Module I

CHAPTER 1

Introduction

Syllabus

Motivation, Basic Concepts, The Retrieval Process, Information System: Components, parts and types on information system; Definition and objectives on information retrieval system, Information versus Data Retrieval. Search Engines and browsers.

Self-learning Topics : Search Engines , Search API

1.1 INTRODUCTION

GQ. Define information retrieval.

- Information is data that has been structured for easier understanding or interpretation. Information retrieval is the process of getting data that has been processed, is not in its raw form, and satisfies your specific requirements.
- Aa an example of Information retrieval system, web search engines, are used to find the relevant documents or web pages.
- Information Retrieval (IR) can be defined as a system that deals with the organization, storage, retrieval, evaluation and access of information.
- An Information retrieval (IR) system is an Information system, a system used to store items of information that need to be processed, searched, retrieved, and distributed to various users
  - (1) The organization and representation of the information should be helpful for the users to provide easy access to information and satisfy his information need.
  - (2) User summarizes his information need in the form of a query using

set of keywords or index terms.

(3) IR system processes this query and returns useful or relevant information to the user.

(4) The main purpose of IR system is to provide a user easy access to documents containing the desired information

(5) Information Retrieval System is a system it is a capable of storing, maintaining from a system. and retrieving of information. This information may be in any of the form i.e. audio, video, text.

(6) Information Retrieval System is mainly focus electronic searching and retrieving of documents.

(7) The first Information retrieval systems originated with the need to organize information in central repositories e.g. libraries.

1.2 OBJECTIVES OF INFORMATION RETRIEVAL SYSTEM

GQ. Discuss the objectives of information retrieval systems?

GQ. List and explain components of IR block diagram.

(1) The objective of an information retrieval system is to enable users to find relevant information from an organized collection of documents in response to the user query.

(2) To provide information to the user in least time with least efforts.

(3) To act as facilitator between information and user.

(4) To provide non-ambiguous search results through proper indexing.

(5) User friendliness.

Fig. 1.2.1 Shows basics of Information Retrieval

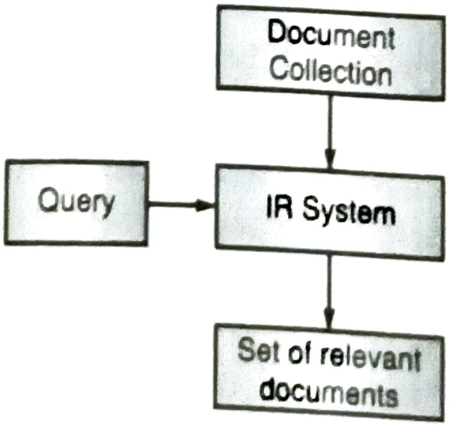

Fig. 1.2.1 : Basics of Information Retrieval

(I) Growth of Information Retrieval

GQ. Discuss growth of information retrieval.
GQ. List out reasons behind success of web.
GQ. Explain retrieval from web with the help of diagram.

(1) Initially primary goal of information retrieval was indexing and searching for useful documents in collection.

(2) Information retrieval has grown at a very large scale in past 20 years because of rapid growth of world wide web (WWW)

(3) Web is growing at very fast pace and becoming a universal repository of human knowledge and culture.

(4) The reasons behind success of web are :

(i) Standard user interface hiding all implementation details from user

(ii) Any user can create his own Web documents and make them point to any other Web documents without restrictions, which in turn making web a new publishing medium accessible to everyone.

(iii) Applications like online shopping and internet banking making users' life easy and also generating revenues.

(5) Due to its growing size and success, the Web has introduced new problem. Searching useful information on the web is difficult and time-consuming task.

(6) Absence of a well-defined underlying data model for the Web makes navigation task difficult.

(7) These difficulties have attracted new interest in IR and have created a place for IR at the center of stage. Fig. 1.2.2 shows information retrieval from Web.

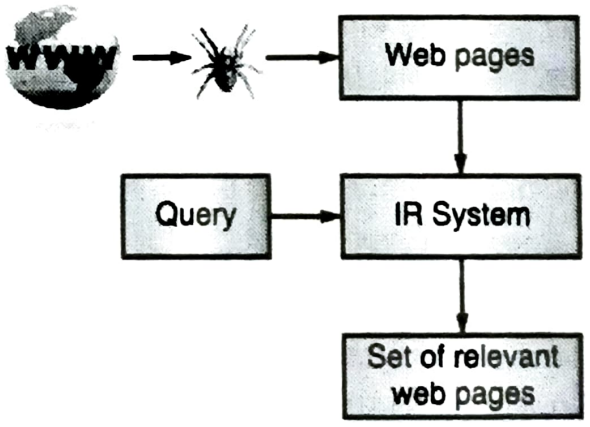

Fig. 1.2.2 : Information Retrieval from Web

(II) Features of an information retrieval system

GQ. List out features of IRS.

An effective information retrieval system must have provisions for :

(1) Prompt dissemination of information
(2) Filtering of information
(3) The right amount of information at the right time
(4) Active switching of information
(5) Receiving information in an economical way
(6) Browsing
(7) Getting information in an economical way
(8) Current literature
(9) Access to other information systems
(10) Interpersonal communications, and
(11) Personalized help.

(III) Basic Concepts

The user task and the logical view of the documents are two important factors to have a direct impact on the effective retrieval of relevant information.

The User Task

GQ. Explain user interaction with IR with the help of a diagram.

- A retrieval system's user must convert his information requirement into a query in the system's given language.
- User provides a set of keywords that accurately express the semantics of the information requirement when using an information retrieval system
- Now imagine a user who has a poorly defined or naturally broad area of interest and looking for information by using an interactive interface to simply look around in the collection for documents related to his requirement.
- While doing this, user might find many documents related to different topics one after another and he continue to browse the documents in the collection rather than searching information on specific topic.

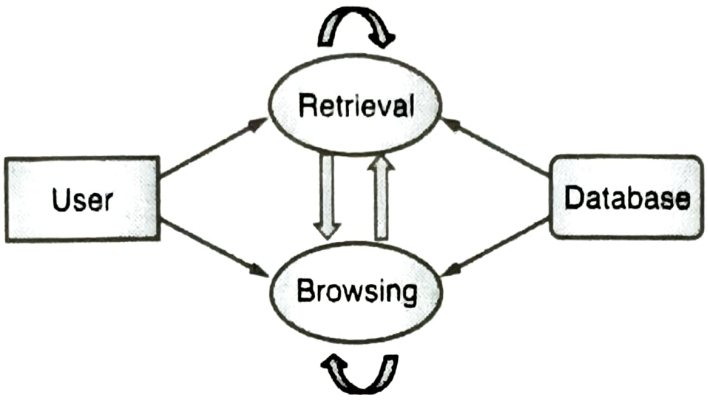

Fig. 1.2.3 : User interaction with Retrieval System Using different tasks

Logical View of Document

GQ. Explain logical view of a document.

- A set of index terms or keywords are typically used to describe the documents in a collection. Such keywords may be chosen by a human subject or may be directly extracted from the document's text
- These representative keywords offer a logical picture of the document whether they are generated manually or automatically.
- It is now possible to represent a document using its entire word-set on modern computers. We refer to this situation as a full text logical view (or representation) of the documents by the retrieval system.
- Even modern computers, nevertheless, might need to narrow the range of representative terms in very big collections.
- Eliminating stopwords, such as articles and connectives, using stemming to reduce different words to their grammatical roots, and identifying noun groups are all ways to achieve this (which eliminates adjectives, adverbs, and verbs). Additionally, compression might be used.
- These operations are referred as text operations (or transformations).
- Text operations make document representation less complicated and enable changing the logical view from one of a full text to one of a set of index words.

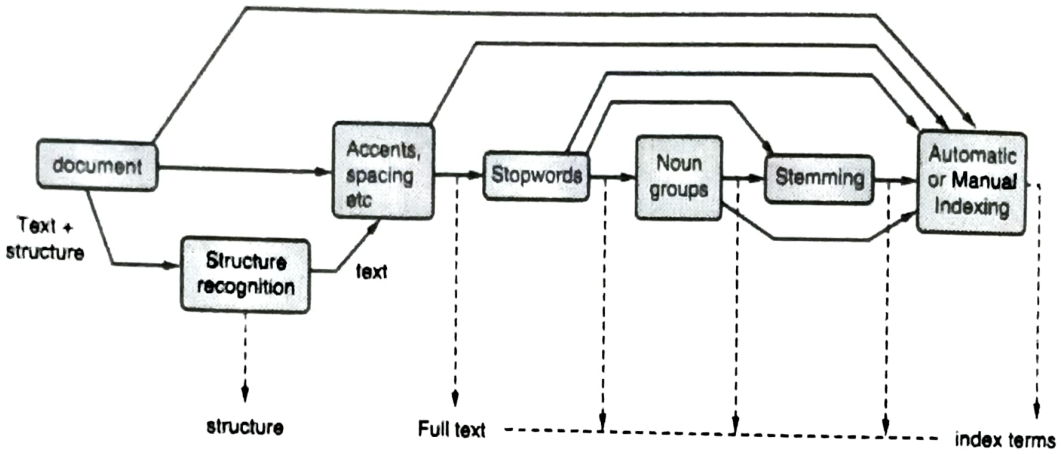

Fig. 1.2.4 : Logical view of a document : from full text to a set of index terms

(IV) The Retrieval Process

GQ. Illustrate the concepts of IRS with architecture view ?

The following Fig. 1.2.5 shows Information Retrieval Process.

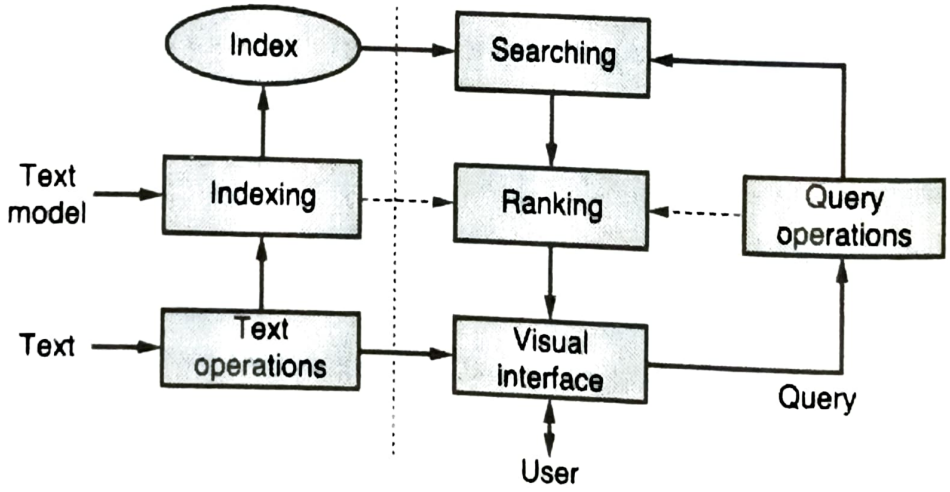

Fig. 1.2.5 : Information Retrieval Process

1.3 INFORMATION RETRIEVAL PROCESS

GQ. Illustrate the concepts of IRS with architecture view?

(1) First of all, the text database must be defined before the retrieval process can even be started. The database manager typically handles this, and they state the following :

(i) The documents to be used,

(ii) The text operations that will be carried out, and

(iii) The text model (i.e., the text structure and what elements can be retrieved).

(2) The text operations alter the source documents and produce a logical view of them. The database manager creates a text index after defining the logical view of the documents using the DB Manager Module.

(3) The retrieval process can start now that the document database is indexed. Once the user specifies information requirement, the same text operations are used to parse and alter the text.

(4) Then, before to generating the actual query that gives a system representation for the user requirement, query operations are applied.

(5) The query is then processed to obtain the *retrieved documents*. The retrieved documents are ranked according to a likelihood of relevance before sending them to user.

(6) The user then looks over the collection of ranked documents in an effort to find relevant information.

(7) At this stage, he might identify a subset of the documents considered to be unquestionably interesting and start a feedback cycle from users.

(8) In such a cycle, the system modifies the query formulation based on the documents the user has chosen. Hopefully, this query has been changed to better reflect the actual user requirement.

1.3.1 Requirements for Information Retrieval

GQ. List out major requirements of IR.

(1) An automated or manually-operated indexing system used to index and search techniques and procedures.

(2) A collection of documents in any one of the following formats: text, image or multimedia.

(3) A set of queries that serve as the input to a system, via a human or machine.

(4) An evaluation metric to measure or evaluate a system's effectiveness.

1.3.2 Three Major Components of Traditional IRS

GQ. Discuss 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

☞ Components of Information Retrieval/ IR Model

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.

(1) Sequential : It contains documents by document data.

(2) 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.

(1) Information System : Components and Types

GQ. What is information system. Discuss its components.

- Information systems are collections of multiple information resources (e.g., software, hardware, computer system connections, the system housing, system users, and computer system information) to gather, process, store, and disseminate information.
- IT is a group of data sets that ensures that business operates smoothly, embracing change, and helping companies achieve their goal.
- Tools such as laptops, databases, networks, and smartphones are examples of information systems.
- Information systems consist of members that gather, store, and process data, with the data being utilized to give information, add to knowledge and create digital products that aid decision-making.
- IT has sets of technological methods and techniques used to store, organize, manage, and retrieve information digitally.

(2) Components of Information Systems

(i) **Resources of people :** (end users and IS specialists, system analyst, programmers, data administrators etc.).

(ii) **Hardware :** (Physical computer equipment and associate device, machines and media).

(iii) **Software :** (programs and procedures).

(iv) **Data :** (data and knowledge bases).

(v) **Networks :** (communications media and network support).

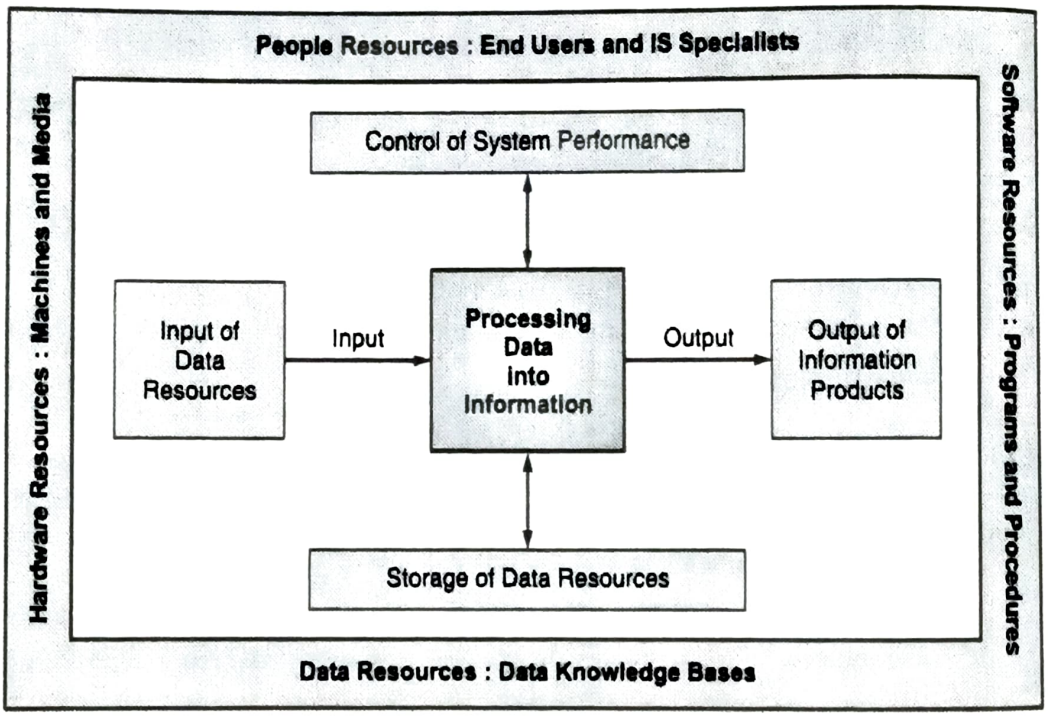

Fig. 1.3.1 : Components of Information System

☞ Types of Information System

GQ. Explain various types of Information Systems.

(1) Operations support system
(2) Transaction Processing System (TPS)
(3) Management Information System (MIS)
(4) Decision Support System (DSS)
(5) Experts System
(6) Office Automation System (OAS)

▶ (1) Operations support system

- In an organization, data input is done by the end user which is processed to generate information products i.e. reports, which are utilized by internal and or external users. Such a system is called operation support system.
- The purpose of the operation support system is to facilitate business transaction, control production, support internal as well as external communication and update organization central database.

- The operation support system is further divided into :
  (i) Transaction-processing system,
  (ii) Processing control system and
  (iii) Enterprise collaboration system.

(2) Transaction Processing System (TPS)

- In manufacturing organization, there are several types of transaction across department. Typical organizational departments are Sales, Account, Finance, Plant, Engineering, Human Resource and Marketing.
- Across which following transaction may occur sales order, sales return, cash receipts, credit sales; credit slips, material accounting, inventory management, depreciation accounting, etc.
- These transactions can be categorized into
  (i) Batch transaction processing,
  (ii) Single transaction processing and
  (iii) Seal time transaction processing.

(3) Management Information System (MIS)

- Management Information System is designed to take relatively raw data available through a Transaction Processing System and convert them into a summarized and aggregated form for the manager, usually in a report format.
- It reports tending to be used by middle management and operational supervisors.
- Many different types of report are produced in MIS. Some of the reports are a summary report, on-demand report, ad-hoc reports and an exception report.
- **Example :** Sales management systems, Human resource management system.

(4) Decision Support System (DSS)

- Decision Support System is an interactive information system that provides information, models and data manipulation tools to help in making the decision in a semi-structured and unstructured situation.

- Decision Support System comprises tools and techniques to help in gathering relevant information and analyze the options and alternatives, the end user is more involved in creating DSS than an MIS.
- **Example :** Financial planning systems, Bank loan management systems.

(5) Experts System

- Experts systems include expertise in order to aid managers in diagnosing problems or in problem-solving. These systems are based on the principles of artificial intelligence research.
- Experts Systems is a knowledge-based information system. It uses its knowledge about a specify are to act as an expert consultant to users.
- Knowledgebase and software modules are the components of an expert system. These modules perform inference on the knowledge and offer answers to a user's question

(6) Office Automation System (OAS)

- OAS consists of computers, communication-related technology, and the personnel assigned to perform the official tasks.
- The OAS covers office transactions and supports official activity at every level in the organization.
- The official activities are subdivided into managerial and clerical activities.

1.3.3 Difference between Information Retrieval and Data Retrieval

GQ. Explain Information versus Data Retrieval in detail.

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.

Sr. No.	Information Retrieval	Data Retrieval
(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.
(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.

1.4 SEARCH ENGINES AND WEB BROWSERS

GQ. Differentiate between Search Engines and browsers.
GQ. What is search engine?

(1) Searching for information on the Web is, for most people, a daily activity. Search and communication are by far the most popular uses of the computer.

(2) A search engine is the practical application of information retrieval techniques to large-scale text collections.

(3) A web search engine is used to find information on the World Wide Web and returns web pages that are accessible online, displaying the results at one place. To retrieve and see information from web pages stored on web servers, web browsers use search engines.

(4) A search engine's primary purpose is to collect and maintain information about several URLs and Web browsers are designed to display the website at the server's current URL.

(5) A web browser uses a graphical user interface to help users have an interactive online session on the Internet.

1.4.1 Search Engine Working

GQ. Explain working of search engine.

- The Search Indexer, Crawler, and Database are the three essential components of a search engine.
- Boolean operators AND, OR, and NOT are used by search engines to limit and widen the results of a search.
- The steps that the search engine takes are as follows :
  (1) Instead of searching for the phrase directly on the web, the search engine first looks for it in the index for predefined databases.
  (2) The information is then found using software to search the database. This software component is known as web crawler.
  (3) Once web crawler finds the pages, the search engine then shows the relevant web pages as a result. These retrieved web pages often include the page title, the amount of material on the page, the first few phrases, etc.
  (4) User can click on any of the search results to open it.

1.4.2 Building Blocks of Search Engine

Major Functions of search engines components are the indexing process and the query process.

GQ. Discuss indexing process with the help of diagram

(1) Indexing Process

- The indexing process builds the structures that enable searching, and the query process uses those structures and a person's query to produce a ranked list of documents. Major components of indexing process are shown in Fig. 1.4.1.
- Text acquisition component is used to identify and make available the documents that will be searched. Text acquisition will require

building a collection by crawling or scanning the Web, a corporate intranet, a desktop, or other sources of information.

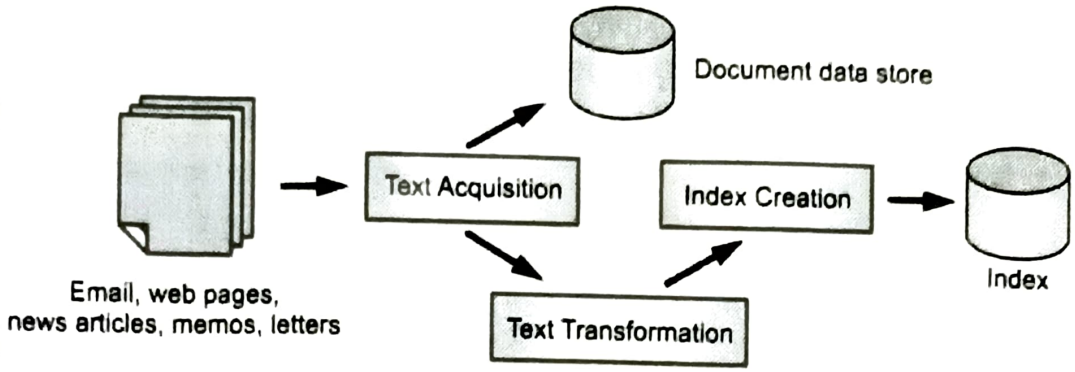

Fig. 1.4.1 : Components of Indexing Process

- The text acquisition component also creates a document data store, which contains the text and metadata for all the documents.
- The text transformation component transforms document into index terms or features.
- The index creation component takes the output of the text transformation component and creates the indexes or data structures that enable fast searching.

2) Query Process

GQ. Explain query process with the help of a diagram.

- Query process uses structures created by indexing process and a person's query to produce a ranked list of documents. Fig. 1.4.2 shows the building blocks of the query process.

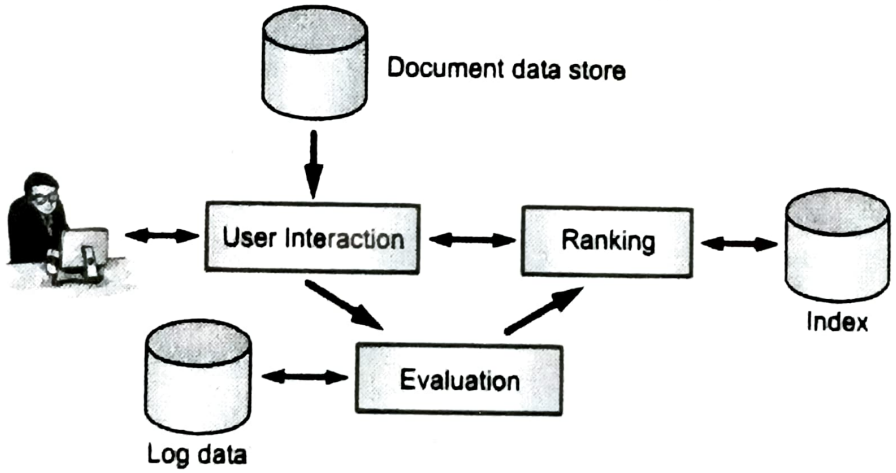

Fig. 1.4.2 : Components of Indexing Process

- The major components are user interaction, ranking, and evaluation as shown Fig. 1.4.2.
- The user interaction component provides the interface between the person doing the searching and the search engine. It accepts the user's query and transforms query into index terms. It also takes the ranked list of documents from the search engine and organizes it into the results shown to the user.
- The document data store is one of the sources of information used in generating the results.
- The ranking component takes the transformed query from the user interaction component and generates a ranked list of documents using scores based on a retrieval model.
- The efficiency of ranking depends on the indexes, and the effectiveness depends on the retrieval model.
- Evaluation component measures and monitors effectiveness and efficiency. It also records and analyzes user behavior using log data.
- The results of evaluation are used to tune and improve the ranking component.

Short Questions and Answers

Q. 1 Define information retrieval.

Ans. :

Information Retrieval is finding material of an unstructured nature that satisfies an information need from within large collections

Q. 2 List and explain components of IR block diagram.

Ans. :

(1) Input – Store Only a representation of the document

(2) A document representative – Could be list of extracted words considered to be significant.

(3) Processor – Involve in performance of actual retrieval function

(4) Feedback – Improve

(5) Output – A set document numbers.

Q. 3 Explain the type of natural language technology used in information retrieval.

Ans. :

Two types

(1) Natural language interface makes the task of communicating with the information source easier, allowing a system to respond to a range of inputs.

(2) Natural Language text processing allows a system to scan the source texts, either to retrieve particular information or to derive knowledge structures that may be used in accessing information from the texts.

Q. 4 What is search engine?

Ans. :

A search engine is a document retrieval system design to help find information stored in a computer system, such as on the WWW. The search engine allows one to ask for content meeting specific criteria and retrieves a list of items that match those criteria

Q. 5 What are the applications of IR?

Ans. :

(1) Indexing (2) Ranked retrieval

(3) Web search (4) Query processing

Q. 6 How to AI applied in IR systems?

Ans. :

Four main roles investigated

(1) Information characterization

(2) Search formulation in information seeking

(3) System Integration

(4) Support functions

Q. 7 Give the functions of information retrieval system.

Ans. :

(1) To identify the information(sources) relevant to the areas of interest of the target users community

(2) To analyze the contents of the sources(documents)

(3) To represent the contents of the analyzed sources in a way that will be suitable for matching user's queries

(4) To analyze user's queries and to represent them in a form that will be suitable for matching with the database

(5) To match the search statement with the stored database

(6) To retrieve the information that is relevant

(7) To make necessary adjustments in the system based on feedback form the users.

Q. 8 List the issues in information retrieval system.

Ans. :

(1) Assisting the user in clarifying and analyzing the problem and determining information needs.

(2) Knowing how people use and process information.

(3) Assembling a package of information that enables group the user to come closer to a solution of his problem.

(4) Knowledge representation.

(5) Procedures for processing knowledge/information.

(6) The human-computer interface.

(7) Designing integrated workbench systems

(8) Designing user-enhanced information systems.

(9) System evaluation.

Q. 9 Define relevance.

Ans. :

- Relevance appears to be a subjective quality, unique between the individual and a given document supporting the assumption that relevance can only be judged by the information user.
- Subjectivity and fluidity make it difficult to use as measuring tool for system performance.

Q. 10 Define indexing & document indexing.

Ans. :

- Association of descriptors (keywords, concepts, metadata) to documents

in view of future retrieval. Document indexing is the process of associating or tagging documents with different "search" terms.

- Assign to each document (respectively query) a descriptor represented with a set of features, usually weighted keywords, derived from the document (respectively query) content.

Q. 11 Discuss the impact of IR on the web.

Ans. :

The impacts of information retrieval on the web are influenced in the following areas.

(1) Web Document Collection

(2) Search Engine Optimization

(3) Variants of Keyword Stuffing

(4) DNS cloaking: Switch IP address

(5) Size of the Web

(6) Sampling URLs

(7) Random Queries and Searches

Q. 12 Define web search and web search engine.

Ans. :

- Web search is often not informational -- it might be navigational (give me the url of the site I want to reach) or transactional (show me sites where I can perform a certain transaction, e.g. shop, download a file, or find a map).
- Web search engines crawl the Web, downloading and indexing pages in order to allow full-text search.
- There are many general purpose search engines; unfortunately, none of them come close to indexing the entire Web.
- There are also thousands of specialized search services that index specific content or specific sites.

Q. 13 What are the components of search engine?

Ans. : Generally, there are three basic components of a search engine as listed below :

(1) Web Crawler

(2) Database

(3) Search Interfaces

Q. 14 What are search engine processes?

Ans. :

Indexing Process

(1) Text acquisition

(2) Text transformation

(3) Index creation

Query Process

(1) User interaction

(2) Ranking

(3) Evaluation

Q. 15 What are the challenges of web?

Ans. :

(1) Distributed data

(2) Volatile data

(3) Large volume

(4) Unstructured and redundant data

(5) Data quality

(6) Heterogeneous data

Chapter Ends...

❑❑❑