Unit- I

Introduction to Information Retrieval Systems

1.1 Definition of Information Retrieval System :

An IR System is a system capable of storage, retrieval, and maintenance of information.
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:
  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).

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 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

$$\text{Precision} = \frac{Number\_\mathrm{Re}\,trieved\_\mathrm{Re}\,levant}{Number\_Total\_\mathrm{Re}\,trieved}$$

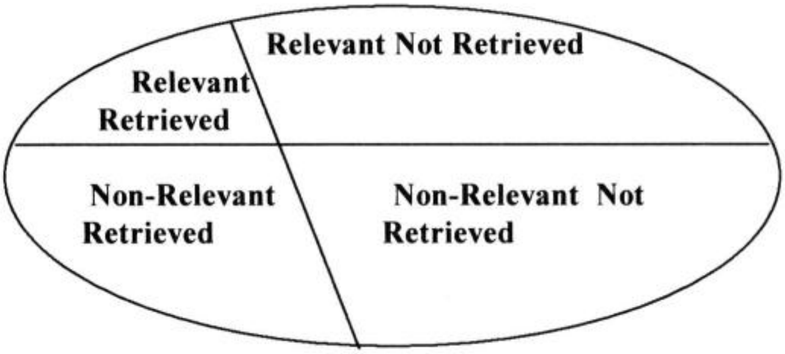

Figure 1.1 Effects of Search on Total Document Space

$$\textbf{Recall} = \frac{Number\_\mathrm{Re}\,trieved\_\mathrm{Re}\,levant}{Number\_Possible\_\mathrm{Re}\,levant}$$

Where *Number_Possible_Relevant* are the number of relevant items in the database. *Number_Total_Retieved* 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.

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.

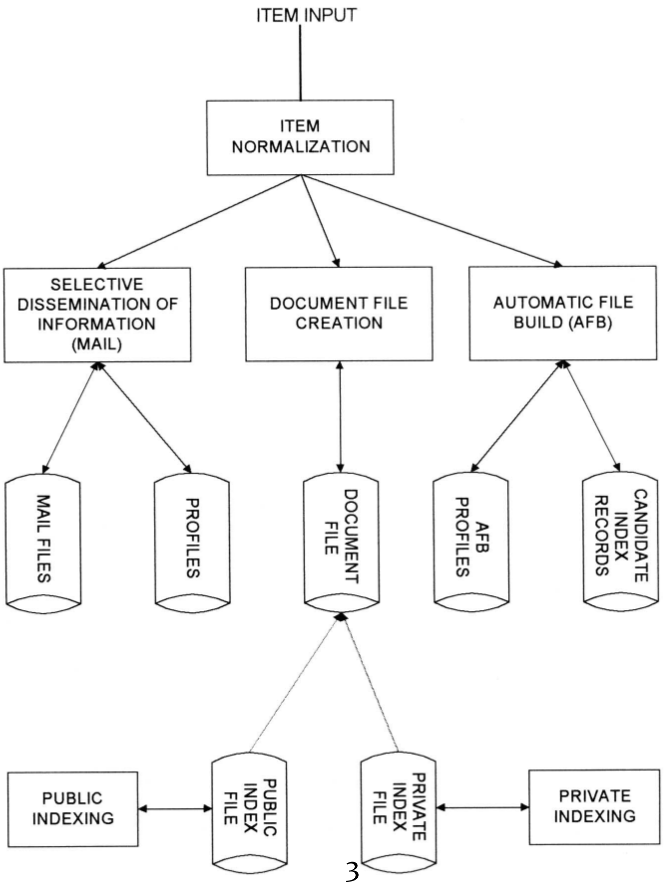

Figure 1.4 Total Information Retrieval System

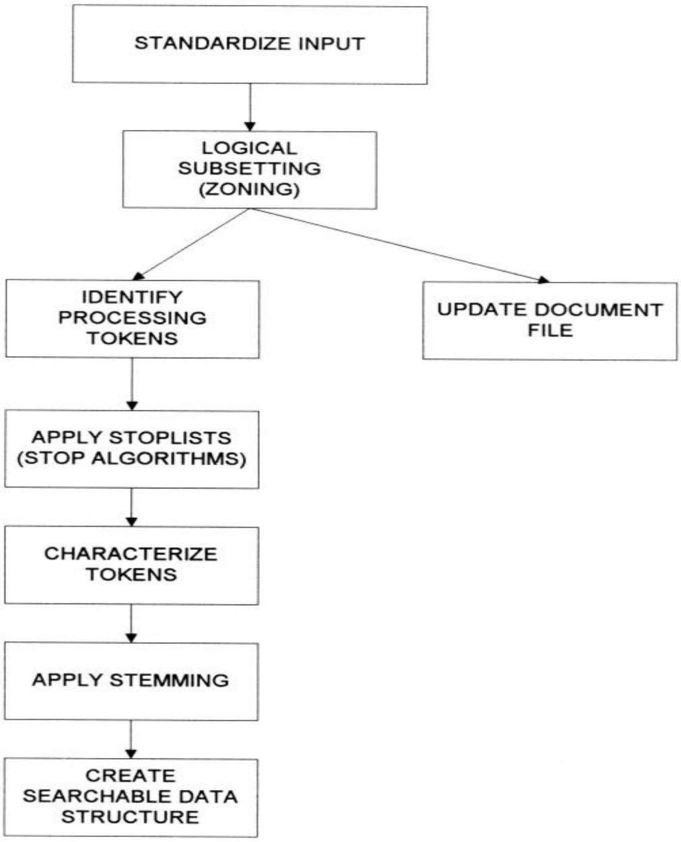
STANDARDIZE INPUT
LOGICAL SUBSETTING (ZONING)
IDENTIFY PROCESSING TOKENS
UPDATE DOCUMENT FILE
APPLY STOPLISTS (STOP ALGORITHMS)
CHARACTERIZE TOKENS
APPLY STEMMING
CREATE SEARCHABLE DATA STRUCTURE

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

1.3.1.1 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…

1.3.1.2 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…)

1.3.1.3 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 (nonsearchable)
• Special processing symbols: hyphen (-)

A word is defined as a contiguous set of word symbols bounded by inter-word symbols.

1.3.1.4 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

1.3.1.5 Characterize Tokens :

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

1.3.1.6 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

1.3.1.7 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

1.3.2 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 .

1.3.3 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 .

1.3.4 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 (or Data Marts).

- ❖ There is a significant overlap between these two systems and an Information Storage and
Retrieval System. All three systems are repositories of information and their primary goal is to "satisfy user information needs"

- ❖ 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 interneting of libraries and information sources (e.g., publishers, news agencies, wire services, radio broadcasts) via the Internet, more focus has been on the concept of an electronic library.

- 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 issues 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.

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.
- **M-form-N**: Find any document containing N of the terms T1,…,TM. May be expressed as a Boolean query
- **Weighting** : A weight is associated with each term.

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.

Proximity Constraints :

Restrict the distance within a documents between two search terms.

- Proximity specifications limit the acceptable occurrences and hence increase the precision of the search.
- Important for large documents.
- **General Format**: *TERM1* within m units of *TERM2*
- *UNIT* may be character, word, paragraph, etc.
- **Direction** operator: specify which term should appear first.
- **Adjacent** operator: m = 1 in forward direction.

Example:

VENETIAN ADJ BLIND

Find documents that discuss “Venetian Blinds” but not “Blind Venetians”.

UNITED WITHIN 5 WORDS OF AMERICAN

Find documents that discuss “United Airlines and American Airlines” but not “United States of America and the American dream”.

NUCLEAR WITHIN 0 PARAGRAPHS OF CLEANUP.

Find documents that discuss “nuclear” and “cleanup” in the same paragraph.

Contiguous Word Phrase Matches :

Treat a sequence of N words as a single semantic unit.

Example: “United States of America”.

CWP is N-ary (not Boolean) operator.
Cannot be expressed as Boolean query.

If only two are specified, then CWP reduces to the adjacent operator
(or the proximity operator with m = 1 in forward direction).

Also called “literal string” or “exact phrase” matching.

Fuzzy (Approximate) matching :

Match terms that are similar to the query term.

- ❖ Fuzzy matching compensates for spelling errors, especially when documents were scanned-in and then subjected to optical character recognition (OCR).
- ❖ Increased recall (more documents qualify because new terms may be matched) at the expense of deceased precision (erroneous matches may introduce nonrelevant documents).

Example: COMPUTER may match COMPITER, CONPUTER, etc.

- o Usually, should not match if the closely-spelled word is legitimate in itself (e.g., COMMUTER. This would help maintain precision.
- o Rules needed to indicate allowed differences (e.g., one character replacement, or one transposition of adjacent characters).

- Similar method may be used to overcome *phonetic* spelling errors.Should be distinguished from *fuzzy set theory* solutions.

Term masking:

Match terms that contain the query term.

- ❖ **Single position mask**: accept any term that will match the query term, once the character in a certain position is disregarded.
  - • **Example**: the term MULTI$NATIONAL will be matched by “multi-national” or “multinational” (but not by “multi national” since it is a sequence of two terms!)
- ❖ **Variable length mask**: accept any term that will match the query term, once a sequence of any number of characters in a certain position is disregarded.
  - • **Suffix**: *WARE will match terms that end with “ware”.
  - • **Prefix**: WARE* will match terms that begin with “ware”. The most common mask (sometimesapplied by default).
  - • **Imbedded**: *WARE* will match terms that contain “ware

Number and data Ranges :

Match numeric or date terms that are in the range of the query term.

- ❖ **Numeric** query terms: >125 (matches all numbers greater that 125) or 125-425 (matched all numbers between 125 and 425).
- ❖ **Date** query terms: 9/1/97 - 8/31/98 (matches all dates between 1 September 1997 and 31 August 1998).
- ❖ In a way, term-masking handles “string ranges”.

Term Expansion :

Expand/restrict the query terms via thesauri or concept hierarchies/networks.

- ❖ **Concept hierarchy**: A hierarchy(tree) of concepts.

  • Replacing a query term(e.g. BOOK) by an ancestor(more general) term (e.g.,PUBLICATION) increases recall and decreases precision.

  • Replacing a query term by a descendant (more specific)term (e.g. PAPERBACK)decreases recall and increase precision.

- ❖ **Concept network:** Terms are related by associations.
  - • Often,associations are specific to the database(the context).
  - • Example:CONCERT is generalized by PERFORMANCE and associated with TICKET
- ❖ Semantic thesaurus:Groups together terms that are similar in meaning (a single level concept hierarchy). A query term is matched by every term in its thesaurus group.
  - – Must avoid expanding with “synonyms” that change the meaning.Like (in the sense of “akin”) might be expanded with “prefer”.

– Expansion may introduce terms not found in the document database.
– Thesauri and concept networks should be expandable by users.

- ❖ Statistical thesaurus:Groups together terms that are statistically related(occur together in the same documents).
  – Terms in a class may have no shared meaning.
  – Specific to a given document database.
  – Must be updated when the document database is updated.

Natural Language :

Describe the information needed in natural language prose.

– **Example**: Find all the documents that discuss oil reserves and current attempts to find oil reserves. Include any documents that discuss the international financial aspects of the old production process. Do not include documents about the oil industry in the United States.
– **Pseudo NL processing**: System scans the prose and extracts recognized terms and Boolean connectors. The grammaticality of the text is not important.
– **Problem**: Recognize the negation in the search statement(“Do not include…”)
– **Compromise**: Use enter natural language sentences connected with Boolean operators.

1.6.2 Browsing :

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.

Item summary

o Typically,each retrieved document is displayed in one status line, and as many documents are displayed as can fit in the screen.

o The status line may contain the relevance factor (if computed), the title,and possibly some other zones.

o Documents may be displayed in more than one line (less documents per screen).

Summary order

- o **Boolean systems:** All retrieved documents equally meet the query criteria. Documents are displayed in arbitrary,or in sorted order(alphabetically by title or chronologically by date).
- o **Relevance:** In systems that compute relevance, retrieved documents

are sorted by relevance. Usually relevance is normalized to a range 0-1 .0-100. A threshold value defines the documents that are not relevant.

1.6.3 Miscellaneous Capabilities:

- **Vocabulary browse**
    - Users enter a term and are positioned in an alphabetically-sorted list of all the terms that appear in the database.
    - With each term the number of documents in which it appears is shown.
    - Assists users who are not familiar with the vocabulary.
    - Help users determine the impact of using individual terms
- **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.
- **Relevance feedback**
    - The old query is replaced by a new query
    - The new query is a transformation of the old query, reflecting feedback about the relevance of the documents retrieved by the first query
- Canned(stored) queries
    - Users tend to reuse previous queries
    - Allows users to store previously-used queries and incorporate
    - Canned queries tend to be large.