NARSIMHA REDDY ENGINEERING COLLEGE

An Autonomous Institution| Affiliated to JNTUH | Approved by AICTE

Accredited by NBA & NAAC with 'A' Grade

your roots to success. Estd.2007

Program Name	**: B.Tech - CSE**
Name Of The Course	**: IRS**
Course Code	**: 23CS511**
Year & Semester	**: III B.Tech. I Sem**
Faculty Name	**: Mr. R.ManojKumar**

UNIT-I

Introduction to Information retrieval Systems

Contents

- **Introduction**
- Definition of Information Retrieval System
- Objectives of Information Retrieval System
- Functional Overview
- Relationship to RDBMS
- Digital Libraries and Data Warehouses
- Summary

Introduction

- Here we will define and differentiate the differences between Information Retrieval and DBMS.

- The importance of differences lies in the inability of a DBMS to provide the functions needed to process "information".

- An information system containing structured data also suffers major functional deficiencies.

Contents

- Introduction
- **Definition of Information Retrieval System**
- Objectives of Information Retrieval System
- Functional Overview
- Relationship to RDBMS
- Digital Libraries and Data Warehouses
- Summary

Definition of IRS

Def:

- An information Retrieval System is a system that is capable of storage, retrieval and maintenance of information.
  - Information here can be composed of text (including numeric and date data), images, audio, video and other multimedia objects.

- An information Retrieval System consists of a software program that facilitates a user in finding the information the user needs.

Definition of IRS

- The gauge of success of an information system is how well it can minimize the overhead for a user to find the needed information

  - **Overhead:** It is the time required to find the information needed, excluding the time for actually reading the relevant data.
  - **Aspects:** Search composition, search execution and reading non-relevant items

Contents

- Introduction
- Definition of Information Retrieval System
- **Objectives of Information Retrieval System**
- Functional Overview
- Relationship to RDBMS
- Digital Libraries and Data Warehouses
- Summary

Objectives of IRS

- **Objective:**

  To minimize the overhead of a user locating needed information.

  - **Overhead** is the time a user spends in all of the steps leading to reading an item containing the needed information. Eg: Query generation, query execution etc.

- The success of an information system is very subjective, based upon what information is needed and the willingness of user to accept overhead.

Objectives of IRS

- Measures:
  - Precision
  - Recall

$$\text{Precision} = \frac{\text{Number\_Retrieved\_Relevant}}{\text{Number\_Total\_Retrieved}}$$

$$\text{Recall} = \frac{\text{Number\_Retrieved\_Relevant}}{\text{Number\_Possible\_Relevant}}$$

Objectives of IRS

Where

Number_Possible_Relevant is 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

Objectives of IRS

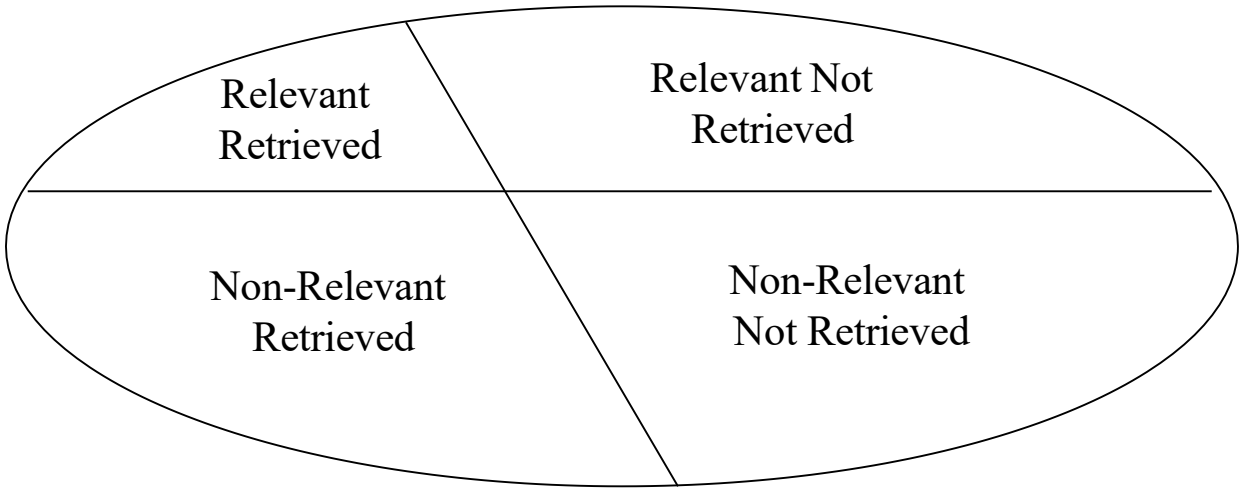


Figure1: Effects of Search on Total Document Space

Objectives of IRS

- Precision measures one aspect of information retrieval overhead for a user associated with a particular search.

- Recall gauges how well a system processing a particular query is able to retrieve the relevant items that the user is interested in seeing.

- Recall is a very useful concept, but due to the denominator (in formula on recall) is non-calculable in operational systems

Objectives of IRS

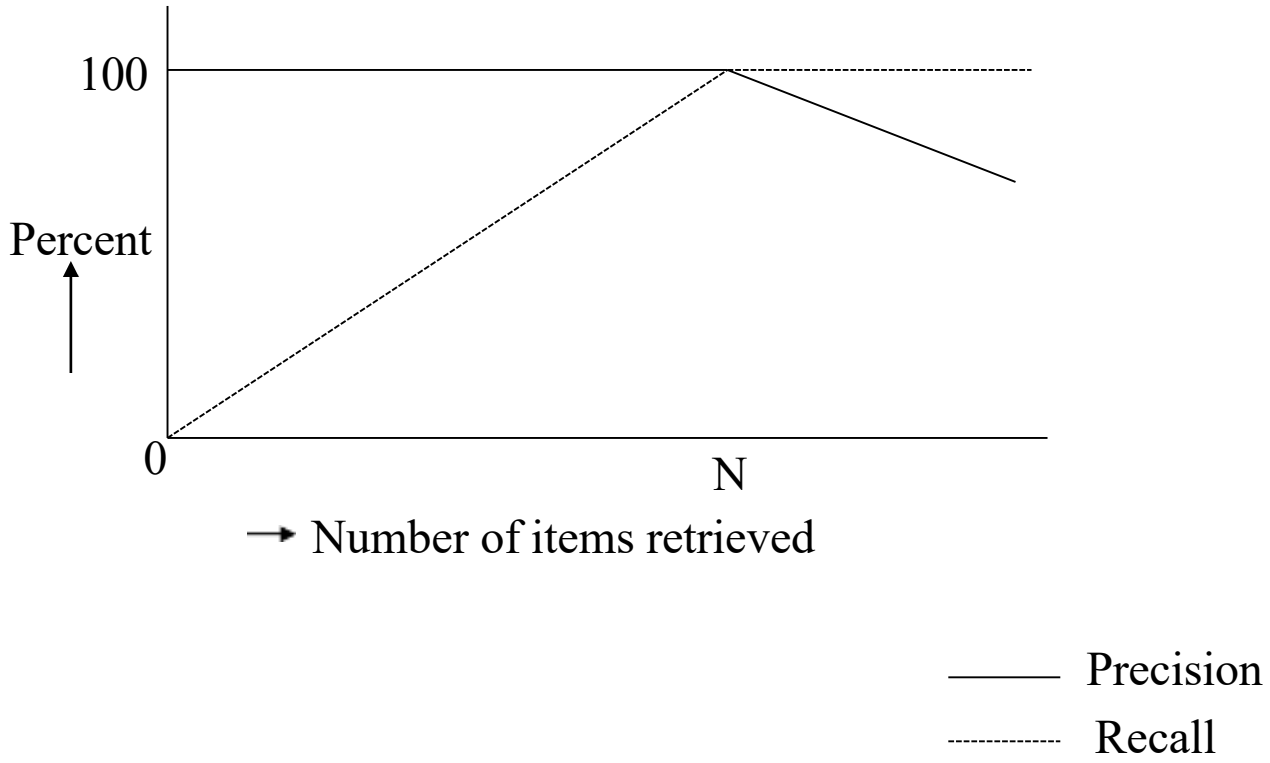


Fig: Ideal Precision and Recall

Objectives of IRS

- Precision starts off at 100 percent and maintains that value as long as relevant items are retrieved.
  - Precision is directly affected by retrieval of non-relevant items and drops close to zero.
- Recall starts off close to zero and increases as long as relevant items are retrieved.
  - Recall is not effected by retrieval of non-relevant items and hence remains at 100 percent.

Objectives of IRS

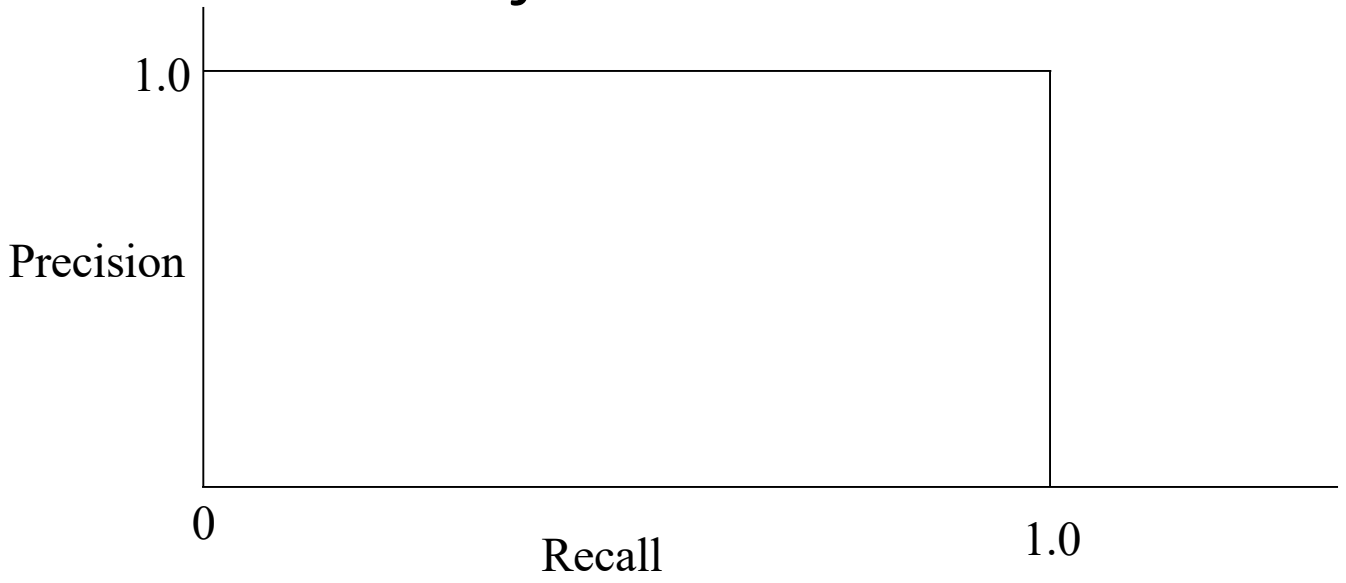


Fig: Ideal Precision/Recall graph

Here every item retrieved is relevant. Thus Precision stays at 100 percent(1.0). Recall continues to increase by moving to the right on the X-axis until it also reaches to 100 percent.

Objectives of IRS

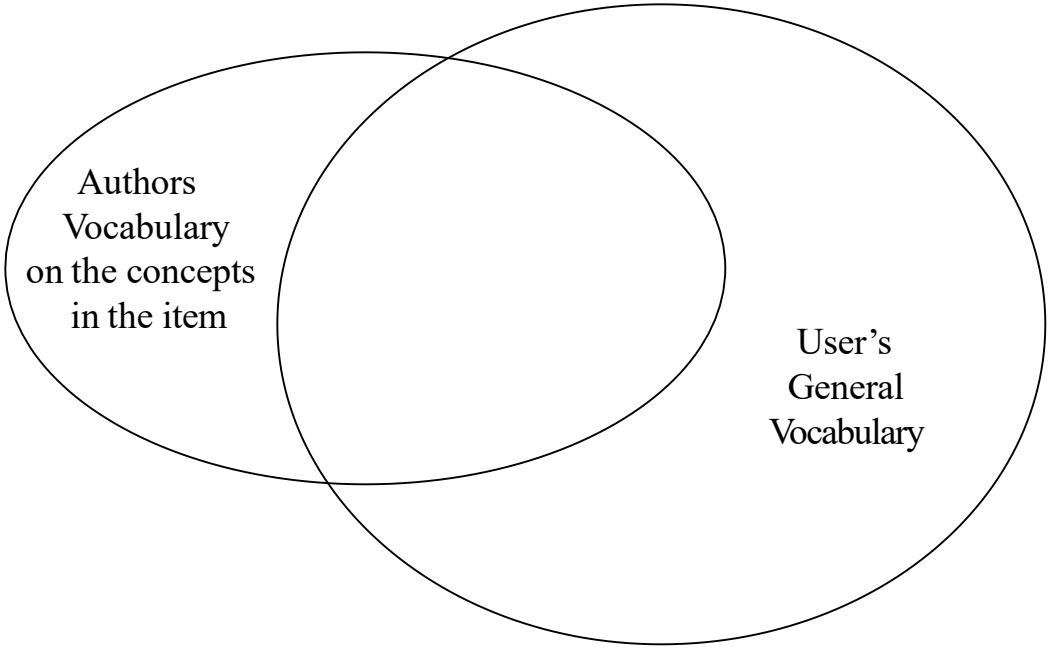


Fig: Vocabulary Domains

Contents

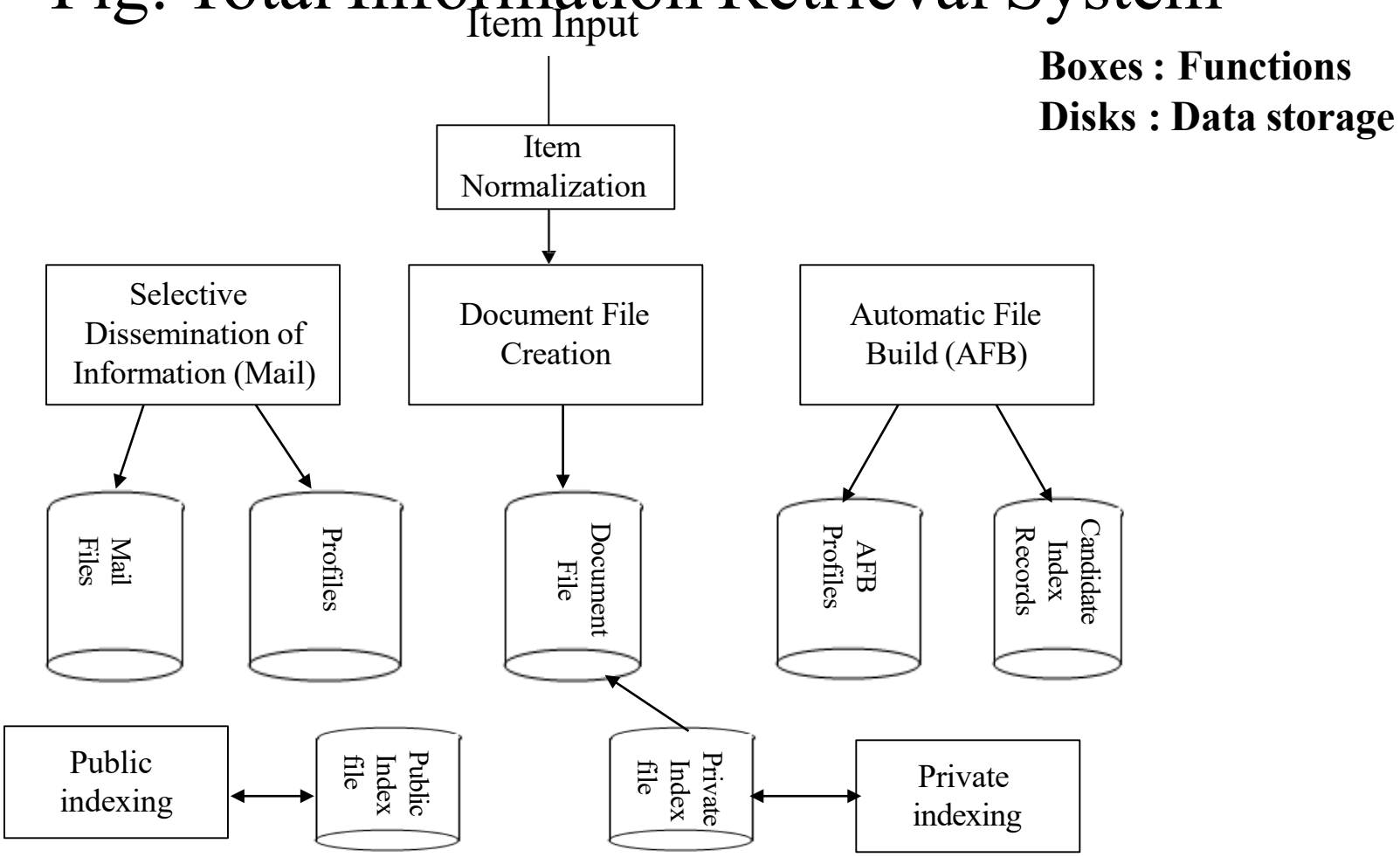
- Introduction
- Definition of Information Retrieval System
- Objectives of Information Retrieval System
- **Functional Overview**
- Relationship to RDBMS
- Digital Libraries and Data Warehouses
- Summary

Functional Overview

- The total information storage and retrieval system consists of four major functional processes:
  - Item Normalization
  - Selective dissemination of information (i.e. Mail)
  - Archival Document Database Search
  - Index database search along with Automatic File Build Process

The next figure shows the logical view of these capabilities in a single integrated information retrieval system.

Fig: Total Information Retrieval System



Boxes : Functions
Disks : Data storage

Functional Overview

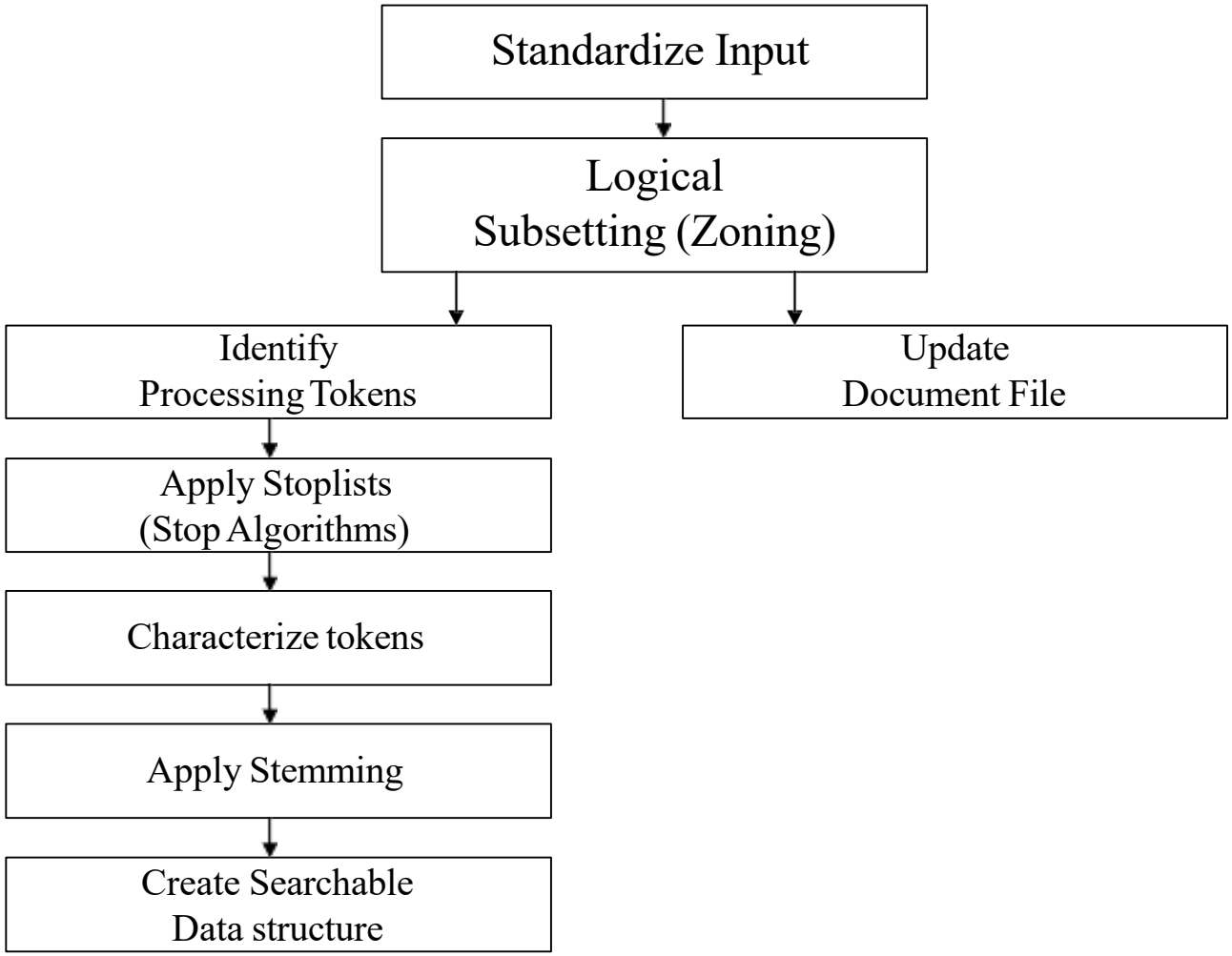
1. Item Normalization

- Normalize the incoming items to a standard format
- Provides logical restructuring of the item.
- Additional operations are needed to create a searchable data structure:
  - Identification of Processing tokens
  - Characterization of tokens
  - Stemming (eg: removing word endings) of tokens.

The processing tokens and their characterization are used to define the searchable text from the total received text.

The following figure shows the normalization process.

Fig: Text Normalization Process



Functional Overview

- Standardizing the input takes the different external formats of input data and performs the translation to the formats acceptable to the system (eg: translation of foreign languages into Unicode)
- One standard encoding that covers English, French, Spanish, etc. is ISO-Latin
- Multimedia adds an extra dimension to the normalization Process.
- If the input is video the likely digital standards will be: MPEG-2, MPEG-1 etc.
- MPEG (Motion Picture Expert Group) are the most universal standards for higher quality video.

Functional Overview

- ***Zoning:*** It is the process to parse the item into logical subdivisions that have meaning to the user
  - Used to increase the precision of a search and optimize the display
- Identify the Processing tokens
  - Consists of determining a word
  - Systems determine words by dividing input symbols into three classes:
    - Valid word Symbols
    - Inter-word symbols
    - Special processing symbols
  - ***Word:*** It is defined as a contiguous set of word symbols bounded by inter-word symbols (eg: of word symbols are Alphabetic characters and numbers, eg of inter-word symbols are: blanks, periods and semicolons)

Functional Overview

- Stop List Algorithm is applied to the list of potential processing tokens.
  - ***Objective of Stop function:*** To save system resources by eliminating from the set of searchable processing tokens those that have little value to the system.
  - Stop lists are commonly found in most systems and consists of words (Processing tokens) whose frequency and/or semantic use make them of no value as a searchable token.
  - The rank frequency law of Ziph
    - **Frequency * Rank = Constant**

      where Frequency = no. of times a word occurs and

      Rank = rank order of the word

Functional Overview

- The next step in finalizing on processing tokens is identification of any specific word characteristics.
- Once the potential processing token has been identified and characterized, most systems apply stemming algorithms to normalize the token to a standard semantic representation
- The decision to perform stemming is a trade-off between precision of a search Vs standardization to reduce system overhead in expanding a search term to similar token representations.
- Once the processing tokens have been finalized based upon the stemming algorithm, they are used as updates to the searchable data structure.

Functional Overview

2. Selective Dissemination of Information (Mail)

- This process 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 item.
- The Mail process is composed of the search process, user statements of interest and user mail files.

Functional Overview

3. Document database Search

- This process provides the capability for a query to search against all items received by the system.
- This process is composed of the search process, user entered queries and Document database which contains all items that have been received, processed and stored by the system.
- The Document database can be very large, hundreds of millions of items or more.
- Typically, items in the document database do not change (i.e not edited) once received.

Functional Overview

4. Index Database Search

- A user may want to save the interested item for future reference. This is accomplished via 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.
- A good analogy to an index file is the card catalog in a library.
- The index database search Process provides the capability tom create indexes and search them.

Functional Overview

- 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. This Process is called a **Combined File Search.**
- In an ideal system the index record could reference portions of items versus the total item

Functional Overview

- Two classes of Index files: Public and Private
  - Every user can have one or more Private Index Files leading to a very large number of files
  - Private Index File
    - References only a small subset of the total number of items in the Document Database.
    - Typically have very limited access lists.
  - Public Index Files
    - Maintained by professional library services personnel and typically index every item in the Document Database.
    - Have access lists that allow any one to search and retrieve data.

Functional Overview

- To assist the users in generating indexes, especially the professional indexers, the system provides a process called Automatic File Build.

- The capability to create Private and Public Index Files is frequently implemented via a Structured DBMS.

Functional Overview

- From a system perspective, the multimedia data is not logically its own data structure.
- It will reside almost entirely in the area described as the Document database
- The correlation between the multimedia and the textual domains will be either via Time or Positional synchronization
  - Time synchronization is an ex. of transcribed text from audio or composite video sources.
  - Positional synchronization is where the multimedia is localized by a hyperlink in a textual item.

Contents

- Introduction
- Definition of Information Retrieval System
- Objectives of Information Retrieval System
- Functional Overview
- **Relationship to DBMS**
- Digital Libraries and Data Warehouses
- Summary

Relationship to DBMS

- Information Retrieval System is software that has the features and functions require to manipulate " information" items
- A DBMS is optimized to handle structured data.
  - Structured data is well defined data represented by tables.
- Information is Fuzzy text.
  - The term fuzzy is used to imply the results from the minimal
    - standards or controls on the creators of the text items

Relationship to DBMS

- The integration of DBMS and Information Retrieval systems is very important.
- One of the first commercial databases to integrate the two systems into a single view is the INQUIRE DBMS (past 15 years)
- A more current example is the ORACLE DBMS that offers an imbedded capability called CONVECTIS
  - It is Information Retrieval system that uses a comprehensive thesaurus which provides the basis to generate “themes” for a particular item

Relationship to DBMS

- The INFORMIX DBMS has the ability to link to RetrievalWare to provide integration of structured data and Information along with functions associated with Information retrieval systems.

Contents

- Introduction
- Definition of Information Retrieval System
- Objectives of Information Retrieval System
- Functional Overview
- Relationship to DBMS
- **Digital Libraries and Data Warehouses**
- Summary

Digital Libraries and DWHs

- There is significant overlap between these two systems and information storage and retrieval systems.
- Digital libraries, DWHs and Information retrieval systems are the repositories of information.
  - ***Goal:*** to satisfy user information needs

Contents

- Introduction
- Definition of Information Retrieval System
- Objectives of Information Retrieval System
- Functional Overview
- Relationship to DBMS
- Digital Libraries and Data Warehouses
- **Summary**

Summary

- This unit places into perspective a total Information Storage and Retrieval System. This perspective introduces new challenges to the problems that need to be theoretically addressed and commercially implemented.

- From a theoretical perspective, efficient scalability of algorithms to systems with gigabytes and terabytes of data, operating with minimal user search statement information and making maximum use of all functional aspects of an information system need to be considered.

Summary

- The dissemination systems or mail files to modify ranking algorithms and combining the search of structured information fields and free text into a consolidated weighted output are examples of potential new areas of investigation.
- Understanding the differences between Digital Libraries and Information Retrieval systems will add an additional dimension to the potential future development of systems.
- The collaborative aspects of digital libraries can be viewed as a new source of information that dynamically could interact with information retrieval techniques.

Information Retrieval System Capabilities

Contents

- **Search Capabilities**
- Browse Capabilities
- Miscellaneous Capabilities
- Standards
- Summary

1. Search Capabilities

- The search capabilities address both Boolean and Natural Language Queries
- The algorithms used for searching are called Boolean, natural language processing and probabilistic.
- Probabilistic algorithms use frequency of occurrence of processing tokens in determining similarities between queries and items.
- The systems such as TOPIC, RetrievalWare and INQUERY allow for natural language queries.

Continued…

- **Objective:**
  - To allow for a mapping between a user's specified need and items in the information database that will answer that need.

1.1 Boolean Logic

- It allows a user to logically relate multiple concepts together to define what information is needed.
- Boolean functions apply to processing tokens identified anywhere within an item.
- Typical Boolean operators are AND, OR and NOT, and are implemented using intersection, set union and set difference procedures

Continued…

- **Ex:** Find any item containing any two of the following terms: ‘AA’, ‘BB’, ‘CC’. This can be expanded into a Boolean search that performs an AND between all combinations of two terms and ‘OR’s that results together ((AA AND BB) or (AA AND CC) or (BB AND CC))

- Most information retrieval systems allow Boolean operations and natural language interfaces

Fig: Use of Boolean Operators

SEARCH STATEMENT	SYSTEM OPEARTION
Computer OR Processor NOT Mainframe	Select all items discussing computers and/or Processors that do not discuss Mainframes
Computer OR (Processor NOT Mainframe)	Select all items discussing computers and/or items that discuss Processors and do not discuss Mainframes
Computer AND NOT Processor OR Mainframe	Select all items that discuss computers and not Processors or Mainframes in the item

Continued…

- 1.2 Proximity
  - It is used to restrict the distance allowed within an item between two search terms.
  - The semantic concept is that the closer two terms are found in a text , the more likely they are related in the description of a particular concept.
  - It is used to increase the precision of a search.
  - The typical format for proximity is:

    TERM 1 within "m" "units" of TERM 2

    The distance operator "m" is an integer number

    "units" are in Characters, Words, Sentences, or Paragraphs

Continued…

- The proximity relationship contains a direction operator indicating the direction (before or after) that the second term must be found within the number of units specified. Default is either direction.
- A special case of Proximity operator is the Adjacent (ADJ) operator that normally has a distance operator.
- Another special case is where the distance is set to zero (within the same semantic unit)

Fig: Use of Proximity

SEARCH STATEMENT

- SYSTEM OPEARTION

"Venetian" ADJ "Blind"

- Would find items that mention a Venetian Blind on a window but not items discussing a Blind Venetain

"United" within five words of "American"

- Would hit on "United States and American interests," "United Airlines and American Airlines" not on "United States of America and the American Dream

- Would find items that have "nuclear" and "clean-up" in the same paragraph.

Continued…

1.3 Contiguous Word Phrases (CWP)

- A CWP is both a way of specifying a query term and special search operator.
- A CWP is two or more words that are treated as a single semantic unit.
- **EX:** "United States of America"
  - It is four words that specify a search term representing a single specific semantic concept (a country)
- A CWP also acts like a special search operator that is similar to the proximity operator but allows for additional specificity

Continued…

- If two terms are specified, CWP and Proximity operator are identical.
- For CWP with more than two terms the only way of creating an equivalent search statement using proximity and Boolean operators is via nested Adjacencies.
- A CWPs are called Literal Strings in WAIS (Wide Area Information Servers) and Exact Phrases in RetrievalWare.
- In WAIS multiple Adjacency (ADJ) operators are used to define a Literal String. **Ex:** "United" ADJ "States" ADJ "of" ADJ "America"

Continued…

- 1.4 Fuzzy Searches
  - Fuzzy searches provide the capability to locate spelling of words that are similar to the entered search term.
  - This function is primarily used to compensate for errors in spelling of words
  - Fuzzy searching increases recall at the expense of decreasing precision (i.e. it can erroneously identify terms as the search term)
  - In the process of expanding a query term fuzzy searching includes other terms that have similar spellings, giving more weight to words in the database.

Continued…

- **EX:** Term entered is "computer"

  Fuzzy search would automatically include the following words from the information database: "computer", "compiter", "conputer", "computter", "compute".

- Fuzzy searching has its maximum utilization in systems that accept items that have been Optical Character Read (OCR)
- In the OCR process a hardcopy item is scanned into a binary image (usually at a resolution of 300 dots per inch or more)
- The OCR process is a pattern recognition process that segments the scanned image into meaningful subregions.
- The OCR process will then determine the character and translate it to an internal computer encoding (ex: ASCII or other)

Continued…

- 1.5 Term Masking

  - It is the ability to expand a query term by masking a portion of the term and accepting as valid any processing token that maps to the unmasked portion of the term.
  - The value of term masking is much higher in systems that do not perform stemming or only provide a very simple stemming algorithm.
  - Two types of search term masking are: fixed length
    variable length

Continued…

- Fixed length term masking is a single position mask. It masks out any symbol in a particular position or lack of that position in a word.
- Variable length "don't cares" allows masking of any number of characters within a processing token. The masking may be in the front, at the end, at both front and end or imbedded.

"* COMPUTER"	Suffix Search
"COMPUTER*"	Prefix Search
"*COMPUTER*"	Imbedded String Search

Fig: Term masking

SEARCH STATEMENT	SYSTEM OPEARTION
Multi$national	Matches "multi-national," "multiynational," "multinational". Does not match "multi national" single is two processing tokens.
computer	Matches , "minicomputer" "microcomputer" or "computer"
comput*	Matches "computers" "computing", "computes"
comput	Matches "microcomputers", "minicomputing", "compute"

Continued…

- 1.6 Numeric and Date Ranges
  - Term masking is useful when applied to words, but does not work for finding ranges of numbers or numeric dates.
  - To find numbers larger than “125” using a term “125*” will not find any number except those that begin with the digits “125”.
  - Systems as part of their normalization process, characterizes words as numbers or dates.
  - A user could enter inclusive (e.g., “125-425” or “4/2/93-5/2/95” for numbers and dates) to infinite ranges (“>125,” “<=233,” representing “Greater Than” or “Less Than or Equal”) as part of query.

Continued…

- 1.7 Concept/Thesaurus Expansion
  - Associated with both Boolean and Natural Language Queries is the ability to expand the search terms via Thesaurus or Concept Class database reference tool.
  - A thesaurus is typically a one-level or two-level expansion of a term to other terms that are similar in meaning.
  - Concept class is a tree structure that expands each meaning of a word into potential concepts that are related to the initial term.

Fig: Thesaurus for term “computer”

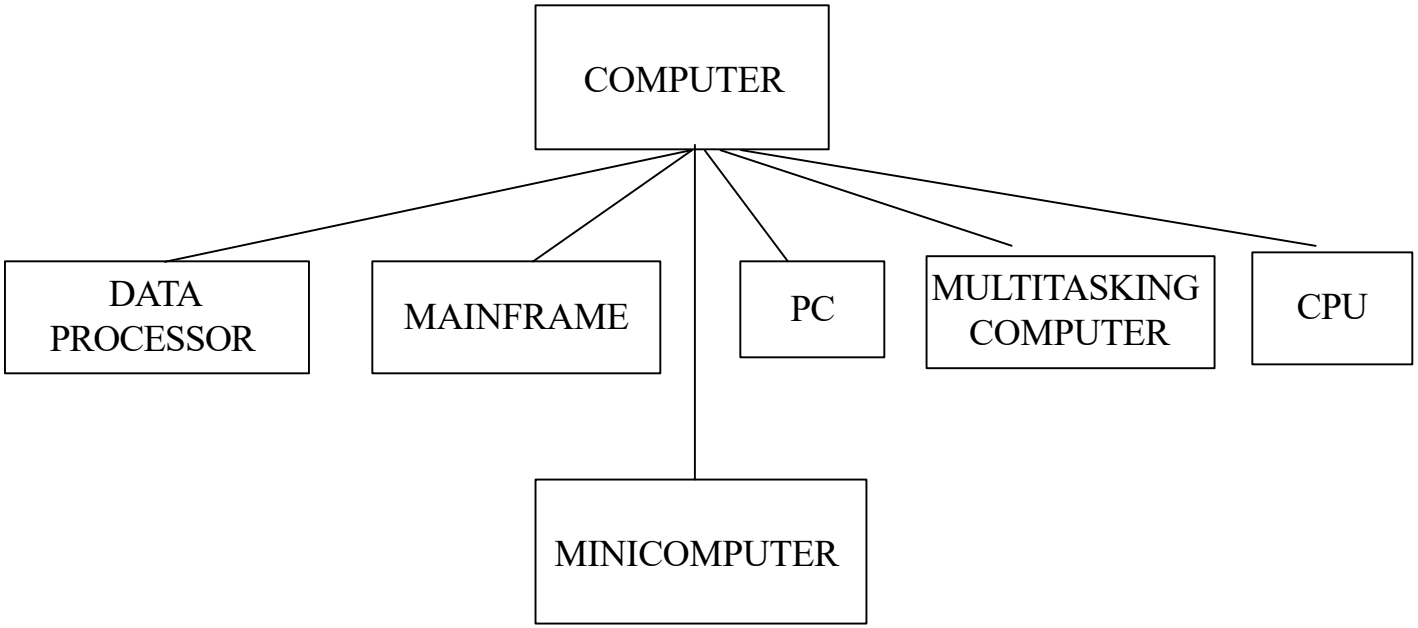
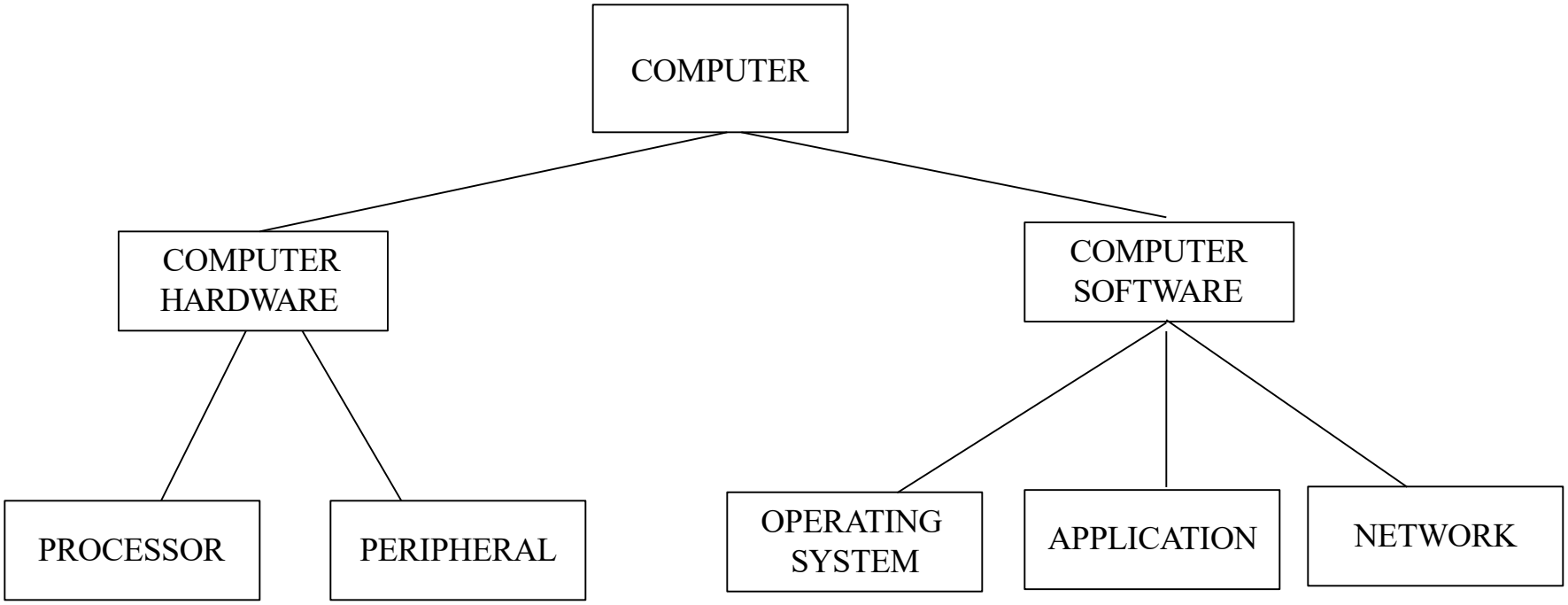


Fig: Hierarchical concept Class Structure for "computer"



Continued…

- 1.8 Natural Language Queries

  - Natural Language Queries allow a user to enter a prose statement that describes the information that the user wants to find.
  - The longer the prose, the more accurate the results returned.
  - The most difficult logic case associated with Natural language queries is the ability to specify negation in the search statement and have the system recognize it as negation.

Continued…

- To accommodate the negation function and provide users with a transition to the natural language systems, most commercial systems have a user interface that provides both a natural language and Boolean logic capability.
- Negation is handled by the Boolean portion of a search.
- Natural language interfaces improve the recall of systems with a decrease in precision when negation is required.

Continued…

- 1.9 Multimedia Queries
  - The user interface becomes far more complex with the introduction of the availability of multimedia items.
  - The current systems only focus on specification of still images.
  - Audio sources are converted to searchable text via audio transcription. This allows queries t be applied to the text.
  - But, like OCR output, the transcribed audio will contain many errors.
  - Thus, the search algorithms must allow for errors in the data. The errors are very different compared to OCR.

Continued…

- OCR errors will usually create a text string that is not a valid word..
- In ASR (Automatic Speech Recognition), all errors are other valid words since ASR selects entries ONLY from dictionary of words.
- Audio also allows the user to search on specific speakers, since speaker identification is relatively accurate against audio sources.
- The correlation between different parts of a query against different modalities is usually based upon time or location. Most common is Time.

Contents

- Search Capabilities
- **Browse Capabilities**
- Miscellaneous Capabilities
- Standards
- Summary

2. Browse Capabilities

- Browse capabilities provide the user with the capability to determine which items are of interest and select those to be displayed.
- There are two ways of displaying a summary of the items that are associated with a query: Line item status

  Data visualization
- From these summary displays, the user can select the specific items and zones within the items for display

Continued….

- The system also allows for easy transitioning between the summary displays and review of specific items.
- If searches resulted in high precision, then the importance of browse capabilities would be lessened.
- Since searches return many items that are not relevant to the user's information need, browse capabilities can assist the user in focusing on items that have the highest likelihood in meeting the need.

2.1 Ranking

- With the introduction of ranking based upon predicted relevance values, the status summary displays the relevance score.
  - **Relevance score:** It is an estimate of the search system on how closely the item satisfies the search statement.
    - Relevance scores are normalized to a value between 0.0 and 1.0
    - The highest value of 1.0 is interpreted that the system is sure that the item is relevant to the search statement.

2.1 Ranking

- Theoretically every item in the system could be returned but many of the items will have a relevance value of 0.0.
- Practically, systems have a default minimum value which the user can modify that stops returning items that have a relevance value below the specified value.
- In many circumstances Collaborative Filtering is providing an option for selecting and ordering output
  - In this, users when reviewing items provide feedback to the system on the relative value of the item being accessed.

2.1 Ranking

- The system accumulates the various user rankings and uses this information to order the output for other user queries that are similar.
- Collaborative filtering has been very successful in sites such as AMAZON.COM, MovieFinder.com and CDNow.com in deciding what products to display to users based upon their queries.

- Information visualization is also being used in displaying individual items and the terms that contributed to the item's selection.

2.2 Zoning

- When the user displays a particular item, the objective of minimization of overhead still applies.
- The user wants to see the minimum information needed to determine if the item is relevant.
- Once the determination is made an item is possibly relevant, the user wants to display the complete item for detailed review.

2.2 Zoning

- Related to zoning for use in minimizing what an end user needs to review from a hit item is the idea of locality and passage based search and retrieval.
- Here the basic search unit is not complete item, but an algorithmic defined subdivision of the item.
- This is known as passage retrieval where the item is divided into uniform sized passages that are indexed and locality based retrieval where passage boundaries can be dynamic.

2.3 Highlighting

- The indication, frequently highlighting, lets the user quickly focus on the potentially relevant parts of the text to scan for item relevance.
- It has always been useful in Boolean systems to indicate the cause of the retrieval. This is because of the direct mapping between the terms in the search and the terms in the item.
- Information visualization appears to be a better display process to assist in helping the user formulate the query than highlighting items.

Contents

- Search Capabilities
- Browse Capabilities
- **Miscellaneous Capabilities**
- Standards
- Summary

3. Miscellaneous Capabilities

- There are many additional functions that facilitate the user's ability to input queries, reducing the time it takes to generate the queries.
  - Vocabulary Browse
  - Iterative searching and search history log
  - Canned queries

3.1 Vocabulary browse

- Vocabulary Browse provides knowledge on the processing tokens available in the searchable database.
- It provides the capability to display 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.
- The user can enter a word or word fragment and the system will begin to display the dictionary around the entered text.

3.1 Vocabulary browse

Below table shows what is seen in vocabulary browse if the user enters **“comput”**

TERM	OCCURRENCES
Computation	265
Comput	1245
Computen	1
Computer	10,800
Computerize	18
Computes	29

Fig: Vocabulary Browse list with entered term “comput”

3.1 Vocabulary browse

- The system indicates what word fragment the user entered and then alphabetically displays other words found in the database.
- The user can continue scrolling in either direction reviewing additional terms in the database.
- Vocabulary browse provides information on the exact words in the database.

3.2 Iterative search and search history log

- Frequently a search returns a Hit file containing many more items than the user wants to review.
- Rather than typing in a complete new query, the results of the previous search can be used as a constraining list to create a new query that is applied against it.
- This has the same effect as taking the original query and adding additional search statement against it in an AND condition.
- This process of refining the results of a previous search to focus on the relevant items is called iterative search.

3.2 Iterative search and search history log

- During a login session, a user could execute many queries to locate the needed information.
- To facilitate locating previous searches as starting points for new searches, search history logs are available.
- The search history log is the capability to display all the previous searches that were executed during the current session.

3.3 Canned Query

- The capability to name a query and store it to be retrieved and executed during a later user sessions is called canned or stored queries.
- A canned allows a user to create and refine a search that focuses on the user's general area of interest one time and then retrieve it to add additional search criteria to retrieve data that is currently needed.
- Queries that start with a canned query are significantly larger than ad hoc queries.

3.4 Multimedia

- To display more aggregate data, textual interfaces sometimes allow for clustering of the hits and then use of graphical display to show a higher level view of the information.