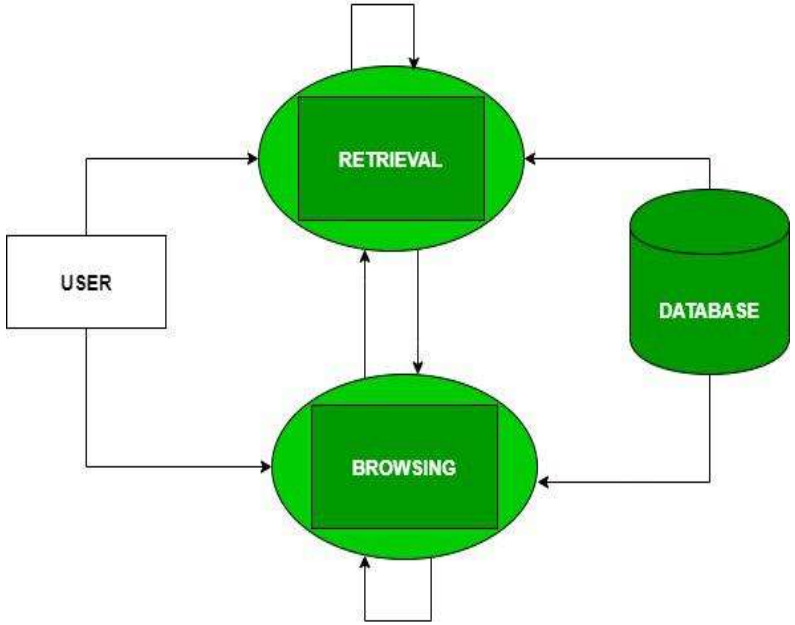
User Task

- The User of a retrieval system has to **translate his information need into a query.**
- With an information retrieval system, this implies **specifying a set of words.**
- With a data retrieval system, a **query expression is used to convey the constraints that must be satisfied by objects in the answer set.**



Retrieval Vs Browsing

Retrieval	Browsing
Information needed (retrieval goal) is focused and crystalized.	Information needed (retrieval goal) is vague and imprecise
Contents of repository are well-known.	Contents of repository are not well-known.
Often user is sophisticated (Database Engineer) .	Often user is naive (layman).

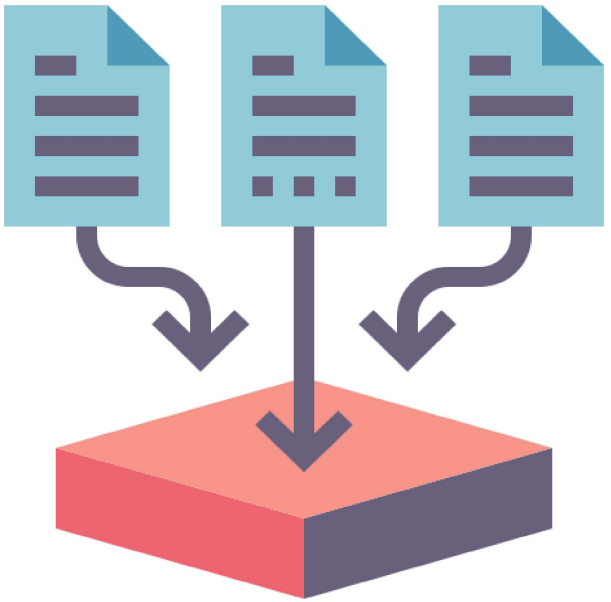


PAST, PRESENT AND FUTURE

- For approximately 4000 years, man has organized information for later retrieval and usage.
  Example is the table of contents of a book.

- In the computer-centered view, the IR problem consists mainly of building up **efficient indexes**

- In the human-centered view, the IR problem consists mainly of **studying the behavior of the user.**



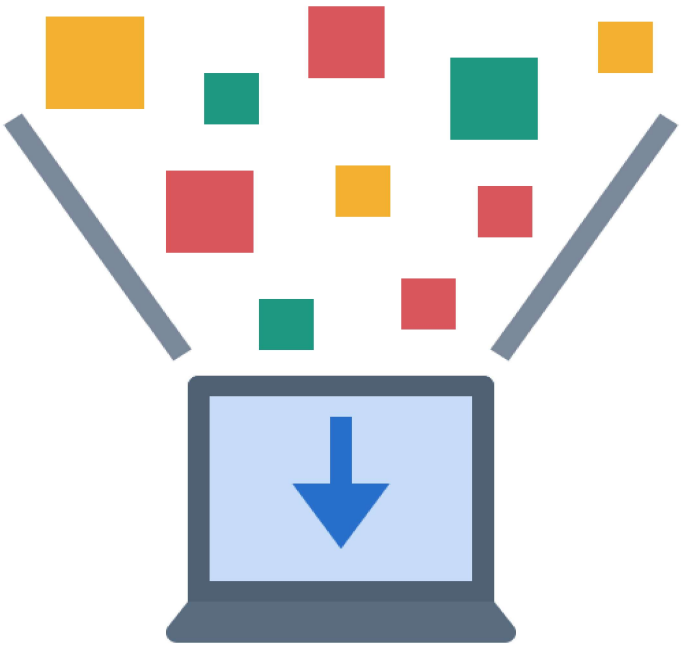
PRACTICAL ISSUES

- One of the major issue is **security.** In an electronic transaction, the buyer usually has to submit to the vendor some form of credit information which can be used for charging for the product or service.

- Two other very important issues are **copyright and patent rights.** It is far from clear how the wide spread of data on the Web affects copyright and patent laws in the various countries.

- Additionally, other practical issues of interest **include scanning**, optical character recognition (OCR), and **cross-language** retrieval (in which the query is in one language but the documents retrieved are in another language).

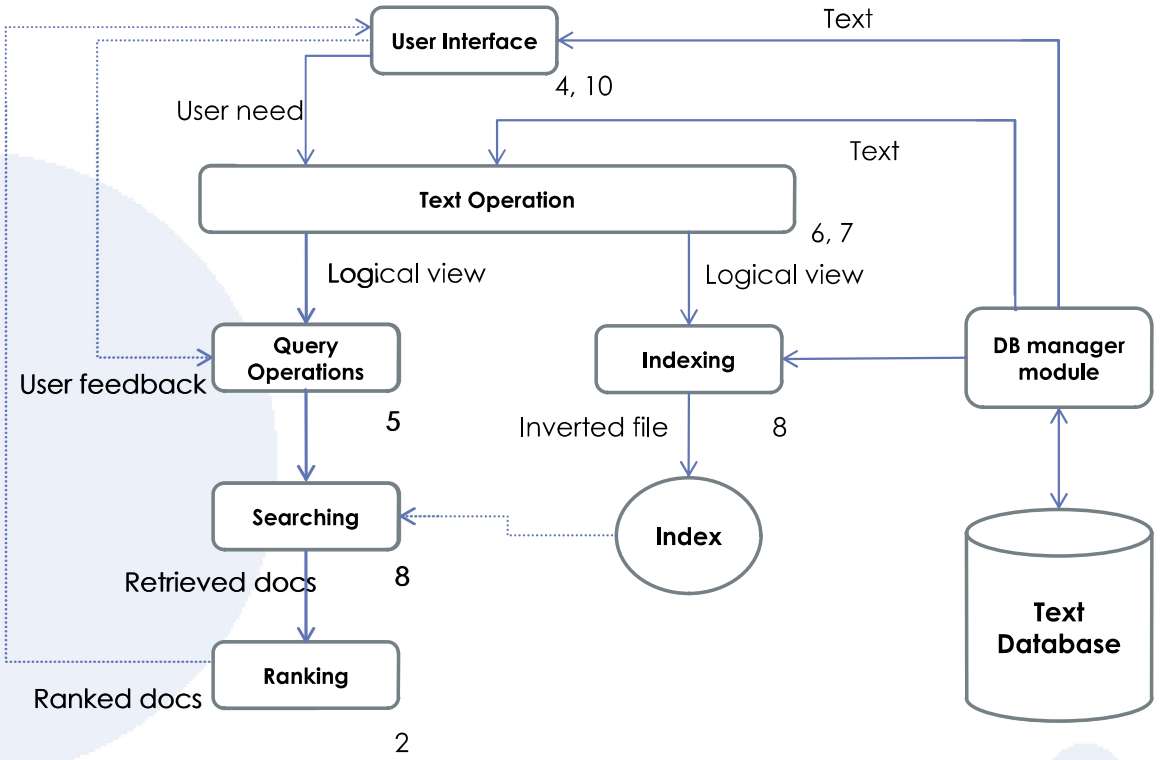
Retrieval Process



- To describe the retrieval process, we use **a simple and generic software architecture.**
- Given that the document database is **indexed**, the retrieval process can be **initiated.**
- The user first specifies a user need which is then **parsed** and **transformed** by the same text operations applied to the text.
- Then, **query operations** might be applied before the actual **query**, which provides a system representation for the user need, is generated.
- **The query is then processed to obtain the retrieved documents. Fast query processing is made possible by the index structure previously built.**

The Retrieval Process

At this point, we are ready to detail our view of the retrieval process. To describe a retrieval process, we use a simple and generic software architecture as shown below:



Assignment 1:

- **Design a Simple Document Search Engine**

  First: Searching the word in the document

Resources

Chapter 1 (Introduction) Book: Modern Information Retrieval

"ACM Press New York - University of California, Los Angeles." [Online]. Available: https://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf. [Accessed: 16-Feb-2023].

S. Rababah, T. User, Yates, Ribeiro, and Wesley, "Modern Information Retrieval - ppt download," *SlidePlayer*. [Online]. Available: https://slideplayer.com/slide/13106234/. [Accessed: 16-Feb-2023].

"What is information retrieval?," *GeeksforGeeks*, 03-Jul-2022. [Online]. Available: https://www.geeksforgeeks.org/what-is-information-retrieval. [Accessed: 16-Feb-2023].

https://www.flexiprep.com/NIOS-Notes/Senior-Secondary/Library-Science/NIOS-Library-Science-Unit-16-Information-Retrieval-System-Part-1-4.html [Accessed: 27 Aug-2023]


Any
Questions