Module 1

CHAPTER 1

Introduction to Information Retrieval

Syllabus

1.1 Introduction to Information Retrieval, Basic Concepts, Information Versus Data, Trends and research issues in information retrieval.

1.2 The retrieval process, Information retrieval in the library, web and digital libraries.

1.1 Defining Data, Information, and Knowledge, Wisdom 1-2
GQ. Define the term: data, information, knowledge, wisdom. 1-2
GQ. Justify "the data is different than the information". 1-2
1.2 Introduction to Information Retrieval 1-4
GQ. What is information retrieval? **(2 Marks)** 1-4
GQ. Define information retrieval **(2 Marks)** 1-4
GQ. How the information retrieval works? **(4 Marks)** 1-4
GQ. Explain the importance of information retrieval in brief. **(4 Marks)** 1-4
1.3 Information Retrieval Process 1-5
GQ. Draw and explain the information retrieval process? **(4 Marks)** 1-5
GQ. Explain the process of information retrieval with suitable example? **(6 Marks)** 1-5
GQ. Discuss the classical problem in information retrieval (IR) model? **(4 Marks)** 1-5
GQ. Explain how to represent the information retrieval model in mathematical terms? **(4 Marks)** 1-5
GQ. Explain with suitable diagram the components of information retrieval in detail. **(6 Marks)** 1-5
GQ. Differentiate between the data retrieval information retrieval. **(4 Marks)** 1-5
GQ. What are the applications of IR? **(4 Marks)** 1-5
GQ. Give the functions of information retrieval system. **(4 Marks)** 1-5
1.3.1 Classical Problem in Information Retrieval (IR) System 1-7
1.3.2 Information Retrieval (IR) Model 1-7
1.3.3 Components of Information Retrieval/ IR Model 1-8
GQ. What are the components of IR? **(4 Marks)** 1-8
1.3.4 Difference between Information Retrieval and Data Retrieval 1-8
1.3.5 User Interaction with Information Retrieval System 1-9
1.4 Past, Present, and Future of Information Retrieval 1-9
1.5 Objectives and Functions of IRS 1-10
GQ. Explain the objectives and functions of Information retrieval (IR) system in details. **(4 Marks)** 1-10
1.6 Issues in Information Retrieval 1-10
GQ. List and explain the issues in information retrieval. **(4 Marks)** 1-11
1.7 Process of IR 1-12
GQ. Explain the process / architecture of IR. **(4 Marks)** 1-12
1.8 Information retrieval in the library, web and digital libraries 1-13
GQ. Write a short note on: IR in library, digital libraries. 1-13
GQ. Discuss the impact of IR on the web. **(4 Marks)** 1-13
GQ. Discuss web information retrieval system. **(4 Marks)** 1-13
❖ **Chapter Ends** 1-15

1.1 DEFINING DATA, INFORMATION, AND KNOWLEDGE, WISDOM

GQ. Define the term: data, information, knowledge, wisdom.

GQ. Justify "the data is different than the information".

- According to Russell Ackoff, a systems theorist and professor of organizational change, the content of the human mind can be classified into five categories:

  1. **Data :** Symbols
  2. **Information :** Data that are processed to be useful; provides answers to "who", "what", "where", and "when" questions
  3. **Knowledge :** Application of data and information; answers "how" questions
  4. **Understanding :** Appreciation of "why"
  5. **Wisdom :** Evaluated understanding.

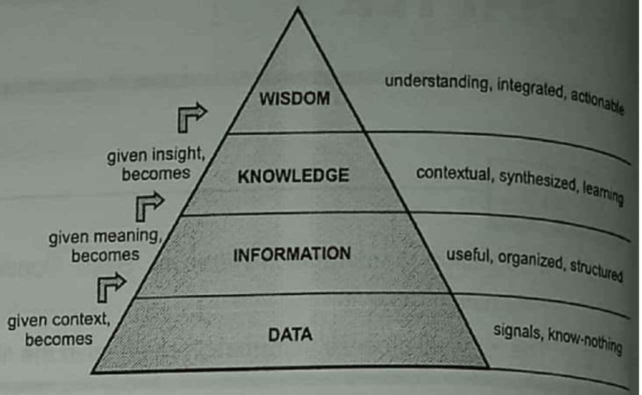


Fig. 1.1.1

A further elaboration of Ackoff's definitions follows :

1. **Data :** This is unprocessed information. It simply exists and has no use (in and of itself). It can exist in any shape or form, whether or not it is usable. It has no intrinsic value. A spreadsheet is a type of computer programme that begins by storing information.

   - Data represents a fact or statement of event without relation to other things.

     Ex : It is raining.

2. **Information :** Data that has been given meaning through a relationship is referred to as information. This "meaning" may or may not be useful.

   A relational database is a type of database that generates information from the data it holds.

   - Information embodies the understanding of a relationship of some sort, possibly cause and effect.

     Ex : The temperature dropped 15 degrees and then it started raining.

3. **Knowledge :** Knowledge is a useful collection of data. The process of learning is deterministic. When someone "memorizes" facts (as many less-motivated test-takers do), they have amassed knowledge. This knowledge is helpful to them, but it does not provide for an integration that would lead to the acquisition of other knowledge.

   - Knowledge represents a pattern that connects and generally provides a high level of predictability as to what is described or what will happen next.

     Ex : If the humidity is very high and the temperature drops substantially the atmospheres is often unlikely to be able to hold the moisture so it rains.

4. **Understanding :** Understanding is a probabilistic and interpolative process. It is both cognitive and analytical in nature. It's the method by which I can take previously acquired knowledge and synthesis new information from it.

- The difference between understanding and knowledge is the difference between "learning" and "remembering". People who have understanding can undertake useful actions because they can synthesize new knowledge, or in some cases, at least new information, from what is previously known (and understood).
- That is, understanding can build upon currently held information, knowledge and understanding itself.
- In computer phrasing, AI systems possess understanding in the sense that they are able to synthesize new knowledge from previously stored information and knowledge.
- **Wisdom :** Wisdom is an extrapolative and non-deterministic, non-probabilistic process. It calls upon all the previous levels of consciousness, and specifically upon special types of human programming (moral, ethical codes, etc.).

It beckons to give us understanding about which there has previously been no understanding, and in doing so, goes far beyond understanding itself. It is the essence of philosophical probing.

Unlike the previous four levels, it asks questions to which there is no (easily-achievable) answer, and in some cases, to which there can be no humanly-known answer period. Wisdom is therefore, the process by which we also discern, or judge, between right and wrong, good and bad.

- Wisdom embodies more of an understanding of fundamental principles embodied within the knowledge that are essentially the basis for the knowledge being what it is. Wisdom is essentially systemic.
- **Example :** It rains because it rains. And this encompasses an understanding of all the interactions that happen between raining, evaporation, air currents, temperature gradients, changes, and raining.

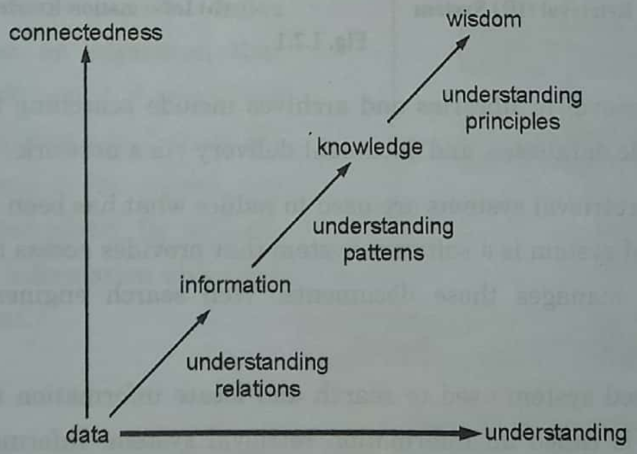
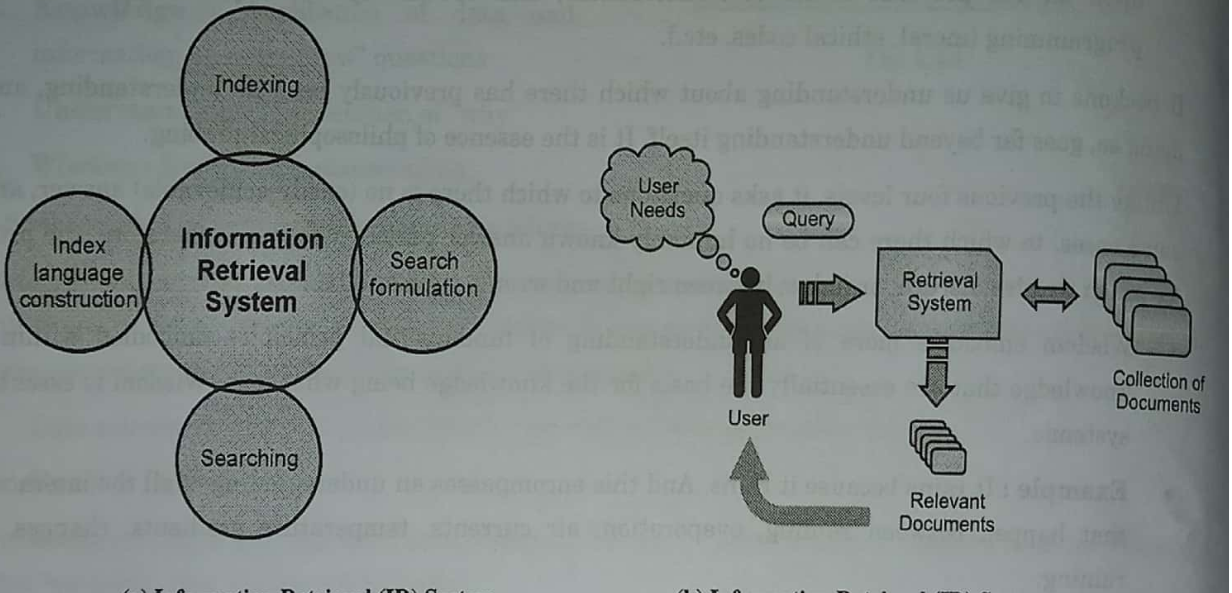


Fig. 1.1.2 : Representation of data, information, knowledge and wisdom

1.2 INTRODUCTION TO INFORMATION RETRIEVAL

GQ. What is information retrieval? (2 Marks)
GQ. Define information retrieval (2 Marks)
GQ. How the information retrieval works? (4 Marks)
GQ. Explain the importance of information retrieval in brief. (4 Marks)

- **Information Retrieval** refers to the process, methods, and procedures of searching, locating, and retrieving recorded data and information from a file or database.
- Information retrieval (IR) is the activity of obtaining information system resources that are relevant to an information need from a collection of those resources.
- Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.



(a) Information Retrieval (IR) System (b) Information Retrieval (IR) System

Fig. 1.2.1

- Modern information retrieval in libraries and archives include searching full-text databases, locating objects from bibliographic databases, and document delivery via a network.
- Automated information retrieval systems are used to reduce what has been called information overload. An Information Retrieval system is a software system that provides access to books, journals, and other documents; stores and manages those documents. Web search engines are the most visible IR applications.
- A print or computer-based system used to search and locate information in a file, database, or other collection of documents is called an information retrieval system. Information retrieval, Recovery of information, especially in a database stored in a computer.
- Two main approaches are matching words in the query against the database index is
  1. By using keyword searching

2. Traversing the database using hypertext or hypermedia links.

- Keyword searching has been the dominant approach to text retrieval since the early 1960's hypertext has so far been confined largely to personal or corporate information-retrieval applications.
- Natural language, hyperlinks, and keyword searching are all part of evolving information-retrieval approaches, as evidenced by modern Internet search engine improvements.

1.3 INFORMATION RETRIEVAL PROCESS

GQ. Draw and explain the information retrieval process? (4 Marks)

GQ. Explain the process of information retrieval with suitable example? (6 Marks)

GQ. Discuss the classical problem in information retrieval (IR) model? (4 Marks)

GQ. Explain how to represent the information retrieval model in mathematical terms? (4 Marks)

GQ. Explain with suitable diagram the components of information retrieval in detail. (6 Marks)

GQ. Differentiate between the data retrieval information retrieval. (4 Marks)

GQ. What are the applications of IR? (4 Marks)

GQ. Give the functions of information retrieval system. (4 Marks)

- An Information Retrieval system is supported by basic processes such as -

  1. Query Handling 2. Indexing and 3. Matching

  All these mentioned processes constitutes the Information Retrieval process.

There are three basic processes an information retrieval system has to support: the representation of the content of the documents, the representation of the user's information need, and the comparison of the two representations. The processes are visualized in Fig. 1.3.1. In the figure, squared boxes represent data and rounded boxes represent processes.

Representing the documents is usually called the indexing process. The process takes place off-line, that is, the end user of the information retrieval system is not directly involved. The indexing process results in a formal representation of the document: the index representation or document representation.

Often, full text retrieval systems use a rather trivial algorithm to derive the index representations, for instance an algorithm that identifies words in an English text and puts them to lower case. The indexing process may include the actual storage of the document in the system, but often documents are only stored partly, for instance only title and abstract, plus information about the actual location of the document.

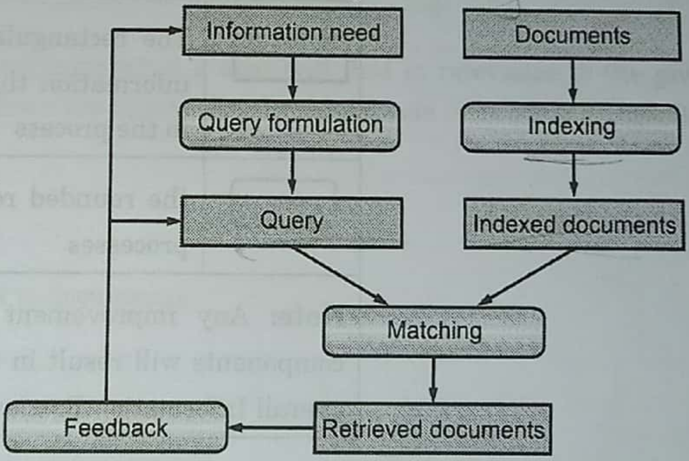


Fig. 1.3.1 : Information Retrieval (IR) Process

The query formulation process refers to the process of representing an information problem or need. The query is the resultant formal representation. In a broad sense, query formulation may refer to the complete interactive dialogue between system and user that leads not only to a suitable query but also to a better understanding of the user's information need by the user. However, in this thesis, query formulation refers to the automatic formulation of the query when no previously retrieved documents are available to guide the search, i.e. the formulation of the initial query.

The automatic formulation of successive queries is referred to as relevance feedback. The user and the system communicate the information need through queries and retrieved sets of documents, respectively. This is not the most natural way of communicating. Humans would use natural language to communicate information needs to one another. A request is a natural language statement of information need. Automatic query formulation takes in the request and generates an initial query.

In practise, this means that some or all of the words in the request are converted to query terms, such as by the rather simple algorithm that lowercases words. Relevance feedback uses a query or request as input and some previously retrieved relevant and non-relevant documents to generate a subsequent query. The matching process refers to the comparison of the query against the document representations. The matching process yields a prioritised list of relevant documents. Users will scroll down this document list looking for the information they require. Ranked retrieval should place relevant documents near the top of the ranked list, reducing the amount of time the user has to spend reading the documents.

The frequency distribution of terms across documents is used by simple but effective ranking algorithms. For example, the words "family" and "entertainment" mentioned in the first section appear relatively infrequently throughout the book, indicating that this book should not be read. The frequency distribution of terms across documents is used by simple but effective ranking algorithms. For example, the words "family" and "entertainment" mentioned in the first section appear relatively infrequently throughout the book, indicating that this book should not be read.

- With the help of the Fig.1.3.1, we can understand the process of information retrieval (IR) –

Notation	Represents
▭	The rectangular boxes represent the information that is supplied as input to the process
▢	the rounded rectangles represent the processes
Note: Any improvement in any of the above components will result in the improvement of the overall Information Retrieval system.	

- The user's **information need** is normally referred to as a **query**. The process of translating the information need into a query is called as **query formulation**.
- In its original form, a query consists of keywords, and the documents containing those keywords are searched for. A query may consist of a single word or a combination of words with multiple operations.

- **Indexing** happens in the back end without the direct involvement of the user and is responsible for the representation of the documents.
- The indexing process includes storing the document either partly or in some case the whole document. The index is always built before the searching begins and is a constant dynamic process.
- The query representation is further matched with the document representation that is stored in the index file and is referred to as the **matching process**. It results in a set of ordered documents based on the relevance and is referred to as the **ranked list**.

1.3.1 Classical Problem in Information Retrieval (IR) System

- The main goal of IR research is to develop a model for retrieving information from the repositories of documents. Here, we are going to discuss a classical problem, named **ad-hoc retrieval problem**, related to the IR system.
- In ad-hoc retrieval, the user must enter a query in natural language that describes the required information. Then the IR system will return the required documents related to the desired information. **For example**, suppose we are searching something on the Internet and it gives some exact pages that are relevant as per our requirement but there can be some non-relevant pages too. This is due to the ad-hoc retrieval problem.

Aspects of Ad-hoc Retrieval

Followings are some aspects of ad-hoc retrieval that are addressed in IR research –

- How users with the help of relevance feedback can improve original formulation of a query?
- How to implement database merging, i.e., how results from different text databases can be merged into one result set?
- How to handle partly corrupted data? Which models are appropriate for the same?

1.3.2 Information Retrieval (IR) Model

- Mathematically, models are used in many scientific areas having objective to understand some phenomenon in the real world.
- A model of information retrieval predicts and explains what a user will find in relevance to the given query. IR model is basically a pattern that defines the above-mentioned aspects of retrieval procedure and consists of the following -
  - A model for documents.
  - A model for queries.
  - A matching function that compares queries to documents.
- Mathematically, a retrieval model consists of -

  D – Representation for documents.

  R – Representation for queries.

  F – The modeling framework for D, Q along with relationship between them.

R (q,di) – A similarity function which orders the documents with respect to the query. It is also called ranking.

1.3.3 Components of Information Retrieval/ IR Model

GQ. What are the components of IR? (4 Marks)

(1) **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.

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

(3) **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.

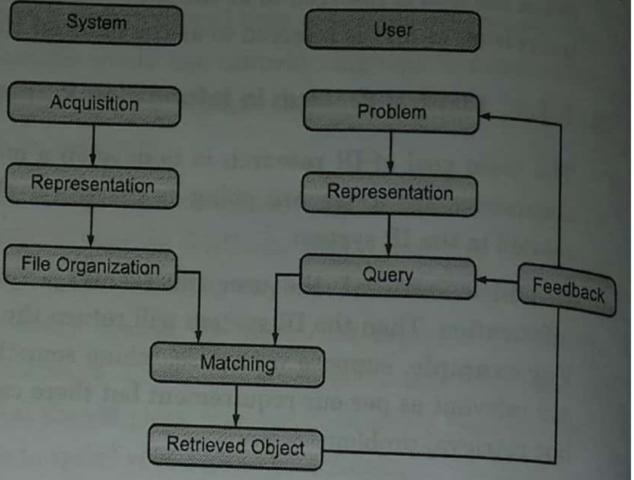


Fig. 1.3.2

(4) **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.

1.3.4 Difference between Information Retrieval and Data Retrieval

Sr. No.	Information Retrieval	Data Retrieval
1.	The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	Data retrieval deals with obtaining data from a database management system such as ODBMS. It is A process of identifying and retrieving the data from the database, based on the query provided by user or application.
2.	Retrieves information about a subject.	Determines the keywords in the user query and retrieves the data.
3.	Small errors are likely to go unnoticed.	A single error object means total failure.
4.	Not always well structured and is semantically ambiguous.	Has a well-defined structure and semantics.
5.	Does not provide a solution to the user of the database system.	Provides solutions to the user of the database system.

Sr. No.	Information Retrieval	Data Retrieval
6.	The results obtained are approximate matches.	The results obtained are exact matches.
7.	Results are ordered by relevance.	Results are unordered by relevance.
8.	It is a probabilistic model.	It is a deterministic model.

1.3.5 User Interaction with Information Retrieval System

1. The User Task

- The information first is supposed to be translated into a query by the user.
- In the information retrieval system, there is a set of words that convey the semantics of the information that is required whereas, in a data retrieval system, a query expression is used to convey the constraints which are satisfied by the objects.
- **Example :** A user wants to search for something but ends up searching with another thing.
- This means that the user is browsing and not searching. The above Fig.1.3.3 shows the interaction of the user through different tasks.

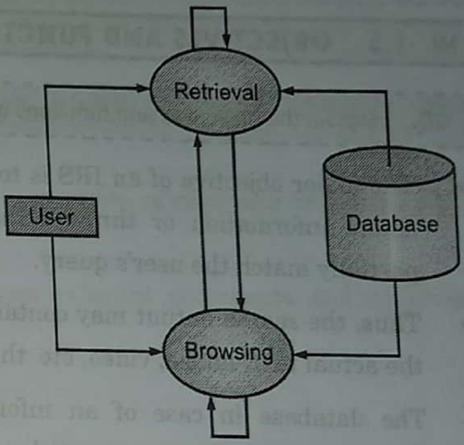


Fig. 1.3.3

2. Logical View of the Documents

- A long time ago, documents were represented through a set of index terms or keywords. Nowadays, modern computers represent documents by a full set of words which reduces the set of representative keywords.
- This can be done by eliminating stopwords i.e. articles and connectives.
- These operations are text operations. These text operations reduce the complexity of the document representation from full text to set of index terms.

1.4 PAST, PRESENT, AND FUTURE of INFORMATION RETRIEVAL

1. 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.

2. **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.

3. **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.

1.5 OBJECTIVES AND FUNCTIONS of IRS

GQ. Explain the objectives and functions of Information retrieval (IR) system in details. (4 Marks)

- The major objective of an IRS is to retrieve the required information whenever needed. It is either the actual information or through the documents containing the information surrogates that fully or partially match the user's query.
- Thus, the search output may contain bibliographic details of the documents that matches the query, or the actual text, image, video, etc. that contain the required information.
- The database in case of an information retrieval system may contain abstracts or full texts of documents, like newspaper articles, handbooks, dictionaries, encyclopedias, legal documents, statistics, etc., as well as audio, images, and video information.
- The major functions of an IRS are :
  (i) To identify the sources of information relevant to the areas of interest of the target users' community;
  (ii) To analyze the contents of the sources (documents);
  (iii) To represent the contents of the analyzed sources for matching with the users' queries;
  (iv) To match the search statement with the stored database;
  (v) To retrieve the information that is relevant
  (vi) To make necessary adjustments in the system based on feedback from the users.

1.6 ISSUES IN INFORMATION RETRIEVAL

The main issues of the Information Retrieval (IR) are Document and Query Indexing, Query Evaluation, and System Evaluation.

1. Expression of User's Information Need
2. Relevance
3. Indexing
4. Evaluation

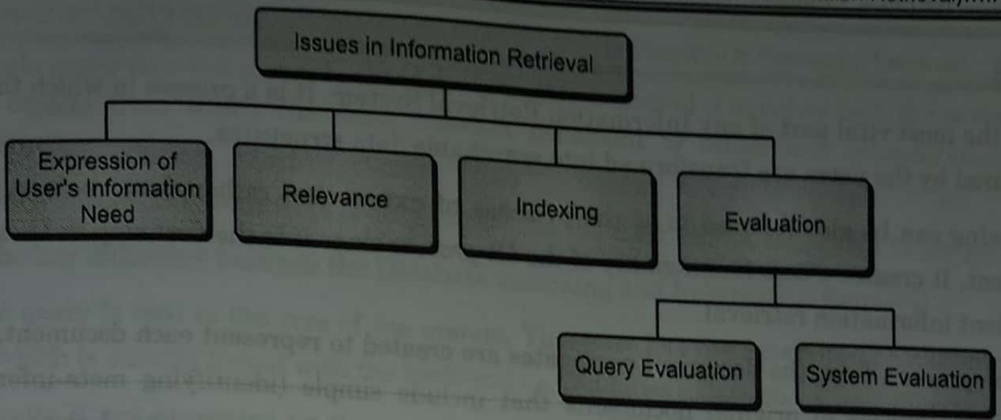

Issues in Information Retrieval
Expression of User's Information Need
Relevance
Indexing
Evaluation
Query Evaluation
System Evaluation

Fig. 1.6.1

GQ. List and explain the issues in information retrieval. (4 Marks)

1. Relevance

- The relevance refer to the retrieval of the information which could be text, audio, image or video from the information sources as requested by a user.
- The relevance of the retrieval results is user centric as the perspective of relevance varies from one user to other.
- Designing information retrieval algorithms to retrieve user relevant documents and achieving better retrieval effectiveness is a real challenge.

2. Expression of User's Information Need

- The expectation of the user posing a query could be to expect the information what he/she had in his/her mind. But the problem lies in whether the user expresses his/her needs correctly and precisely.
- An exact match of the user query to the document may not fetch the relevant documents.
- The terms used to express the user need in the form of query may not be present in the vocabulary/thesaurus/knowledge source and in literature this is reported as vocabulary mismatch problem or sparse data problem.
- Though the query given by a user is expanded using the vocabulary.
- The vocabulary must be updated to reflect the terms, phrases currently practiced/used by the user community.
- Another reason which reduces relevance is that most of the information retrieval systems ignore linguistic relevance and they fetch documents based on the statistical properties.
- Hence the design of information retrieval systems should take into consideration the linguistic features and user context to fetch more relevant documents /information even though the user query is expressed with less preciseness.

3. Indexing

- It is the most vital part of any Information Retrieval System. It is a process in which the docum required by the users are transformed into searchable data structures.
- Indexing can be also referred to as the **process of extraction** rather than analysis of partic content. It creates a core functionality of the IR process since it is the first step in IR and assis efficient information retrieval.
- In the process, first, the document surrogates are created to represent each document. Secondl requires analysis of original documents that include simple (identifying meta-information author, title, subject etc.) and complex (linguistic analysis of content) data. Indexes are the structures that are used to make the search faster.
- Main goal of **Document and Query Indexing** is to find important meanings and creating internal representation. The factors to be considered are accuracy to represent semant exhaustiveness, and facility for a computer to manipulate.

4. Evaluation

Evaluation in Information Retrieval is the process of systematically determining a subject's me worth, and significance by using certain criteria that are governed by a set of standards.

(i) **Query Evaluation :** In the retrieval model how can a document be represented with the select keywords and how are documents and query representations compared to calculate a sco Information Retrieval (IR) deals with issues like uncertainty and vagueness in informati systems.

- o **Uncertainty :** The available representation does not typically reflect true semantics of objec such as images, videos etc.
- o **Vagueness :** The information that the user requires lacks clarity, is only vaguely expressed a query, feedback or user action.

(ii) **System Evaluation :** System Evaluation tells about the importance of determining the impact information given on user achievement. Here, we see if the efficiency of the particular syster related to time and space.

⏭ 1.7 PROCESS of IR

GQ. Explain the process / architecture of IR. (4 Marks)

The working of information Retrieval process is explained below

- The Process of information retrieval starts when a user creates any query into the system through some graphical interface provided.
- These user-defined queries are the statements of needed information. For example, queries fork by users in search engines.

- In IR single query does not match to the right data object instead it matches with the several collections of data objects from which the most relevant document is taken into consideration for further evaluation.
- The ranking of relevant documents is done to find out the most related document to the given query.
- This is the key difference between the Database searching and Information Retrieval.
- After the query is sent to the core of the system. This part has the access to the content management module which is directly linked with the back-end i.e. the large collections of data objects.
- Once results R are generated by the core system then it is returned to the user by some graphical user interfaces.
- The process repeats and results are modified until the user satisfied for what he is actually looking for.
- The Following Fig. 1.7.1 sketches processing of textual queries performed by an Information Retrieval system.

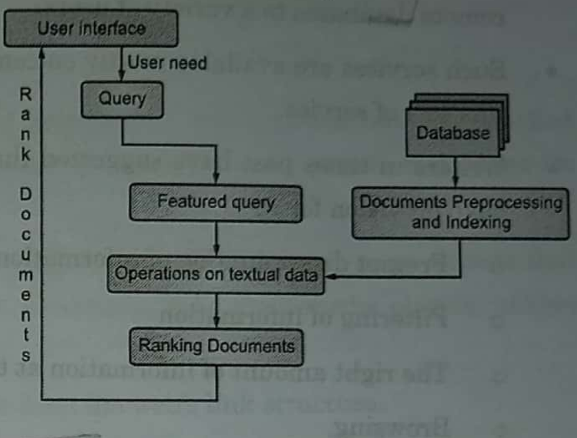


Fig. 1.7.1 : The process of information retrieval

1.8 INFORMATION RETRIEVAL IN THE LIBRARY, WEB AND DIGITAL LIBRARIES

GQ. Write a short note on: IR in library, digital libraries.

GQ. Discuss the impact of IR on the web. **(4 Marks)**

GQ. Discuss web information retrieval system. **(4 Marks)**

- For centuries libraries have been organizing reading materials on shelves for easy access. However, systematic methods that have been widely adopted for the organization of library materials and their recordings for use by readers came into being a little more than a century ago.
- Today's information professionals should know and be conversant with the traditional information retrieval tools and methods like classification, cataloguing, and vocabulary control as well as the traditional manual indexing systems.
- This is because these traditional methods show the process of evolution of information retrieval and most importantly, many recent developments in information retrieval in web and digital library environments have their roots in these traditional tools and methods.
- Different measures are currently taken for informing users about various materials accessible through a given digital or hybrid library.
- The information retrieval system serves as a bridge between the world of creators or generation of information and the users of that information.
- Two broad categories of information retrieval have been identified :
  - In-house Information retrieval

- ○ Online Information retrieval

- In-house Information retrieval systems are set up by a particular library or information centre to serv mainly the users within the organization. An example of an in-house database is the library catalogue.
- Online public access catalogue (OPAC) provides facilities for library users to carry out online catalogu searches, and then check the availability of the item required.
- By online information retrieval systems, we mean those that have been designed to provide access t remote databases to a variety of users.
- Such services are available mostly on commercial basis, and there are a number of vendors that handl this sort of service.
- Writers in times past have suggested that an effective and reliable information retrieval system mus have provision for :
  - ○ Prompt dissemination of information
  - ○ Filtering of information
  - ○ The right amount of information at the right time
  - ○ Browsing,
  - ○ Getting information in an economical way
  - ○ Current literature
  - ○ Interpersonal communication and
  - ○ Personal help.
- In a typical library environment there are two categories of users, the library and information personnel and the end users.
- Library and information personnel often act as intermediaries and they may also act as end users seeking for information for their own use or for decision making.
- All information retrieval systems should be user oriented. As such the interest of every user should be given due importance at every point of information storage and retrieval.

Web IR

- Web IR refers to the application of IR to the internet.
- Users in traditional IR specify queries in some query language that represent their information needs.
- The system selects and displays the set of documents in its collection that appear to be the most relevant to the query.
- Users can then refine their queries to get a better answer.
- In the web environment, user intents are not as static and consistent as they are in traditional IR. The information need on the web is associated with a given task that is unknown in advance and may differ significantly from user to user, even if the query specification is the same.

- The task identification and mental process of deriving a query from an information need are critical aspects of web IR.
- Web IR is related to web mining, which is the automated discovery of useful and interesting information on the internet.
- Web mining is currently being developed in three major research directions based on the type of data being mined:
  - web content mining
  - web structure mining
  - web usage mining.
- Another type of data, document change, page age, and information regency, has recently sparked research interest: it has a temporal dimension and allows for the analysis of the growth and dynamics of the Web over time.
  - **Web content mining** is the process of extracting useful information from web page content that can be found in a variety of formats, including text, metadata, links, multimedia objects, hidden and dynamic pages, and semantic data.
  - **Web structure mining** attempts to infer knowledge from the web's link structure.
  - **Web usage mining** attempts to investigate web user behavior by analyzing data generated by user interaction and automatically recorded in web server logs. Web usage mining applications typically aim to learn user profiles or navigation patterns. The goal of web usage mining is to predict the next user request based on the analysis of previous requests.
  - IR is the automatic retrieval of all relevant documents while retrieving as few non-relevant documents as possible. Some argue that Web resource or document discovery (IR) is an example of Web content mining, while others associate web mining with intelligent IR. In reality, the primary goals of IR are to index text and search for useful documents in a collection, and current IR research includes modelling, document classification and categorization, user interfaces, data visualization, filtering, and so on. Web document classification or categorization, which could be used for indexing, can be considered an example of Web mining. In this sense, Web mining is a component of the (Web) IR process.

Chapter Ends...

❑❑❑