

Unit- I

Introduction to Information Retrieval Systems

- 1.1 Definition of Information Retrieval System
- 1.2 Objectives of Information Retrieval Systems
- 1.3 Functional Overview
- 1.4 Relationship to Database Management Systems
- 1.5 Digital Libraries and Data Warehouses
- 1.6 Information Retrieval Systems Capabilities
 - 1.6.1 Querying
 - 1.6.2 Browsing
 - 1.6.3 Miscellaneous capabilities

1.1 Definition of Information Retrieval System:

- IRS stands for Information Retrieval Systems
- An IR System is a system capable of storage, retrieval, and maintenance of information.
- An IRS System facilitates user in find the information that user need

Types of Searches in IRS:-

- Web Search
- Desktop Search
- Library search

Information: text, image, audio, video, and other multimedia objects Focus on textual information here. An IR system facilitates a user in find the information the user needs.

• Item(Data):

The smallest complete textual unit processed and manipulated by an IR system Depend on how a specific source treats information

• Success measure (Objectives of an IR System) :

Minimize the overhead for finding information

• Overhead:

The time a user spends in all of the steps leading to reading an item containing needed information, excluding the time for actually reading the relevant data

- Query generation
 - Search composition
 - Search execution
 - Scanning results of query to select items to read
- An Information Retrieval System consists of a software program that facilitates a user in finding the information the user needs.
 - The system may use standard computer hardware or specialized hardware to support the search sub function and to convert non-textual sources to a searchable media

- (e.g., transcription of audio to text).

Query Generation:

The user entering the keyword like Eg :Sachin Tendulkar, Best Restaurants,.....etc
It will be providing relevant information to the user

Search Composition:

If the given text or query, It will be available or not finding the user required Information

Search Execution:

Automatically start Execution procedure for relevant Information

Scanning Results of Query for Reading Item:

The user Required Information starts scanning Results of Query, Required Information Generated Infront of User's Screen

- An Information Retrieval system consists of a software Program that facilitates a user in the information the user needs
- The system may use standard computer hardware or specialized hardware to support the search sub Function to convert Non-Textual sources to searchable Media

1.2 Objectives of Information Retrieval Systems

The general objective of an IR system is ,

- To minimize the overhead of a user locating needed information
 - The Main objective of an Information Retrieval system is to reduce the overhead of a user locating required Information
- Over head is Expressed as the time a user spends in all of the steps leading to reading an item containing the needed Information
- The success of an IRS is how well it can Minimize the user
- The overhead for a user to find the needed information
- Overhead from users Perspective is the time required to find the Information
- Thus, search Composition, Search Execution & reading Non-relevant Items are all aspects of IR Overhead

Relevant Item:

In IRS the term "Relevant" Item is used to Represent an Item containing the needed Information

Ex: JPG, bmp

From a user Perspective "Relevant "& Needed

Measures Used:

The two major Measures commonly Associated with Information Systems are

- Precision
- Recall

Precision:

The ability to retrieve Top ranked Documents that are Mostly Relevant

Ex: Key Exactly Matched Ranked Work

Recall:

The ability of the search to find all of the relevant items

Ex: Search computer it will shows as hard Disk , Keyboard, Monitor

Precision: $\frac{\text{Number-Retrieved-Relevant}}{\text{Number-Total-Retrieved}}$

Recall: $\frac{\text{Number-Retrieved-Relevant}}{\text{Number-Possible-Retrieved}}$

Enter a Documents effect of search on Total Document Space

- The two major measures commonly associated with information systems are “precision” and “recall”
- Support of user search generation
- How to present the search results in a format that facilitate the user in determining relevant items

The two major measures commonly associated with information systems are precision and recall. When a user decides to issue a search looking for information on a topic, the total database is logically divided into four segments shown in Figure 1.1. Relevant items are those documents that contain information that helps the searcher in answering his question. Non-relevant items are those

items that do not provide any directly useful information. There are two possibilities with respect to each item: it can be retrieved or not retrieved by the user’s query. Precision and recall are defined as:

Figure 1.1 Effects of Search on Total Document Space

$$\text{Precision} = \frac{\text{Number_Retrieved_Relevant}}{\text{Number_Total_Retrieved}}$$

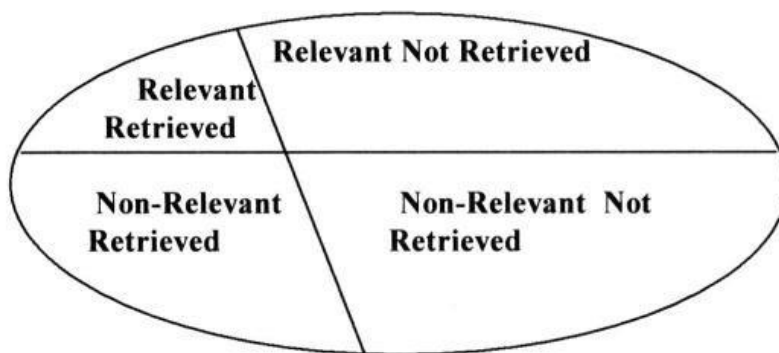


Figure 1.1 Effects of Search on Total Document Space

$$\text{Recall} = \frac{\text{Number_Retrieved_Relevant}}{\text{Number_Possible_Relevant}}$$

Searching an Item:

- For a search looking into an a Topic, the total Database is Logically divided 4 Segments
- Where *Number_Possible_Relevant* are the number of relevant items in the database.
 - *Number_Total_Retrieved* is the total number of items retrieved from the query.
 - *Number_Retrieved_Relevant* is the number of items retrieved that are relevant to the user’s search need.

- Precision Measures one Aspect of Information retrieved overhead for a user Associated with a particular Search

Two More Objectives of IR Systems :

- Support of user search generation How to specify the information a user needs
- Language ambiguities – “field”
- Vocabulary corpus of a user and item authors Must assist users automatically and through interaction in developing a search specification that represents the need of users and the writing style of diverse authors
- How to present the search results in a format that facilitate the user in determining relevant items ,
 - A) Ranking in order of potential relevance
 - B) Item clustering and link analysis.

1.3 Functional Overview:

A total Information Storage and Retrieval System is composed of four major functional processes:

- Item normalization,
- Selective dissemination of information (i.e., “mail”),
- Archival document database search, and
- An index database search along with the
- Automatic file build process that supports index files.

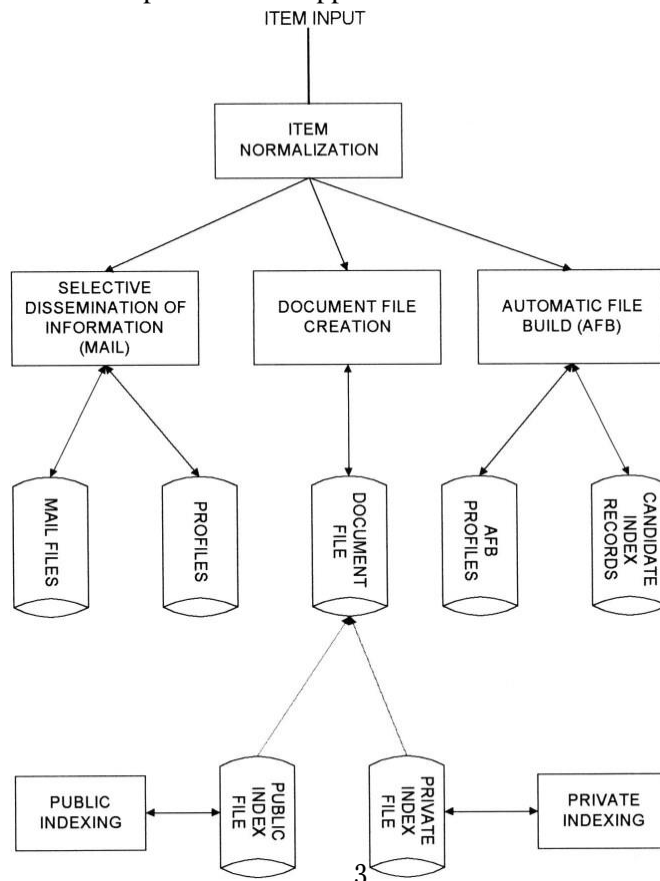


Figure 1.4 Total Information Retrieval System

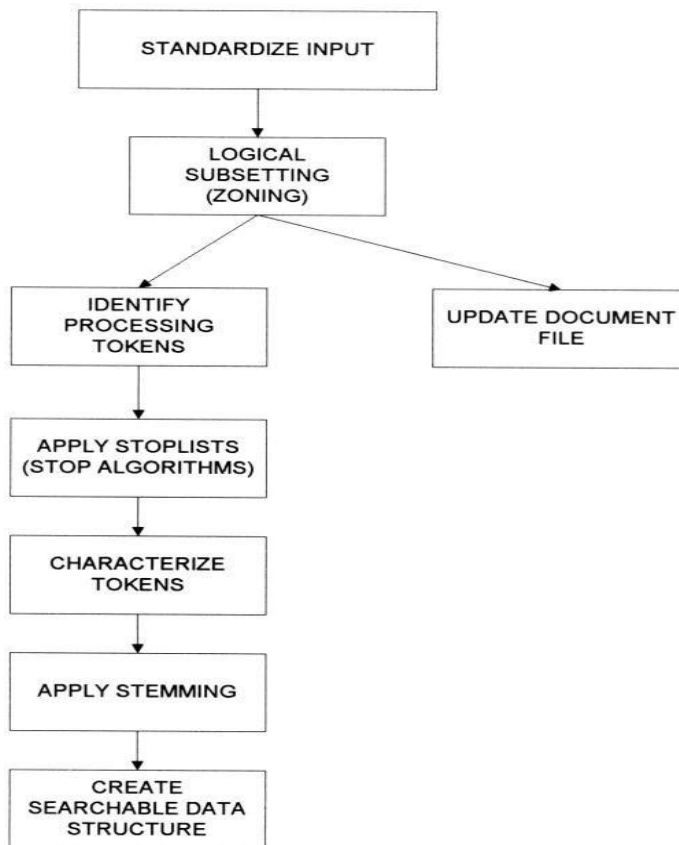


Figure 1.5 The Text Normalization Process

1.3.1 Item Normalization:

- Normalize incoming items to a standard format
 - Language encoding
 - Different file formats...
- Logical restructuring – zoning
- Create a searchable data structure (Indexing)
 - Identification of processing tokens
 - Characterization of the tokens – single words, or phrase
 - Stemming of the tokens

Standardize Input:

- Standardizing the input takes the different external format of input data and performs the translation to the formats acceptable to the system.
- Translate foreign language into Unicode Allow a single browser to display the languages and potentially a single search system to search them
- Translate multi-media input into a standard format
Video: MPEG-2, MPEG-1, AVI, Real Video...
Audio: WAV, Real Audio
Image: GIF, JPEG, BMP...

Logical Subsetting (Zoning) :

- Parse the item into logical sub-divisions that have meaning to user Title, Author, Abstract, Main Text, Conclusion, References, Country, Keyword...
- Visible to the user and used to increase the precision of a search and optimize the display The zoning information is passed to the processing token identification operation to store the information, allowing searches to be restricted to a specific zone display the minimum data required from each item to allow determination of the possible relevance of that item (Display zones such as Title, Abstract...)

Identify Processing Tokens :

- Identify the information that are used in the search process – *Processing Tokens (Better than Words)*
- The first step is to determine a word
Dividing input symbols into three classes
- Valid word symbols: alphabetic characters, numbers
- Inter-word symbols: blanks, periods, semicolons (non-searchable)
- Special processing symbols: hyphen (-)
A word is defined as a contiguous set of word symbols bounded by inter-word symbols.

Stop Algorithm:

- Save system resources by eliminating from the set of searchable processing tokens those have little value to the search Whose frequency and/or semantic use make them of no use as searchable token
- Any word found in almost every item
- Any word only found once or twice in the database
Frequency * Rank = Constant
Stop algorithm v.s. Stop list

Characterize Tokens :

- Identify any specific word characteristics Word sense disambiguation Part of speech tagging
- Uppercase – proper names, acronyms, and organization Numbers and dates

Stemming Algorithm :

- Normalize the token to a standard semantic representation Computer, Compute, Computers, Computing
 - Comput
- Reduce the number of unique words the system has to contain
 - ex: “computable”, “computation”, “computability”
 - small database saves 32 percent of storages
 - larger database : 1.6 MB □ 20 % 50 MB □ 13.5%
- Improve the efficiency of the IR System and to improve recall -> Decline precision

Create Searchable Data Structure:

- Processing tokens -> Stemming Algorithm -> update to the Searchable data structure
- Internal representation (not visible to user)
Signature file, Inverted list, PAT Tree...
- Contains
 - Semantic concepts represent the items in database
 - Limit what a user can find as a result of the search

Functional Overview – Selective Dissemination of Information :

- Provides the capability to dynamically compare newly received items in the information system against standing statements of interest of users and deliver the item to those users whose statement of interest matches the contents of the items
- Consist of ,
 - Search process
 - User statements of interest (Profile)
 - User mail file
- A profile contains a typically broad search statement along with a list of user mail files that will receive the document if the search statement in the profile is satisfied
As each item is received, it is processed against every user’s profile When the search statement is satisfied, the item is placed in the mail file(s) associated with the process
User search profiles are different than ad hoc queries in that they contain significant more search terms and cover a wider range of interests .

Document Database Search :

- Provides the capability for a query to search against all items received by the system
Composed of the search process, user entered queries and document database.
- Document database contains all items that have been received, processed and store by the system. Usually items in the Document DB do not change May be partitioned by time and allow for archiving by the Time partitions.
- Queries differ from profiles in that they are typically short and focused on a specific area of interest .

Index Database Search:

- When an item is determined to be of interest, a user may want to save it (file it) for future reference Accomplished via the index process.
- In the index process, the user can logically store an item in a file along with additional index terms and descriptive text the user wants to associate with the item. An index can reference the original item, or contain substantive information on the original item Similar to card catalog in a library.
- The Index Database Search Process provides the capability to create indexes and search them
- The user may search the index and retrieve the index and/or the document it references
- The system also provides the capability to search the index and then search the items referenced by the index records that satisfied the index portion of the query
Combined file search
- In an ideal system the index record could reference portions of items versus the total item
- Two classes of index files: public and private index files Every user can have one or more private index files leading to a very large number of files, and each private index file references only a small subset of the total number of items in the Document database Public index files are maintained by professional library services personnel and typically index every item in the Document database
- The capability to create private and public index files is frequently implemented via a structured Database Management System (RDBMS)
- To assist the users in generating indexes, the system provides a process called Automatic File Build (Information Extraction)
 - Process selected incoming documents and automatically determines potential indexing for the item
 - Authors, date of publication, source, and references

The rules that govern which documents are processed for extraction of index information and the index term extraction process are stored in Automatic File Build Profiles. When an item is processed it results in creation of Candidate Index Records -> for review and edit by a user

Prior to actual update of an index file.

1.4 Relationship to Database Management Systems :

There are two major categories of systems available to process items:

Information Retrieval Systems and Data Base Management Systems (DBMS).

1. An Information Retrieval System is software that has the features and functions required to manipulate “information” items versus a DBMS that is optimized to handle “structured” data.
2. Structured data is well defined data (facts) typically represented by tables. There is a semantic description associated with each attribute within a table that well defines that attribute. For example, there is no confusion between the meaning of “employee name” or “employee salary” and what values to enter in a specific database record. On the other hand, if two different people generate an abstract for the same item, they can be different. One abstract may generally discuss the most important topic in an item. Another abstract, using a different vocabulary, may specify the details of many topics. It is this diversity and ambiguity of language.
3. With structured data a user enters a specific request and the results returned provide the user with the desired information. The results are frequently tabulated and presented in a report format for ease of use. In contrast, a search of “information” items has a high probability of not finding all the items a user is looking for. The user has to refine his search to locate additional items of interest. This process is called “iterative search.
4. From a practical standpoint, the integration of DBMS’s and Information Retrieval Systems is very important. Commercial database companies have already integrated the two types of systems. One of the first commercial databases to integrate the two systems into a single view is the INQUIRE DBMS.

1.5 Digital Libraries and Data Warehouses :

Two other systems frequently described in the context of information retrieval are,

- Digital Libraries and
- Data Warehouses

There is a significant overlap between these two systems and an Information Storage and Retrieval System.

All these systems are repositories of information and their primary goal is to “satisfy user information needs”

Digital Library:

A Digital Library enables users to Interact effectively with Information distributed across a network

These network Information systems support search & Display of Items from organized collections

- ❖ As such, libraries have always been concerned with storing and retrieving information in the media it is created on.
- ❖ As the quantities of information grew exponentially, libraries were forced to make maximum use of electronic tools to facilitate the storage and retrieval process. With the worldwide Internet of libraries and information sources (e.g., publishers, news agencies,....etc)via Internet, more focus has been on the concept of an electric library

List of Softwares For Digital Libraries

- KOHA
- BIBLIOTEQ
- PMP

- Indexing is one of the critical disciplines in library science and significant effort has gone into the Establishment of Indexing and cataloging Standards
 - Migration of many of the library products to a digital format Introduces both opportunities and challenges the full text of items available for search makes the index process
 - Another important library service is a source of search Intermediaries to assist users in finding Information

Information storage and Retrieval Technology has addressed a small subset of the issue associated with Digital Libraries the focus has been on the search and retrieval of Textual data with no concern for establishing standards on the contents of the system.

DATAWAREHOUSES:

A Data warehouse is a type of Data Management System that is designed to enable and support Business Intelligence Activities, Especially Analytics, Data warehouses are solely Intended to perform queries and Analysis and often contain Large amounts of Historical Data.

List of Softwares For DATAWAREHOUSES

- Amazon Red shift
- Microsoft Azure
- Google Big query
- Snowflake
- A Data warehouse is Relational Database that is designed for query and analysis rather than transaction processing It includes historical data derived from transaction data from single &Multiple sources
- A Data warehouse is a group of Data specific to the entire organization, not only to particular group of users
 - It is not used for daily operations and transaction processing but used for making decisions.

1.6 Information Retrieval Systems Capabilities :

The capabilities in the information retrieval systems are,

- Querying
- Browsing
- Miscellaneous capabilities

1.6.1 Querying:

Communicate a description of the needed information to the system.

Main paradigms:

- Query term sets
- Query terms connected with Boolean operations
- Weighted terms
- Relaxation or restriction of term matching
- Term expansion
- Natural language

Query Term Sets :

Describe the information needed by specifying a set of query terms.

- System retrieves all documents that contain *at least one* of the query terms.
- Documents are ranked by the number of terms they include :
 - Documents containing all query terms appear first
 - Documents containing all query terms but one appear second
 - Documents containing only one query term appear last

Boolean Queries :

Describe the information needed by relating multiple terms with Boolean operators.

- **Operators** : AND, OR, NOT (sometimes XOR).
- Corresponding **set operations** : intersection, union, difference. Operate on the sets of documents that contain the query terms.
- **Precedence** : NOT, AND, OR; use parentheses to override; process left-to- right among operators with same precedence.

Example: This example uses standard operator precedence (Note: the combination **AND NOT** is usually abbreviated **NOT**)

COMPUTER OR SEVER AND NOT MAINFRAME

Select all documents that discuss computers, or documents that discuss servers that do not discuss mainframes.

(COMPUTER OR SERVER) AND NOT MAINFRAME

Select all documents that discuss computers or servers, do not select any documents that discuss mainframes.

COMPUTER AND NOT (SERVER OR MAINFRAME)

Select all documents that discuss computers, and do not discuss either servers or mainframes.

Weighting:

A Weight is associated with each term

Weighted Term:

Term weighting is a procedure that takes place during the text indexing process in order to assess the value of each term to the document, Term weighting is the assignment of numerical values to terms that represent their importance in a document in order to improve retrieval effectiveness

Natural Language :

Natural language, understood as a tool that people use to express themselves, has specific properties that reduce the efficacy of textual information retrieval systems. These properties are linguistic variation and ambiguity. By linguistic variation we mean the possibility of using different words or expressions to communicate the same idea. Linguistic ambiguity is when a word or phrase allows for more than one interpretation.

1.6.2 Browsing :

Browsing can be defined as an interactive search activity in which the direction of the search is determined by the user on the basis of immediate feedback from the system being browsed. Most users of most information retrieval systems exhibit browsing behavior no matter what the underlying system structure.

Determine the retrieved documents that are of interest

- The query phase ends, and the browse phase begins, with a summary display of the result. Summary displays use either
 - Line item status
 - Data visualization
- Powerful browsing capabilities are particularly important when precision is low.

1.6.3 Miscellaneous Capabilities:

- **Vocabulary browse**
 - Vocabulary browse provides the capability to display in Alphabetical sorted order words from the document Database
 - Logically all unique words (Processing Tokens) in the Database are kept in sorted order along with the count of the number of unique items in which the word is found
 - It makes the search Procedure Easier
 - It corrects entered the word “computen” when they really Meant”computer”
- **Iterative search(query refinement)**
 - The result of a previous search is subjected to a new query
 - Same as repeating the previous query with additional conditions.
- **Canned(stored) queries**
 - Users tend to reuse previous queries
 - Allows users to store previously-used queries and incorporate
 - Canned queries tend to be large.