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 :Sachine 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

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

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

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 possibility
es 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

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

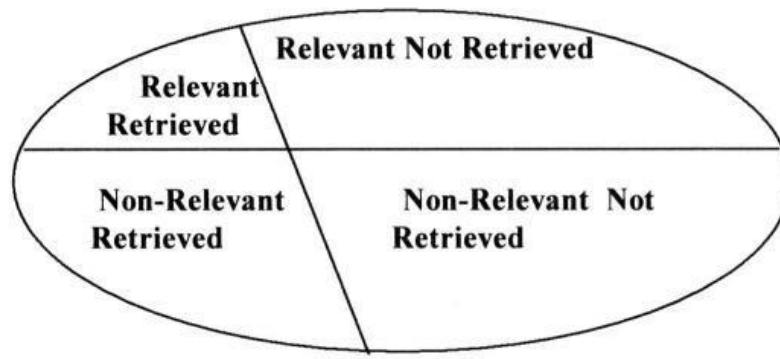


Figure 1.1 Effects of Search on Total Document Space

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

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_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.

- 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.

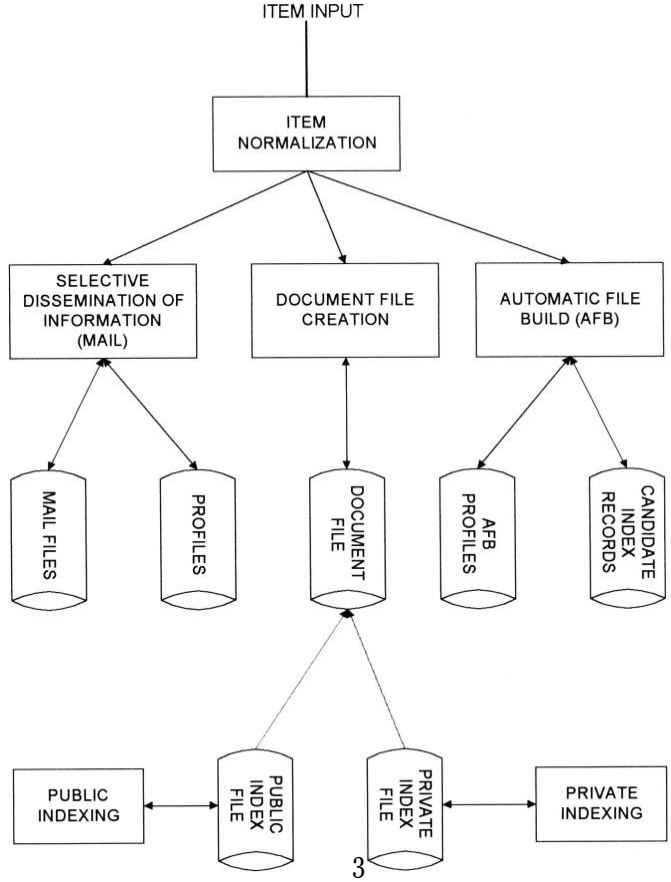


Figure 1.4 Total Information Retrieval System

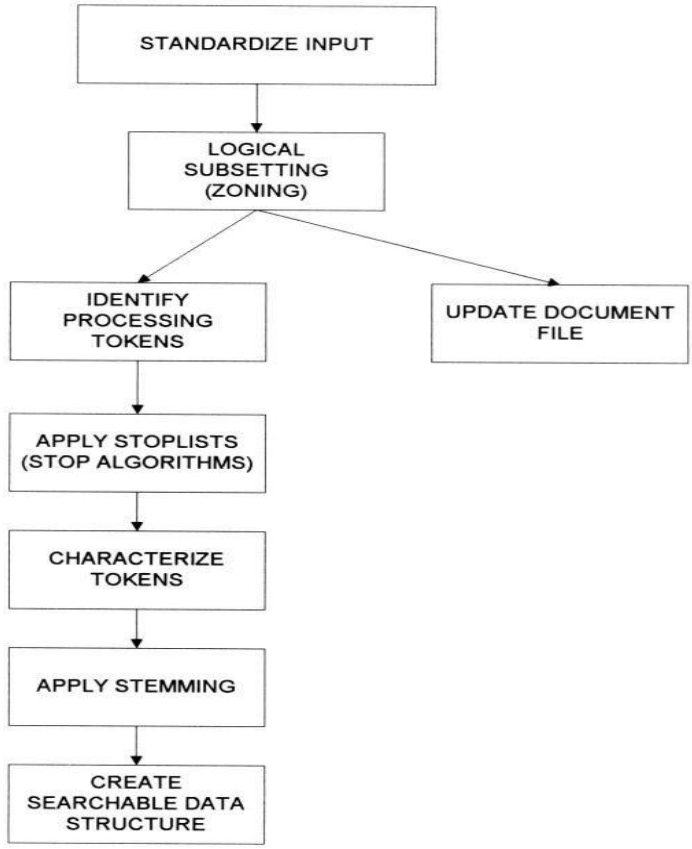


Figure 1.5 The Text Normalization Process

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 Analytices, 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.

Information Retrieval Systems
UNIT-2

CATALOGING AND INDEXING
CATALOGING:
It is also a Systematic arrangement of Items in an Alphabetical or other logical order Including brief Description

- A Catalogue is the record of the collection in the Library
- A library Catalogue is a list of books and other reading material available in a particular Library
- The card Catalogue has been a familiar sight to Library users for generations

  But, it has been effectively replaced by the online public Access catalog

  TYPES OF CATALOGUES:

  - Author Catalogues
  - Title Catalogues
  - Author/Title Catalogues
  - Subject Catalogues

Author Catalogues:
The Author Catalogues contain entires with Author names as the heading, Authors may be persons or Corporate bodies and the term Author is normally extended to Included writers, Illustrators ,performers ,producers, translators, &others with some Intellectual or Artistic responsibility for a Work
Eg: Vikas publishing pvt ltd, SIA Publications

Title Catalogues:
The Title Catalogue has entries with title as the heading some libraries and Information centers make title entries for all items being Indexed, but in other situations title entries are made selectively for only one Material

Author/Title Catalogues:
The Author/Title Catalogues contain both title and author Entries As both titles and Authors names are in Alphabetical order,It is Easy to file together Authors Names and Titles as headings
Subject Catalogues:
Subject Catalogues have an Indication of the Subject of the Documents being Indexed as their headings, The Entries are arranged in an appropriate System order
EX:
Car, Lawyers, These entries are arranged Alphabetically according to the subject heading

INDEXING:

Indexing is an Important process in Information Retrieval Systems

It forms the core Functionality of the IR Process Since, It is the first step in IR and assists in efficient Information Retrieval ,Indexing reduces the documents to the Informative terms contained in them

The transformation from received item to searchable data structure is called indexing.

- Process can be manual or automatic.
- Creating a direct search in document data base or indirect search through indexfiles.
- Concept based representation: instead of transforming the input into a searchable format some systems transform the input into different representation that is concept based .Search ? Search and return item as per the incoming items.

History of indexing:

It shows the dependency of information processing capabilities on manual and then automatic processing systems .

- Indexing originally called cataloging : oldest technique to identity the contents of items to assist in retrieval.
- One of the technique as similar as cataloging &Indexing ,the technique as both are Systematic Arrangement of items in an Alphabetical
- Items overlap between full item indexing , public and private indexing of files

Objectives of Indexing :

The public file indexer needs to consider the information needs of all users of library system . Items overlap between full item indexing , public and private indexing of files
•Users may use public index files as part of search criteria to increase recall.

•They can constrain there search by private index files

•The primary objective of representing the concepts within an item to facilitate users finding relevant information.

•Users may use public index files as part of search criteria to increase recall.

•They can constrain there search by private index files

•The primary objective of representing the concepts within an item to facilitate users finding relevant information.

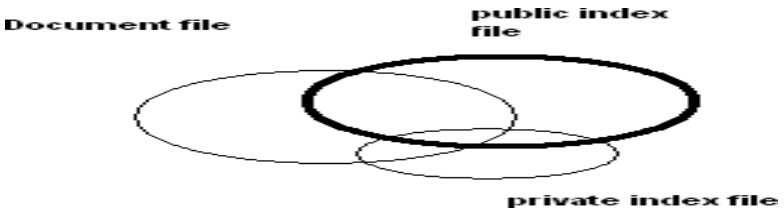


Fig:

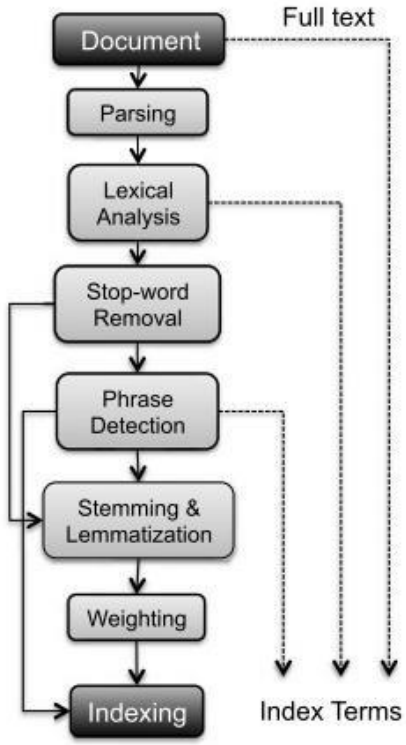
- 1.Decided the scope of indexing and the level of detail to be provided. Based on usage scenario of users.
- 2. Second decision is to link index terms together in a single index for a particular concept.

Indexing process:

Indexing Process is the collecting, Parsing, &Storing of data to facilitate fast and accurate Information retrieval Index Design incorporates Interdisciplinary Concepts from Linguistics, Cognitive Psychology, Mathematics, Informatics &Computer Science

TEXT PROCESSING

Fig. 2.3 Text processing phases in an IR system



Text process phases

1. **Document Parsing:** Documents come in all sorts of languages, character sets, and formats; often, the same document may contain multiple languages or formats, e.g., a French email with Portuguese PDF attachments. Document parsing deals with the recognition and "breaking down" of the document structure into individual components. In this pre processing phase, unit documents are created; e.g., emails with attachments are split into one document representing the email and as many documents as there are attachments.

2. **Lexical Analysis**: After parsing, lexical analysis tokenizes a document, seen as an input stream, into words. Issues related to lexical analysis include the correct identification of accents, abbreviations, dates, and cases. The difficulty of this operation depends much on the language at hand: for example, the English language has neither diacritics nor cases, French has diacritics but no cases, German has both diacritics and cases. The recognition of abbreviations and, in particular, of time expressions would deserve a separate chapter due to its complexity and the extensive literature in the field For current approaches

3. **Stop-Word Removal:** A subsequent step optionally applied to the results of lexical analysis is stop-word removal, i.e., the removal of high-frequency words. For example, given the sentence "search engines are the most visible information retrieval applications" and a classic stop words set such as the one adopted by the Snowball stemmer,1 the effect of stop-word removal would be: "search engine most visible information retrieval applications".

4. **Phrase Detection:** This step captures text meaning beyond what is possible with pure bag- of-word approaches, thanks to the identification of noun groups and other phrases. Phrase detection may be approached in several ways, including rules (e.g., retaining terms that are not separated by punctuation marks), morphological analysis , syntactic analysis, and combinations thereof. For example, scanning our example sentence "search engines are the most visible information retrieval applications" for noun phrases would probably result in identifying "search engines" and "information retrieval".

5. **Stemming and Lemmatization:** Following phrase extraction, stemming and lemmatization aim at stripping down word suffixes in order to normalize the word.

Stemming as Removing words ending , In particular stemming is a heuristic process that "chops off" the ends of words in the hope of achieving the goal correctly most of the time; a classic rule based algorithm for this was devised by Porter ,

According to the Porter stemmer, our example sentence “Search engines are the most visible information retrieval applications” would result in: “Search engine are the most visible inform retrieval application”.

- Lemmatization is a process that typically uses dictionaries and morphological analysis of words in order to return the base or dictionary form of a word, thereby collapsing its inflectional forms (see, e.g., [278]). For example, our sentence would result in “Search engine are the most visible information retrieval application” when lemmatized according to a WordNet-based lemmatizer

6. **Weighting:** The final phase of text pre processing deals with term weighting. As previously mentioned, words in a text have different descriptive power; hence, index terms can be weighted differently to account for their significance within a document and/or a document collection. Such a weighting can be binary, e.g., assigning 0 for term absence and 1 for presence.

SCOPE OF INDEXING

- When perform the indexing manually, problems arise from two sources the author and the indexer the author and the indexer.
- Vocabulary domain may be different the author and the indexer.
- This results in different quality levels of indexing.
- The indexer must determine when to stop the indexing process.
- Two factors to decide on level to index the concept in a item.
- The exhaustively and how specific indexing is desired.
- Exhaustively of index is the extent to which the different concepts in the item are indexed.
- For example, if two sentences of a 10-page item on microprocessors discusses on-board caches, should this concept be indexed
- Specific relates to preciseness of index terms used in indexing.
- For example, whether the term "processor" or the term "microcomputer" or the term "Pentium" should be used in the index of an item is based upon the specificity decision.
- Indexing an item only on the most important concept in it and using general index terms yields low exhaustively and specificity.
- Another decision on indexing is what portion of an item to be indexed Simplest case is to limit the indexing to title and abstract(conceptual) zone .
- General indexing leads to loss of precision and recall. PREORDINATION AND LINKAGES
- Another decision on linkages process whether linkages are available between index terms for an item.
- Used to correlate attributes associated with concepts discussed in an item. this process is called preordination.
- When index terms are not coordinated at index time the coordination occurs at search time. This is called post co-ordination , implementing by "AND" ing index terms.
- Factors that must be determined in linkage process are the number of terms that can be related.
- Ex., an item discusses 'the drilling of oil wells in Mexico by CITGO and the introduction of oil refineries in Peru by the U.S.'

AUTOMATIC INDEXING

Automatic Indexing is the computerized process of Scanning Large volumes of Documents against a controlled Vocabulary,Taxonomy or Ontology and using those controlled terms to quickly and effectively Index large Electronic Document depositories

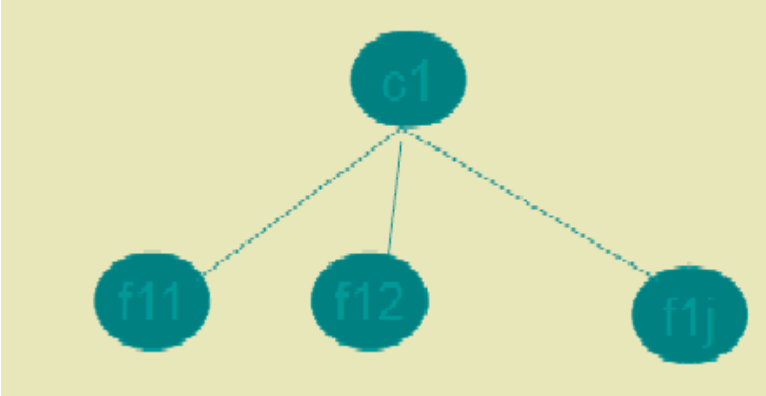
- Case Total document indexing.
- Automatic Indexing requires few seconds based on the processor and complexity of algorithms to generate indexes.'
- Index resulting form automated indexing fall into two classes , weighted and un weighted.
- Weighted indexing system: a attempt is made to place a value on the index term associated with concept in the document. Based on the frequency of occurrence of the term in the item.
- Un weighted indexing system : the existence of an index term in a document and some times its word location are kept as part of searchable data structure.
- Values are normalized between 0 and1.
- The results are presented to the user in order of rank value from highest number to lowest number.

Indexing By term

- Terms (vocabulary) of the original item are used as basis of index process.
- There are two major techniques for creation of index statistical and natural language.
- Statistical can be based upon vector models and probabilistic models with a special case being Bayesian model(accounting for uncertainty inherent in the model selection process).
- Called statistical because their calculation of weights use information such as frequency of occurrence of words.
- Natural language also use some statistical information, but perform more complex parsing to define the final set of index concept.
- Other weighted systems discussed as vectorised Information system.
- The system emphasizes weights as a foundation for information detection and stores these weights in a vector form.
- Each vector represents a document. And each position in a vector represent a unique word(*processing token*) in a database..
- The value assigned to each position is the weight of that term in the document.
- 0 indicates that the word was not in the document.
- Search is accomplished by calculating the distance between the query vector and

document vector.

- Bayesian approach: based on evidence reasoning(drawing conclusion from evidence)
- Could be applied as part of index term weighing. But usually applied as part of retrieval process by calculating the relationship between an item and specific query.
- Graphic representation each node represents a random variable arch between the nodes represent a probabilistic dependencies between the node and its parents.
- Two level Bayesian network
- " c"" represents concept in a query
- "f" representing concepts in an item



- Another approach is natural language processing.
- DR-LINK(document retrieval through linguistics knowledge)
- Indexing by concept
- Concept indexing determines a canonical set of concept based upon a test set of terms and uses them as base for indexing all items. *Called latent semantics indexing.*
- Uses neural NW strength of the system word relationship (synonyms) and uses the information in generating context vectors.
- Two neural networks are used one to generated stem context vectors and another one to perform query.
- Interpretation is same as the weights.
- Multimedia indexing:
- Indexing video or images can be accomplished at raw data level.

INFORMATION EXTRACTION

There are two processes associated with information extraction:

- 1.determination of facts to go into structured fields in a database and
- 2. Extraction of text that can be used to summarize an item.
- The process of extracting facts to go into indexes is called Automatic File Build.
- In establishing metrics to compare information extraction, precision and recall are applied with slight modifications.

- Recall refers to how much information was extracted from an item versus how much should have been extracted from theitem.
- It shows the amount of correct and relevant data extracted versus the correct and relevant data in the item.
- Precision refers to how much information was extracted accurately versus the total information extracted.
- Additional metrics used are over generation and fallout.
- Over generation measures the amount of irrelevant information that is extracted.
- This could be caused by templates filled on topics that are not intended to be extracted or slots that get filled with non-relevant data.
- Fallout measures how much a system assigns incorrect slot fillers as the number of
- These measures are applicable to both human and automated extraction processes.
- Another related information technology is document summarization.
- Rather than trying to determine specific facts, the goal of document summarization is to extract a summary of an item maintaining the most important ideas while significantly reducing the size.
- Examples of summaries that are often part of any item are titles, table of contents, and abstracts with the abstract being the closest.
- The abstract can be used to represent the item for search purposes or as a way for a user to determine the utility of an item without having to read the complete item.

DATA STRUCTURES

- Introduction to DataStructures
- StemmingAlgorithms
- Inverted FileStructure
- N-Gram DataStructure
- PAT DataStructure
- Signature FileStructure
- Hypertext and XML DataStructures

Introduction to Data Structures:

A Data structure is a specialized format for organizing processing, retrieving& storing data

- The knowledge of data structure gives an insight into the capabilities available to the system.
- Each data structure has a set of associated capabilities.

1. Ability to represent the concepts
2. Supports location of those concepts Introduction

- Two major data structures in anyIRS:

1. One structure stores and manages received items in their normalized form is called document manger
2. The other data structure contains processing tokens and associated data to support search.

Item Normalization:
The Item normalization is the Incoming items to a standard format whatever user is searching a Item ,It is not exactly user keyword converts into system understandable format

Document File Creation:
A Document file format is a text or binary file format for storing documents on a storage media especially for use by computers

Document Manager:
A Document Manager is a system used to receive, track manage and store Documents and reduce paper, Most of capable of keeping a record of the various versions created and modified by different users , In the case of management of Digital documents such systems are based on computer programs.

Document Search Manager:
The searching for Information in a document searching for documents themselves and also searching for the meta data that describes data &for databases of text images or sounds.

Processing Tokens:

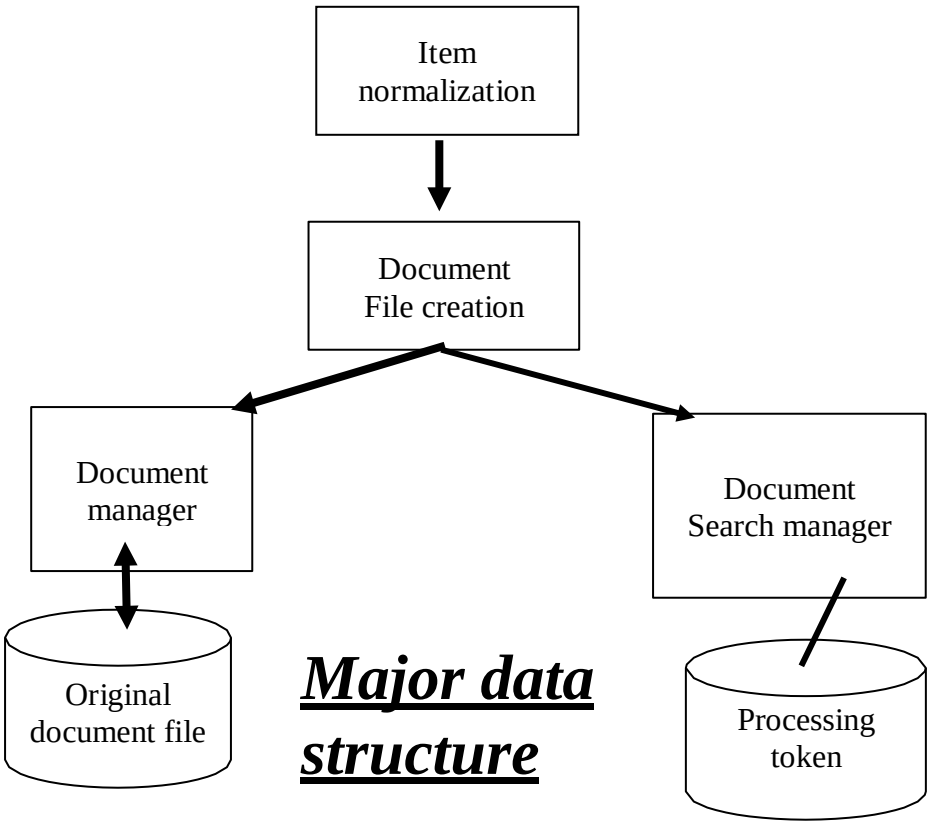
Identify the information that are used in the search process-processing tokens

Divide Input symbols into three classes

Valid word symbols : Alphabets, Numbers &special characters

Inter word Symbols: Blanks,Semi-colons(Non-searchable)

Special Processingsymbols:-Hyphen(-)



Result of a search are references to the items that satisfy the search statement which are passed to the document manager for retrieval.

Focus : on data structure that support search function

STEMMING ALGORITHMS:

Stemming is nothing but cutting &trimming

The concept of stemming is introduced in the 1960's

The main goal of stemming was to Improve performance and require less system resources by reducing the number of unique words that a system has to contain

The stemming algorithms are used to improve the efficiency of the information system &to Improve recall

- Reduce precision
- Increase Recall
- Reduce diversity
- Increase search efficiency
- Conflation

Stemming variations:

- Table lookup stemming
- Poster stemming
- Dictionary stemming(K-Stemming)
- Successor stemming

Table lookup stemming:

Uses Large Data structures

Ex: Retrieval ware

- K-Stemming Example INQUERY
- Combine rules+ dictionary words
- Iterative Nature
- Removes large prefixes&suffixes

Poster stemming Algorithm:

The poster stemming Algorithm is based up on a set of conditions of the stem, suffix &prefix and associated actions given the condition

Conditions are

The measure M of stem is a function of sequences of Vowels(A,E,I,O,U,Y)followed by a consonant

- If V is a sequence of vowels and C is a sequence of consonants, then M is: the number
- Where the initial C and final V are optional and M is the Number

  C(VC)M V

  - ***<X>**stem ends with a letter
  - ***V*** stem contains Vowel
  - ***d** stem ends with double consonant
  - Uses wild chord characters

Dictionary Stemming:

- Also called K-Stemming
- Dictionary based
- Used in Inquery called In query K-Stems
- Avoid collapsing word with different meanings

- Uses NLP Dictionaries
- Example British->Britan
- Extract meaning of words only it finds

Uses 6 major Data files

1. Dictionary of words

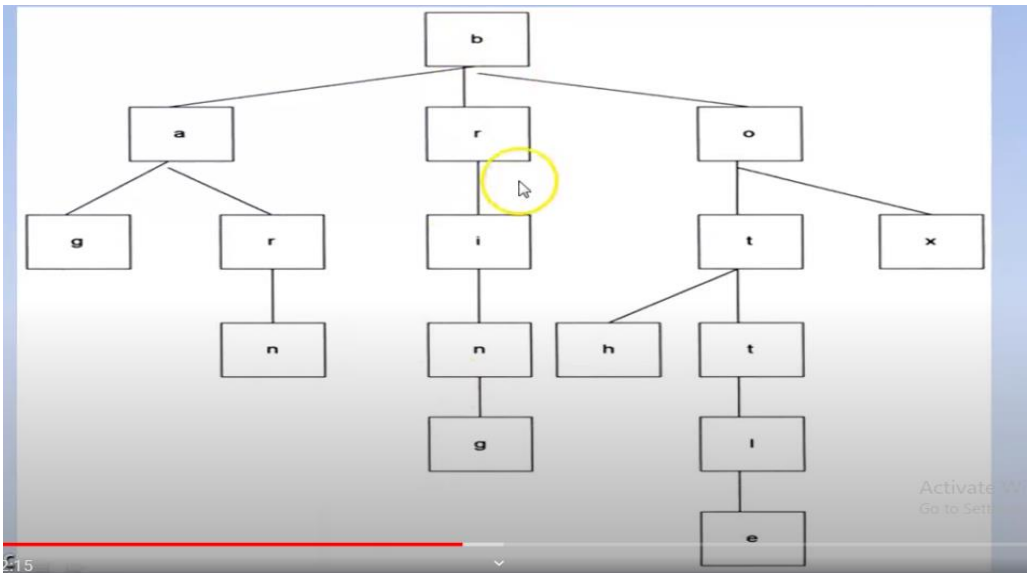
2. Supplement of words

3. Exceptions list of those words that should retain an “e” at the end(Eg:”Suites”to”suite”but “suited” to “suit”)

- Direct-conflation-allows definition of direct conflation via word pairs that override the stemming algorithm
- Country-Nationality-conflationsbetween Nationalities &countries(“British” maps to “Britian”)
- Proper nouns-list of proper nouns that should not be stemmed

Successor Stemming:

- It contains symbol tree for words
- Constructs symbol tree based on words
- Represents both prefix &suffix
- It Implements 3 methods
- Cut off
- Peak& plateau
- Complete word



Symbol tree for terms bag, barn, bring, box, bottle

Conclusion Stemming:

- Good efficient
- Depends on nature of vocabulary
- Stemming is as effective as Manual conflation
- Stemming can affect retrieval(recall)and where effects were Identified
- They were positive
    - It has a potential to increase recall. STEMMING ALGORITHMS
    - Stemming algorithm is used to improve the efficiency of IRS and improve recall.

Inverted File Structure

The most common Data structure used in both Database Management and information Retrieval Systems is the Inverted File Structure

- Inverted File Structures are composed of three basic files
- Document file
- Dictionary
- Inversion Lists

Features of Inverted File:

- Increases Precision
- Zoning used
- Ranking also used
- Used to store concepts &relationships
- NLP Used(Natural Language Processing)

Increases Precision:

The ability to retrieve Top ranked Documents that are Mostly relevant ,the scope has been increasing for Priority

Zoning used:

It is logical sub setting Information has to store specific zone

Ranking also used:

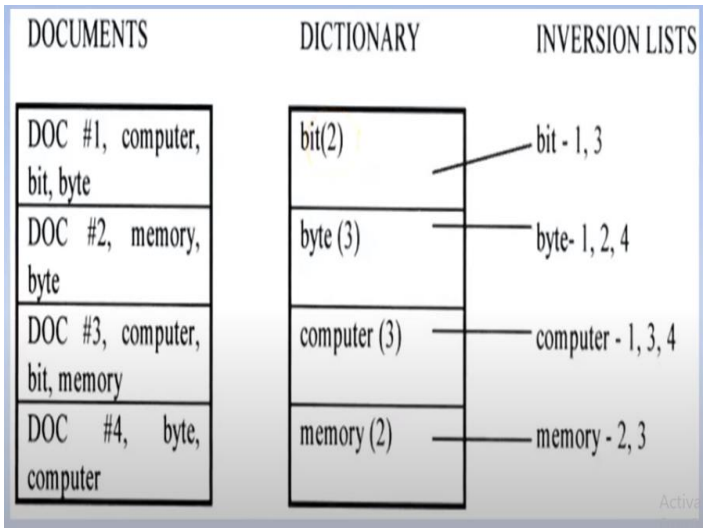
The public will be giving ranking based on particular file, Document, PDF, WORD…..etc

Store concepts &relationships:

It is maintaining as relationships between Database management system and Information retrieval systems

Natural Language Processing:

Natural Language processing as based on public emotions &views like Happy, sad.



- Inversion lists structures are used because they provide optimum performance in searching large Databases
- Inversion list file structures as well suited to store concepts &their relationships

N-Gram Data Structures

N-Gram Data Structures:

- N-gram is a one of the Data structure
- N-grams can be viewed as a special Technique for conflation (stemming) and as a Unique
- It has been logical mapping for searching of searchable Items
- N-grams are fixed length consecutive series of "n" characters

  Specializations:

  - Special Data structure
  - Ignore words(Ignore words &sentences)
  - Input as continuous Data
  - Logical Linkages
  - N-gram=N-Length
  - Trigram=3 Letters

Special Data structure:
It is one of the Data structure

Ignore words:
It is ignores a words &sentences Repeating a words once or twice

Input as continuous Data:
The data has to flow sequence manner &systematic order

N-gram:
It is Indicating as N-gram is equal to N-Length of characters

Trigram:
It is Indicating as trigram is equal to 3 letters of characters

It was Implemented a formula

MAX Segn =(λ)n

- Inversion Lists Document vectors are used
- Here Maximum number of n-grams used of Unique
- Retail trigrams are Ret, eta, tai etc
- Disadvantages is Longer N-grams results poor result
- N-gram characters strength is vey poor

Example:

- Se ea Col ol lo on ny
- Sea col olo lon ony
- #sea # #colo colon olony long#
- Inter words means symbols like non searchable
- Bigrams =2(no Inter word symbols)
- Trigrams=3(it will acceptable as Inter word symbols &No Inter word symbols)

Advantages:

- The first use of N-Grams dates to world war-II, when it was used by Cryptographers
- Another Major use of N-Grams in particular trigrams is in spelling error Detection &corrections
- Frequency of occurrence of N gram pattern also can be used for Identifying the language of an Item
- Because of the processing token bounds of N-gram data structures, optimized performance techniques can be applied in mapping items to an N-gram searchable structure & in query processing
- There is no semantic meaning in a particular n-gram since It is a fragment of processing token and may not represent a concept
- Thus n-grams are a poor representation of concepts &their relationships

PAT data structure:

PAT is one of the Data structure, Practical algorithm to retrieve Information coded in alpha numeric, PAT tree is a data structure that allows very efficient searching with processing

- PAT structure or PAT tree or PAT array : continuous text input data structures(string like N- Gram datastructure).
- The input stream is transformed into a searchable data structure consisting of substrings, all substrings are unique.
- Each position in a input string is a anchor point for a substring.
- Using n-grams with inter word symbols included between valid processing tokens equates to continous text Input Data structure that is being Indexed in contiguous "n"characters Tokens
- Different view of Addressing a continous Text
- Input data structure comes from PAT Trees &PAT Arrays
- The Input stream is transformed into a searchable Data structure consisting of substrings
- In creation of PAT Trees each position in the Input string is the anchor point for a sub-string that starts at that point and Include all new text up to the end of the Input
- All substrings are unique
- This view of text lends itself to many different search processing structures
- Substring can start at any point in the text and can be uniquely by its starting location &length
- A PAT Tree is unbalanced, binary digital tree defined by the Sistrings
- The individual bits of the Sistring decide the branching patterns with zero branching left and one branching right
- PAT Trees also allow each node in the tree to specify which bit is used to determine the branching via bit position
- We have to eliminate Sistring a text wherever we want position of Text
- Text
- Economics for Warsaw is Complex
- Sistring=1
- Conomics for Warsaw is Complex
- Sistring =2

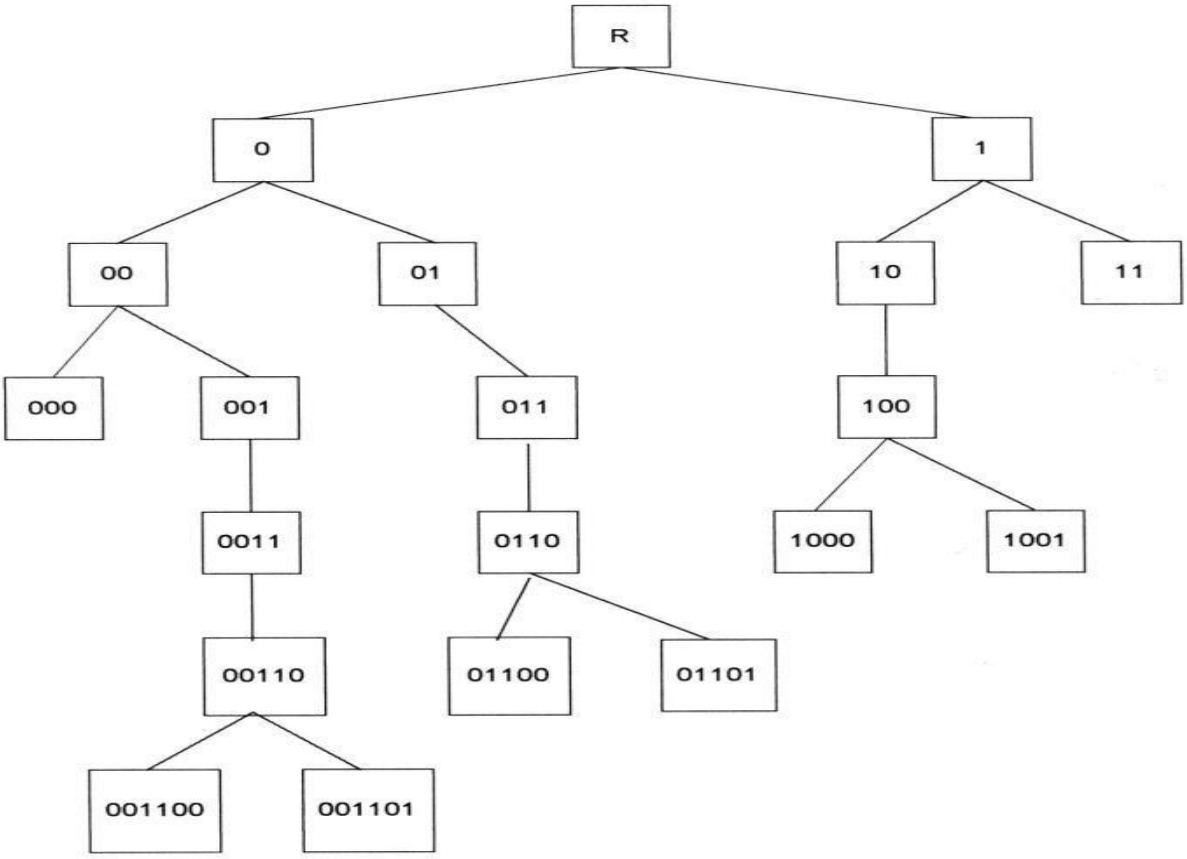
- Onomics for Warsaw is Complex
- Sistring=4
- Omics for Warsaw is Complex
- :
- :
- :
- :
- Sistring=9
- For Warsaw is Complex
- :
- :
- :
- :
- :
- :
- Sistring=25
- Ex

- In creation of PAT trees each position in the input string is the anchor point for a sub-string that starts at that point and includes all new text up to the end of the input.
- Binary tree, most common class for prefix search, But Pat trees are sorted logically which facilitate range search, and more accurate then inversion file.
- PAT trees provide alternate structure if supporting strings search.

    - The key values are stored at the leaf nodes (bottom nodes) in the PATTree.
        - For a text input of size "n" there are "n" leaf nodes and "n-1" at most higher level nodes.

          It is possible to place additional constraints on sistrings for the leafnodes

The full PAT binary tree is



Signature file structure:

- the goal of a signature file structure is to provide a fast test to eliminate the majority of items that are not related to a query
- because file structure is highly compresses and unordered, It requires significantly less space than an Inverted file structure
- New items can be concatenated to the end of the structure
- When items are deleted from Information Databases, It leaves deleted Items in place and mark them as deleted
- Signature file structure is a linear scan of the Compressed of Items producing a response time linear with respect to a file size

Application(s)/Advantage(s)

- Signature files provide a practical solution for storing and locating information in a number of different situations.
- Signature files have been applied as medium size databases, databases with low frequency of terms, WORM devices, parallel processing machines, and distributed environments

HYPERTEXT AND XML DATA STRUCTURES:

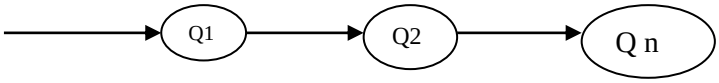
- The Hypertext Data structure is used Extensively in the Internet environment and requires an electronic media storage for the item
- Hypertext allows one item to reference another Item via an Embedded pointer Each separate Item is called a node and the reference pointer is called a link
- Each node is displayed by a viewer that is defined for the file Type associated with the node
- For Example(Html) defines the Internal Structure for Information Exchange across the world wide web on the Internet
- A Document is Composed of the text of the Item a long with Html Tags that describes how to display The Document
- Tags are formatting or structural keywords contained between less than greater than symbols (Eg:-<title>,meaning display prominently)

- The advent of the Internet and its exponential growth and wide acceptance as a new global information network has introduced new mechanisms for representing information.
- This structure is called hypertext and differs from traditional information storage data structures in format and use.
- The hypertext is Hypertext is stored in HTML format and XML.

- Bot of these languages provide detailed descriptions for subsets of text similar to the zoning.
- Hypertext allows one item to reference another item via a embedded pointer.
- HTML defines internal structure for information exchange over WWW on theinternet.
- XML: defined by DTD, DOM, XSL,etc.

HIDDEN MARKOV MODELS:

The Hidden Markov Models is used for searching as Textual Queries has Introduced a new Paradigm for search, the output of one term of query = = Input of another query



- The Hidden Markov models one Input is generated again it produce as output & that output has creating as one Input ,it is come across as chain process
- The statistical process that can generate output that is equivalent to the set of queries that would consider the document relevant
- The general definition that a HMM(Hidden Markov Models)is a defined by the output that is produced by passing some unknown key via state transitions through a noisy channel output is the query and the unknown keys are the relevant Documents
- Channel is the mismatch between the author's way of expressing idea's and the Users ability to specify his query
- The development for HMM(Hidden Markov Models) approach begins with applying Bayes Rule to the conditional Probability

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$$

- The users ideas are Transformed channel is the Transmission of messages Transmission of the queries users idea's are Transformed and act as Input as Another Queries

Disadvantages of HMM:

The biggest problem in using this approach is to estimate the transition probability Matrix and the output for every Document

If there was a large training Database of queries and the relevant documents then the problem could be solved using Estimation-Maximization Algorithms

III UNIT

AUTOMATIC INDEXING:

Automatic Indexing is the computerized process of Scanning Large volumes of Documents against a controlled Vocabulary, Taxonomy or Ontology and using those controlled terms to quickly and effectively Index large Electronic Document depositories

- Case Total document indexing.
- Automatic Indexing requires few seconds based on the processor and complexity of algorithms to generate indexes.'
- Index resulting form automated indexing fall into two classes , weighted and un weighted.
- Weighted indexing system: a attempt is made to place a value on the index term associated with concept in the document. Based on the frequency of occurrence of the term in the item.
- Un weighted indexing system : the existence of an index term in a document and some times its word location are kept as part of searchable data structure.
- Values are normalized between 0 and1.
- The results are presented to the user in order of rank value from highest number to lowest number.

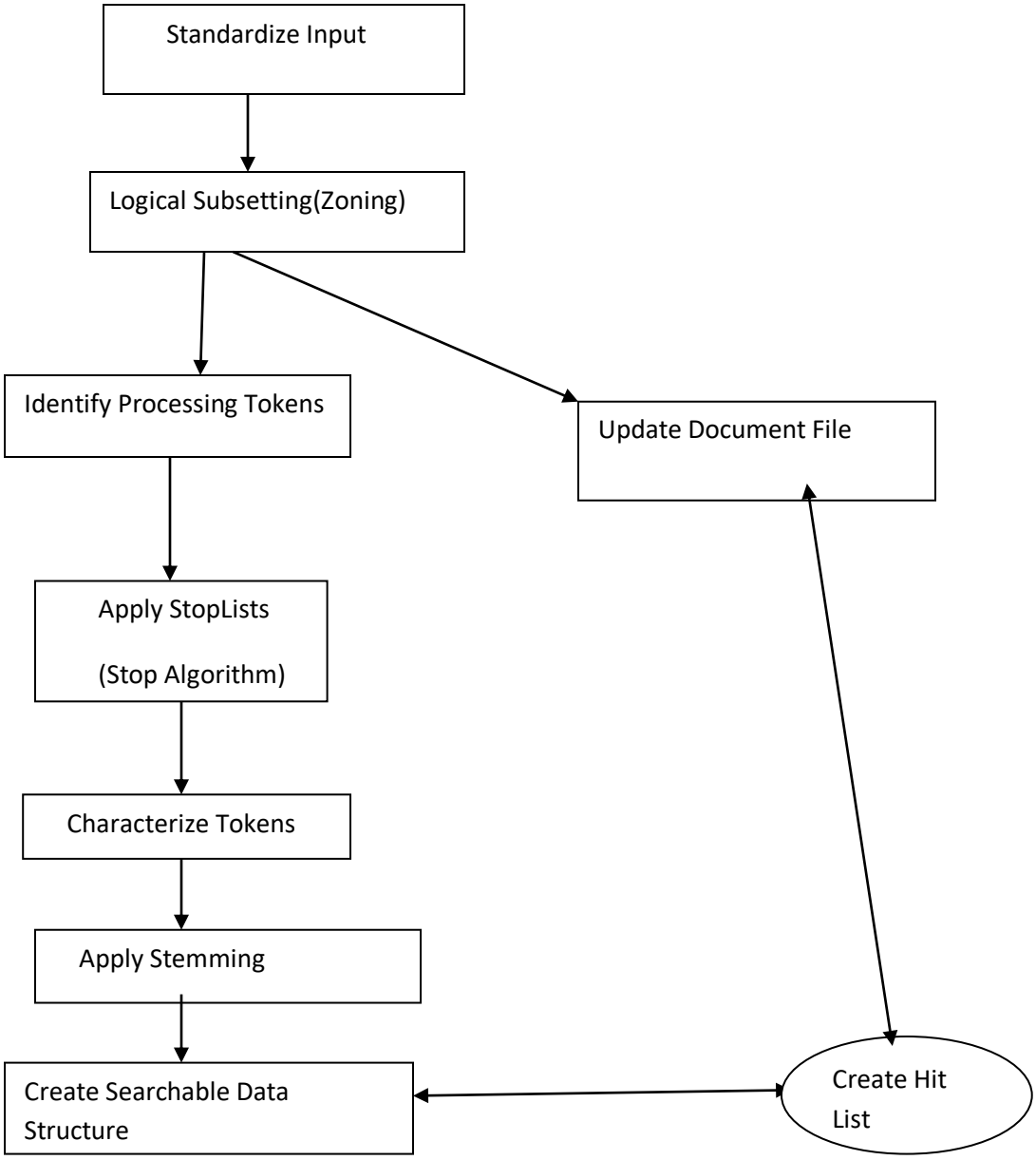
Indexing By term

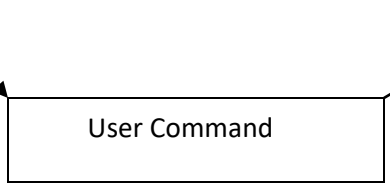
- Terms (vocabulary) of the original item are used as basis of index process.
- There are two major techniques for creation of index statistical and natural language.
- Statistical can be based upon vector models and probabilistic models with a special case being Bayesian model(accounting for uncertainty inherent in the model selection process).
- Called statistical because their calculation of weights use information such as frequency of occurrence of words.
- Natural language also use some statistical information, but perform more complex parsing to define the final set of index concept.

Types of Classes in Automatic Indexing:

- Automatic Indexing is the process of Analyzing an Item to extract the information to be permanently kept in an Index
- This process is associated with the generation of the searchable Data structure Associated with an Item
- The indexing process is shown in the following fig
- The left side of the figure Including Identify Processing Tokens, Apply stop lists, Characterize Tokens, Apply stemming and create searchable Data structure is all part of the Indexing Process

Data flow in Information Processing System





- All Systems go through an Initial stage of Zoning and Identifying the processing tokens used to create the Index
- Filters, such as Stop lists and stemming Algorithms are frequently applied to reduce the number of Tokens to be processed
- The Next step depends up on the search strategy of a particular system
- The search strategies can be classified as statistical natural Language &Concept
- An Index is the Data structure created to support the search strategy

Standardize Input:

- Standardizing the input takes the different external format of input data and performs the translation to the formats acceptable to the system.
- That particular formats System should be Acceptable

Logical Sub-setting (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

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 (-)

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

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 :

Stemming is cutting &Trimming of words

- It has maintaining as systematic arrangement of words
- It is monitoring as phrases, grammatically errors …etc

Create Searchable Data Structure:

- It has in-Built of files
- It has searchable of Data structure
- It is internal representation of user(not visible to user)
- It has contains Semantic concepts represent the items in database Limit what a user can find as a result of the search

List of the classes of Automatic Indexing:

- Statistical Indexing
- Natural Language Processing
- Concept Indexing
- Hypertext Indexing
- An index is the data structure created to support the search strategy
- The statistical strategies cover the broadest range of Indexing Techniques & most prevalent in commercial systems

Statistical Indexing:

The Statistical Indexing uses frequency of occurrence of events to calculate a number to Indicates relevance of an Item

- This is to assist in calculating a relevance value of each item for ranking
- The Documents are found by a normal Boolean search and the Statistical Calculations are performed on the Hit File ranking the output (eg:-Term frequency Algorithms)

There are two types of frequencies

- Document Frequency
- Term frequency

Document Frequency:

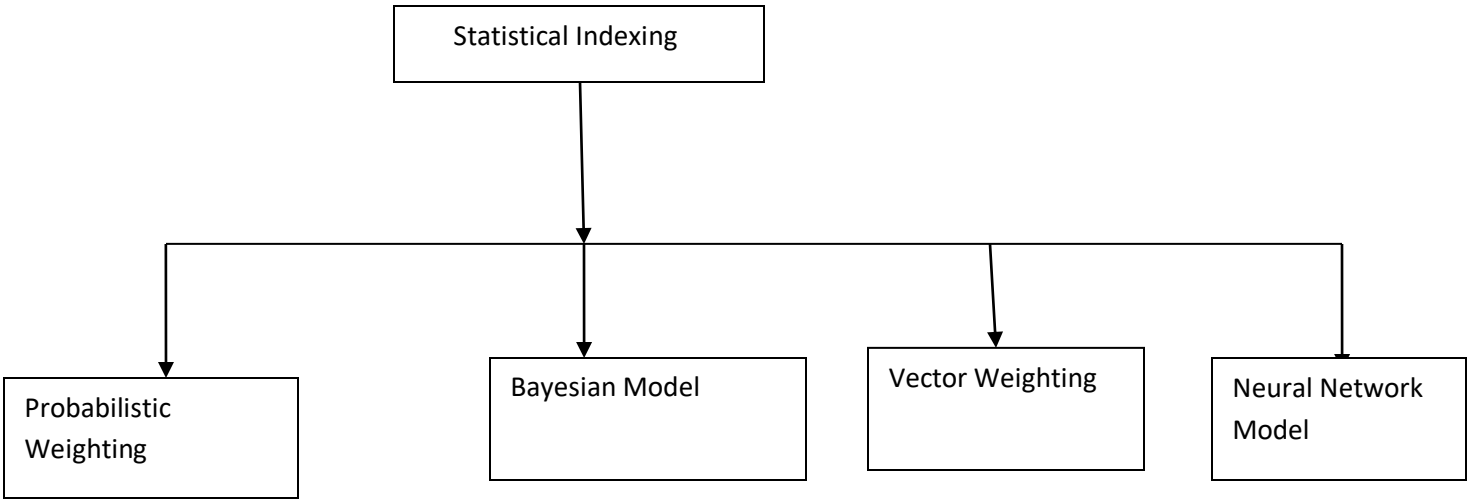
it is identifies as particular documents items is existing or not

Term Frequency:

It is identifies number of times of lines or words ,it is calculating as terms

These calculations are performed by Hit file Document frequency &term frequency producing based on the Ranking Output

- Statistical strategies cover the broadest range of Indexing techniques and the most prevalent in commercial systems
- The basis for a statistical approach is use of frequency of occurrence of events
- the events usually are related to occurrences f processing tokens (words/Phrases)within the documents and within the database
- The words/phrases are the Domain of searchable values
- The static approach stores a single statistic, such a how often each words occurs in an item that is used in generating relevance scores after standard Boolean Search



Probabilistic Weighing:

The Probabilistic approach is based up on direct Application of the theory& Probability to Information Retrieval Systems

- This has the advantage of being able to use the developed formal theory of probability to direct the algorithmic development
- This is summarized by the probability Ranking Principle(PRP)
- It stores the information that are used in calculating a probability that a particular Item satisfies in relevant to a particular Query
- We are going to apply two types of techniques
- HYPOTHESIS
- PLAUSIBLE COROLLARY

HYPOTHESIS:

It contains the collection of list of words with rankings based on occurrences

PLAUSIBLE COROLLARY:

The most promising source of techniques for estimating the probabilities of

usefulness for output ranking in IR Standard Probability Theory & Statistics

- It also leads to an in variant result that facilitates Integration of results from different Databases
  - It can be represented as the following equation

  Log(O(R| Qi,Dj,t k))=C0+C1 V1 +……..+CnVn

- The Log O is the logarithm of the odds(log odds) of relevance for term which is present in Document &Query
- The Logarithm that the query is relevant to the Document is the sum of the log odds for all terms

Bayesian model:

For Overcoming the restrictions in a vector model is to use Bayesian Approach to Maintaining Information on processing Tokens

- The Bayesian model provides a conceptually simple yet complete model for Information Systems
- The Bayesian approach is based up on Conditional Probabilities
- Bayesian approach stress Information used in generating a relative Confidence level of an Items relevance to a query
- It produces a good relative relevance value than producing and absolute probability

Vector Weighting:

One of the earliest using Statistical approaches in Information retrieval was the smart system at Cornell University

- It is implemented this vector weighting
- The system is based upon a vector Model
- The semantics of every item are represented as a vector

- A Vector is a One-dimensional set of values where the Order/Position of each value in the set is fixed and represents a particular Domain
- In Information retrieval each position in the vector typically represents a Processing token
- There are two approaches to the Domain of value in the Vector
    1. Binary
    2. Weighted
- In the Binary Approach, the domain contains the value of one or zero with one representing the existence of the processing token in the Item
- In the weighted approach, the Domain is typically the set of all Real positive Numbers

Neural Network Model:

- The Neural Networks are dynamic learning structures under concept Indexing where they are used to determine concept classes
- Improve Recall
- Concepts classes hierarchies and domain specific systems are best Examples

Natural Language:

The natural Language is nothing but User has giving emotions as well as user Expressing his views

Eg: "Happy or sad Good or Bad"

- The goal of Natural Language processing is to use the semantic Information(semantic information means semantic Analaysis of Natural Language captures the meaning of the given text while taking into account context Logical Structuring of sentences &grammar roles) in addition to the statistical Information to enhance the indexing of the Item
- This Improves the Precision of searches, reducing the number of False hits a user reviews
- The semantic Information is extracted as a result of processing the language rather than treating each word as an Independent Entity
- The simplest output of this process results in generation of Phrases that becomes Indexes to an Item
- Statistical approaches use Proximity as the Basis behind determining strength of word relationships in generating phrases
- Natural Language processing can also combine the concepts into higher level concepts sometimes referred to as thematic representations
- The goal of Indexing is to represent the semantic concepts of an Item in the Information system to support finding relevant Information
- Term Phrases allow additional specification and focusing of the concept to provide better Precision &reduce the user overhead of retrieving non-relevant Items
- One of the earliest Statistical Approaches to determing term phases was use of cohesion factor between terms

  Cohesion=Size-factor*(PAIR-FREQkh/TOTFk*TOTFh)

  - Size factor is a normalization factor based up on the size of the vocabulary

- PAIR-FREQ k, h is the total frequency of Co-Occurrence of the pair Term k, Term h in the Items Collection
- Natural Language processing can reduce errors in determining Inter-Item dependencies and using that Information to create the term phrases used in the Indexing Process

Concept Indexing:

- concept Indexing uses the words within an Item to correlate to concepts discussed in the Item
- this is a generalization of the specific words to values used to Index the Item
- when generating the concepts classes Automatically, There may not be a name Applicable to the concept but just a statistical Significance
- Recall is Improved
- It can be used with concept classes using neural networks
- An Example of applying a concept approach is the convection system
- The convection system uses neural network algorithm(A neural network is a method in Artificial Intelligence that teaches computers to process data in a way that is inspired by the human brain)
- The convection system uses neural network algorithms and terms in a similar context of other terms
- The process of mapping from a specific term to a concept that the term represents is complex because a term may represent Multiple different concepts to different degrees
- The basis behind the generation of the concept approach is a neural network model
- The convections system uses neural network Algorithm &terms

Hypertext Linkage Indexing:

The Hypertext is a Data structures are generated Manually

- Hypertext is using in Information Retrieval systems purpose, it is also comes under storing & Retrieving of a Data
- If user is using a page it will be navigate to another page
- Hypertext Linkages are creating an additional Information Retrieval Dimension
- Traditional Items can be viewed as two Dimensional Constructs
- The text of the items is one dimension representing the items
- The internet at the current Time there are three classes of mechanisms to help find the Information
- Manually generated Indexes
- Automatically generated Indexes and web crawlers(A web crawler called a Spider-bot is an Internet-bot that systematically browses the worldwide web and that is typically operated by search engines for the purpose of web Indexing (Intelligent agents)
- It is a special class of indexing can be defined by creation of hypertext Linkages
- These linkages provide virtual threads of concepts between Items versus directly defining the concept within an Item.

Introduction of Clustering:

Clustering is used in information Retrieval systems to enhance the efficiency and effectiveness of the retrieval process, Clustering is achieved by partitioning the Documents in a collection into classes such that Documents that are Associated with each other are assigned to the same cluster

- Clusters it is provide a grouping of similar objects into a class under a more general title
- Clustering also allows linkage between clusters to be specified
- An Information Database can be viewed as being composed of a number of Independent Items Indexed by a series of Index terms
- Adding some grouping of objects
- Clustering in IRS is of two Types
    - Term clustering
    - Document Clustering

Term Clustering:

A Term may be word or group of words or a single paragraph it will be called as term

- It is used to create a statistical Thesaurus(Thesaurus coming from the Latin word meaning “treasure” is similar to a dictionary in that it store words)
- Increase recall by expanding searches with related terms(Query Expansion)

Documents Clustering:

- The Document clustering is nothing but we are going to clusters the Documents what are he terms is there on what are the terms Existing in number of Documents
- Used to create documents clusters

- The search can retrieve items similar to an Item of Interest, even if the query would not have retrieved the item (Resultant set Expansion)
- Result-set clustering

Define the Domain for Clustering

Thesaurus: The Domain may be medical or education

It should be relevance of similar of terms Documents set of Items to be clustered Identify those objects to be used in the clustering process and reduce the potential data that could Induce errors in the clustering Process

Determine the attributes of the objects to be clustered

Thesaurus: To determine the specific words in the objects to be used

Documents: May focus on specific zone within the items that are used to determine similarity

Reduce Errors Association

Determine the Relationships between the attributes whose co-occurrence is objects suggest those objects should be in the same class

Thesaurus: Determine which words are synonyms and the strength of their relationships

Documents:-Define a similarity function based on word Co-occurrences that determine the similarity between two times

Apply some algorithm to determine the classes to which each object will be assigned

Guide lines on the characteristics of the Classes in Clustering

A well-defined semantic definition should exit for each class

There is a risk that the name assigned to the semantic definition of the class could also be misleading

- The size of the classes should be within the same order of Magnitude
  Within a class, one object should not dominate the class
- Whether an object can be assigned to multiple classes or just one must be decided at creation name

Additional Decisions for Theasurus

Word coordination approach:-specify If Phrases as well as Individual Terms are t be clustered

Word Relationships:-

Equivalence, Hierarchical, Non-Hierarchical

Parts-Wholes:-Aggregation and composition

Collocation:- Statistical Measures that relates words that co-occur in the Same(Sentence, Phrase, Paragraph)

Paradigmatic T:-Paradigmatic relates words with the same semantic base such as ”Formula” & “equation , Axonomy &Synonym

Thesaurus Generation:

The Collection of terms can be generated or Cluster it can done Manually or Automatical

- Automatically generated Thesauri contain classes that reflect the use of words
- The classes do not naturally have a name, but are just a groups of Statistically Similar Term Clustering

- The more Frequently two terms Co-occur in the Same Items, the more Likely they are about the same concept
- Each and every items it has to identified number of repeat times as well as possible location of Documents

Thesaurus Generation(Manual Clustering):

- A keyword out of context(KWOC)is used to represent frequency of Items in Respective Documents
- Keyword in Context(KWIC)displays a possible term in its Phrase Context
- It is Structured to Identify easily the Location of the term under consideration in the Sentence
- Keyword and Context(KWAC)displays the keywords followed by their context
- One sentence it will be four number of words

EX:

KWOC

Term Ids	**Documents**
1	Chips, Computer, Memory
2	Memory, Design
3	Computer, Chips, Design
4	Memory, Chips, Computer

TERM	FREQUENCY	ITEM IDS
Chips	3	DOC1,DOC3, DOC4
Computer	3	DOC1,DOC3, DOC4
Design	2	DOC2,DOC3
Memory	3	DOC1,DOC2, DOC3

KWOC

TERM	FREQ	ITEM Ids
chips	2	doc2, doc4
computer	3	doc1, doc4, doc10
design	1	doc4
memory	3	doc3, doc4, doc8, doc12

KWIC

chips/	computer design contains memory
computer	design contains memory chips/
design	contains memory chips/ computer
memory	chips/ computer design contains

KWAC

chips	computer design contains memory chips
computer	computer design contains memory chips
design	computer design contains memory chips
memory	computer design contains memory chips

Figure 6.1 Example of KWOC, KWIC and KWAC

In the Figure 6.1 the character "/" is used in KWIC to indicate the end of the phrase. The KWIC and KWAC are useful in determining the meaning of homographs.

Once the terms are selected they are clustered based upon the word relationship guidelines and the interpretation of the strength of the relationship. This is also part of the art of manual creation of the thesaurus, using the judgment of the human analyst

Automatic Term Clustering:

There are many techniques for the automatic generation of term clusters to create statistical thesauri. When the number of clusters created is very large

- The basis for automatic generation of a Thesaurus is set of Items that represents the Vocabulary to be Included in the Thesaurus

- The processing tokens(words)in the set of Items are the attributes to be used to create the clusters
- The Automated Method of clustering Documents is based up on the Clustering, where each cluster is defined by set of words &Phrases
- They all use as their basis of the concept that more Frequently two terms co-occur in the same Items, the more likely they are about the same concept
- They differ by the completeness with which terms are correlated

Item Clustering:

The Clustering of various kinds of items in Multiple number of Documents, Item it may be Phrase, word, collection of words Sentence, Diagram or a Picture

- Item Clustering can be done two ways
- Manual term clustering as well Automatic Term clustering
- In Manual term Clustering Requires large space time& Computational Overhead
- In Automatic term clustering –one Primary Category &Several Secondary Categories
- It is very Efficient
- It is same as term clustering
- It is also same as Term Complete relation Method here we Implement Item Complex Relation Method
- Similarity between Documents is based on two Items that have terms in common
- The similarity Function is performed between rows of the Item Matrix
- Based on the threshold value binary Item Matrix is Calculated
- Default Threshold Value is 10
  SIM (Item i, Item j)=€(term i, k)(term j, k)

Ex: 10 is greater than are equal

Item &Item relationship Matrix

Item Id's	Item1	Item2	Item3	Item4	Item5
1		11	3	6	22
2	11		12	10	36
3	3	12		6	9
4	6	10	6		11
5	22	36	9	11	

The based on numbers 10 is greater than are equal to zero it is going converting as one's &zero's greater than are equal to 10 it should be as indicates '1', less than 10 it should be Indicates as '0'

Item id's	Item1	Item2	Item3	Item4	Item5
1		1	0	0	1
2	1		1	1	1
3	0	1		0	0
4	0	1	0		1
5	1	1	0		

HIERARCHY OF CLUSTERS

Hierarchy Clustering:

The Hierarchy is defined as set of clustering items that clustered arranged in hierarchy manner that means Tree Manner root will be there second level of the elements it will be arranged in Hierarchal manner later remaining Elements are arranged in Hierarchal manner
Hierarchical clustering can be divided into two types
1. Hierarchical Agglomerative Clustering(HAC)
2. Hierarchical divisive clustering

Hierarchical Agglomerative clustering(HAC):-

The start with un-clustered items and perform pair-wise similarity measures to determine the clusters Hierarchical or(it is often treated as making Clusters which are Un clusters they can be clusters based on similarity, They can be arranged Tree manner &hierarchal Structure)

Hierarchical divisive clustering:-

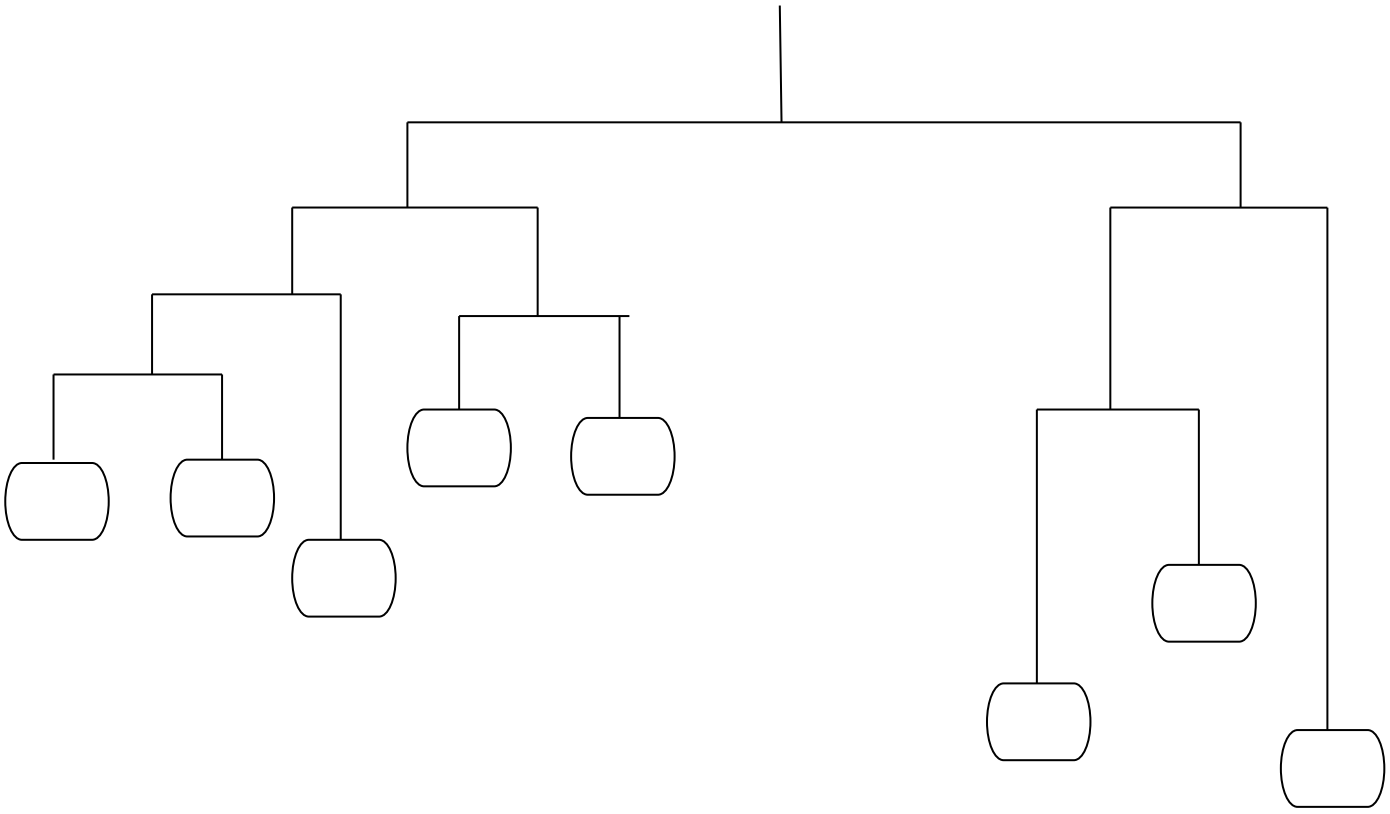
We start with Large cluster and we breaking it down into smaller cluster

Objectives of creating a Hierarchy of Clusters:-

- It Reduce the overhead of search
- It perform top-down searches of the centroid of the clusters in the hierarchy & trim those branches that are not relevant
- It is provide for visual representation of Information space
- Visual cues on the size of clusters and strengths of the linkage between clusters
- Expand the retrieval of relevant Items
- User once having Identified an item of Interest can request to see other items in the cluster

- The user can Increase the specificity of Items by going to children clusters or by Increasing the generality of Items being reviewed by going to parent clusters

Dendogram for Visualizing Hierarchical Clusters



This figure about of Structure of Hierarchy clustering

UNIT-IV

Search Statements&Binding:

The Search statements are the statements of an Information and generated by users to specify the concepts, they are trying to locate in items

- It uses traditional Boolean logic &or Natural Language
- While generating the search statement, the user have the ability to weight or assign an Important to different concepts in the statements
- At this point the binding is to the vocabulary& past experiences of the user
- The search statement is the users attempt to specify the conditions needed to subset logically the total Item Space to that cluster of Items that contains the information needed by the user
- The next level of binding comes when the search statements is parsed for use by a specific search system
- The search System translates the query to its own Meta Language
- This process is similar to the Indexing of Item processes

  Ex:-

  Statistical systems determine the processing tokens of Interest and the weights assigned to each processing token based up on the frequency of occurrence from the Search Statement
  - Natural Language Systems determine the Syntactical and Discourse Semantics using Algorithms Similar to those used in Indexing
  - Concept Systems map the search statements to the set of concepts used to Index Items
  - Some of the Statistics used in weighing are based upon the current contents of the Database
  - Some examples are Documents frequency &Total Frequency for a Specific term
  - Parenthesis are used in the second binding step to Indicate Expansion by a Thesaurus

INPUT	Binding
"Find me information on the impact of the oil spills in Alaska on the price of oil"	User search statement using vocabulary of user
impact, oil (petroleum), spills (accidents), Alaska, price (cost, value)	Statistical system binding extracts processing tokens
impact (.308), oil (.606), petroleum (.65), spills (.12), accidents (.23), Alaska (.45), price (.16), cost (.25), value (.10)	Weights assigned to search terms based upon inverse document frequency algorithm and database

Figure 7.1 Examples of Query Binding

The length of search Statements directly affect the ability of Information Retrieval Systems to find relevant Items

The longer the search query, the easier it is for the system to find items

Similarity Measures and Ranking

The Searching is concerned with calculating the Similarity between a users Search Statement and the items in the Database

- Restricting the similarity Measure to Passages gains Significant Precision with Impact on recall
- Once Items are Identified as possibly relevant to the users query, It is best to present the most likely relevant items first Ranking is a Scalar number that represents how similar an Item is to the Query
- Searching is Concerned with calculating the Similarity between a users search statement and the items in the Database
- Variety of different similarity measures can be used to calculate the similarity the item and the search Statement

- A Characteristic of a similarity formula is that the results of the Formula Increase

  Increases as the Items become More Similar
  - The value is Zero if the Items are totally Dissimilar
  - SIM(Item i, Itemj)=z(Termix)(Termja)
  - This formula uses the summation of the product of the various terms of two Items when treating the Index as a vector
  - The problem with the simple Measures in the normalization needed to account for variances in the length of Items

    Similarity Formula by Salton in SMART System

To determine the weight an Item has with respect to the Search Statement, The Cosine Formula is used to calculate the distance between the vector for the Item and the vector for the Query

$$SIM(DOC_i, QUERY_j) = \frac{\sum_{k=1}^{n} (DOC_{i,k} * QTERM_{j,k})}{\sqrt{\sum_{k=1}^{n} (DOC_{i,k})^2 * \sum_{k=1}^{n} (QTERM_{j,k})^2}}$$

The DICE Formula

The Measure Simplifies the denominator from the Jaccard Formula and Introduces a factor of 2 in the Numerator

- The Normalization in the Dice Formula is also Invariant to the number of terms in common

$$SIM(DOC_i, QUERY_j) = \frac{2 * \sum_{k=1}^{n} (DOC_{i,k} * QTERM_{j,k})}{\sum_{k=1}^{n} DOC_{i,k} + \sum_{k=1}^{n} QTERM_{j,k}}$$

The use of a similarity Algorithm returns the complete Database as Search Results

Many of items have Similarity Close or equal to zero

- The Threshold defines the Items in the resultant Hit File from the Query

Query Threshold Process:

Vector:- American, geography, Lakes, Mexico, Painter, oil, Reserve, Subject

DOC1:- geography of Mexico suggests oil reserves are available Vector(0,1,0,2,0,3,1,0)

DOC2:- American Geography has takes available everywhere vector(1,3,0,0,0,0,0,0)

DOC3:-Painters Suggest Mexico Lakes as Subjects Vector(0,0,1,3,3,0,0,2)

Based on Existing of Occurrence &Based frequency Occurrence numberical values

Relevance Feedback:

The Thesaurus and Semantic networks provide Utility in generally Expanding a users search statement to Include Potential related Search terms

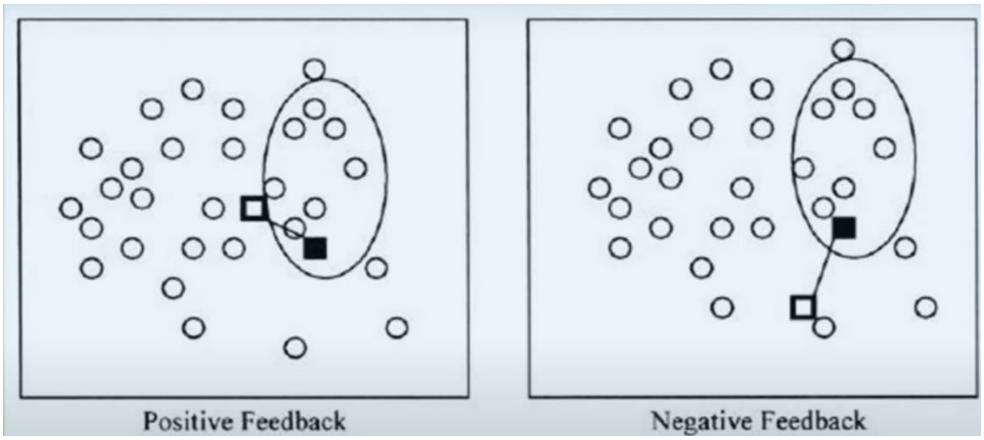
- But this still does not Correlate to the vocabulary used by the authors that contributes to a particular Database
- There is also a significant risk that the Thesaurus does not Include the latest words being used
- In an Interactive System, users can Manually Modify an Inefficient Query as well as the system Automatically Expand the Query Via Thesaurus
- The Relevant Items are used to reweight the Existing Query terms and possible Expand the users Search Statement with terms
- The relevance feedback concept was that new Query Should be based on the old Query Modified to Increase the weight of terms in relevant Items and decrease the weight of terms that are in non-relevant Items

The formula used for Relevance Feedback

$$Q_n = Q_o + \frac{1}{r}\sum_{i=1}^{r} DR_i - \frac{1}{nr}\sum_{j=1}^{nr} DNR_j$$

where

Q_n	= the revised vector for the new query
Q_o	= the original query
r	= number of relevant items
DR_i	= the vectors for the relevant items
nr	= number of non-relevant items
DNR_j	= the vectors for the non-relevant items.



Impact of Relevance Feedback

Selective Dissemination of Information Search:

The Selective Dissemination of Information, frequently Called dissemination Systems are becoming more prevalent with the growth of Internet

- A Dissemination system is sometimes called as Push Pull system
- Push is nothing but whenever you search statement ,Pull is nothing but Extract on Information based on Statement(or)back to the Information
- In a dissemination system, the user defines a profile and as new Information is added to the system, it is Automatically compared to the user's Profile
- If it is considered a Match, it is asynchronously sent to the user mail file
- The difference between the two functions lie in the Dynamic nature of the profiling process, The size and diversity of the search statements and the number of Simultaneous searches per Item
- In the Search System an Existing Database Exists
- These can be used for weighting factors in the Indexing process and the similarity Comparison
- Profiles are relatively static search statements that cover a diversity of topics
- One of the first commercial search Techniques for dissemination was the logic on Message Dissemination System(LMDS)
- The system originated from a system created by chase, Rosen and Wallace(CRW)
- It was designed for speed to support the search of Thousands of profiles with the items arriving every 20 seconds
- Another approach to dissemination uses a Statistical Classification Technique and Explicit Error Minimization to determine the decision criteria for selecting items for a particular profile

Weighted Searches of Boolean Systems:

The two major approaches to generating Queries are Boolean and Natural Language Queries are easily represented within statistical Models and are usable by the Similarity Measures Discussed

- The Issues arise when Boolean Queries are Associated with weighted Index Systems
- Some of the Issues are Associated with how the logic (AND,OR, NOT) Operators Function with Weighted values and How weights are Associated with the Query terms
- If the operators are Interpreted in their normal Interpretation, they act too restrictive or too general AND and OR Operators Respectively
- Closely related to the strict definition problem is the lack of ranking that is missing from a pure Boolean Process
- Some of the early work addressing this problem recognized the Fuzziness associated with mixing Boolean and weighted systems(Brookstein-78,Brookstein-80)
- To Integrate the Boolean and weighted systems Model (Fuzzy set=Fuzzy sets are sets whose Elements have degrees of membership
- Fuzzy sets were Introduced "Zadeh" Fuzzylogic was Introduced with 1965 Mathematician
- The degree of Membership for AND and OR operations are defined

$$DEG_{A\cap B} = \min(DEG_A, DEG_B)$$

$$DEG_{A\cup B} = \max(DEG_A, DEG_B)$$

$$SIM(QUERY_{OR}, DOC) = C_{OR\,1} * \max(DOC1_1, DOC_2, \ldots, DOC_n) + C_{OR2} * \min(DOC_1, DOC_2, \ldots, DOC_n)$$

$$SIM(QUERY_{AND}, DOC) = C_{AND1} * \min(DOC_1, DOC_2, \ldots, DOC_n) + C_{AND2} * \max(DOC1_1, DOC_2, \ldots, DOC_n)$$

Considering all item weighs versus the Maximum/Minimum approach the Similarity Measure is Calculated as

$$SIM(QUERY\ DOC) = \sum_{i=1}^{n} r^{i-1} d_i \ / \ \sum_{i=1}^{n} r^{i-1}$$

$$Q_{OR} = (A_1,a_1) \text{ OR } (A_2,a_2) \text{ OR } \ldots \text{ OR } (A_n,a_n)$$

$$Q_{AND} = (A_1,a_1) \text{ AND } (A_2,a_2) \text{ AND } \ldots \text{ AND } (A_n,a_n)$$

Searching the Internet & Hypertext

The Internet has multiple different mechanisms that are the basis for searching of items. The primary techniques are associated with servers on the Internet that create indexes of items on the Internet and allow search of them. Some of the most commonly used nodes are YAHOO, AltaVista and Lycos. In all of these systems there are active processes that visit a large number of Internet sites and retrieve textual data which they index.

The primary design decisions are on the level to which they retrieve data and their general philosophy on user access. LYCOS (http://www.lycos.com) and AltaVista automatically go out to other Internet sites and return the text at the sites for automatic indexing (http://www.altavista.digital.com). Lycos returns home pages from each site for automatic indexing while Altavista indexes all of the text at a site.

- The retrieved text is then used to create an index to the source items storing the Uniform Resource Locator (URL) to provide to the user to retrieve an item. All of the systems use some form of ranking algorithm to assist in display of the retrieved items. The algorithm is kept relatively simple using statistical information on the occurrence of words within the retrieved text
- Closely associated with the creation of the indexes is the technique for accessing nodes on the Internet to locate text to be indexed. This search process is also directly available to users via Intelligent Agents. Intelligent Agents provide the capability for a user to specify an information, need which will be used by the Intelligent Agent as it independently moves between Internet sites locating information of interest.

  There are six key characteristics of intelligent agents

Autonomy:

The search agent must be able to operate without interaction with a human agent. It must have control over its own internal states and make independent decisions. This implies a search capability to traverse information sites based upon pre-established criteria collecting potentially relevant information.

Communications Ability:

The agent must be able to communicate with the information sites as it traverses them. This implies a universally accepted language defining the external interfaces

Capacity for Cooperation:

This concept suggests that intelligent agents need to cooperate to perform mutually beneficial tasks.

Capacity for Reasoning:

There are three types of reasoning scenarios

Rule-based - where user has defined a set of conditions and actions to be taken

Knowledge-based - where the intelligent agents have stored previous conditions and actions taken which are used to deduce future actions

Artificial evolution based - where intelligent agents spawn new agents with higher logic capability to perform its objectives.

Adaptive Behavior - closely tied to 1 and 4 , adaptive behavior permits the intelligent agent to assess its current state and make decisions on the actions it should take

Trustworthiness- The user must trust that the intelligent agent will act on the user's behalf to locate information that the user has access to and is relevant to the user.

Introduction of Information Visualization:

The Visualization is the transformation of Information into a visual form which enables the user to observe and Understand the Information

- The functions that are available with Electronic display and Visualization of Data

- Modify representations of Data and Information or Display Conditions(Changing colors scales)
- Use the same representation while showing changes in data
- Animate the display to show changes in space &Time
- Create Hyperlinks under user control to establish relationships between data
- The Information visualization addresses how the results of search may be optimally display to the users to facilitate their understanding of what the search has provided and their selection of most likely Items of Interest to read Cognitive(The action or process of acquiring Knowledge and understanding through Experience &The Sense)
- The Engineering derives design principles for Visualization techniques
- It is Attention memory and information processing of the human visual system

There are many areas that information visualization and presentation can help the user:

A. The reduce the amount of time to understand the results of a search and likely clusters of relevant information

B. yield information that comes from the relationships between items versus treating each item as independent
C. perform simple actions that produce sophisticated information search functions

- Visualization can be divided into two broad classes
    - Link Visualization
    - Attribute visualization

Link Visualization:

It displays relationships among items

Attribute Visualization:

It reveals Content relationships across Large Numbers of Items

Cognition &Preception

Cognition:- cognition means to Store the Information**(** The action or process of acquiring knowledge and understanding through Thought, Experience& the Senses)

Preception:- Preception means Receiving Information**(**the ability to see, hear, or become aware of something through the senses)

- The user Machine Interface has primarily focused on a paradigm of a typewriter
- As computers display became Man-Machine Interfaces focused on treating the display as an Extension of paper with the focus on consistency of operations
- The Advent of WLMP(Windows, Icons, Menus and Pointers)Interfaces and the Evolution of the user what is talking place in the computer Environment
- Extending the HCL(Human Computer Interface) to Improve the Information Flow. Thus reducing wasted user overhead in Locating needed Information
- Although the Major focus is on enhanced visualization of Information,other Senses are along being looked at for Future Interfaces
- The Audio Sense has been always been part of Simple Alerts in Computers
- The sounds are now being replaced by Speech in both Input& output Interfaces
- The Touch Senses is being addressed in the Experiments using Virtual Realty
- A Significant portion of the brain is devoted to vision and supports the Maximum Information transfer Function from the environment to a humanbeing

  Other Measures:

  - **Proximity-**Near by figures are grouped together
  - **Similarity-** Similar figures are grouped together
  - **Continunity**-figures are Interpreted as Smooth Continuous patterns rather than discontinuous Concatenations of Shapes
    (Eg:-Circle with its Diameter drawn is perceived as two continuous shapes, a circle& a Line, Versus two half Circles Concatenated together

Closure-gaps within a figure are filled into create a Whole(Eg:-using Dashed lines to represent a Square does not prevent understanding it a square)

Connectedness:- Uniform and Linked Spots,Lines or Areas are perceived as a Single Unit

- Shifting the Information Processing Load from Slower Cognitive processes to faster perceptual Systems Significantly Improves the Information carrying Interfaces between Humans &computers

Information Visualization Technologies

The main focus on Information Retrieval Systems are as follows

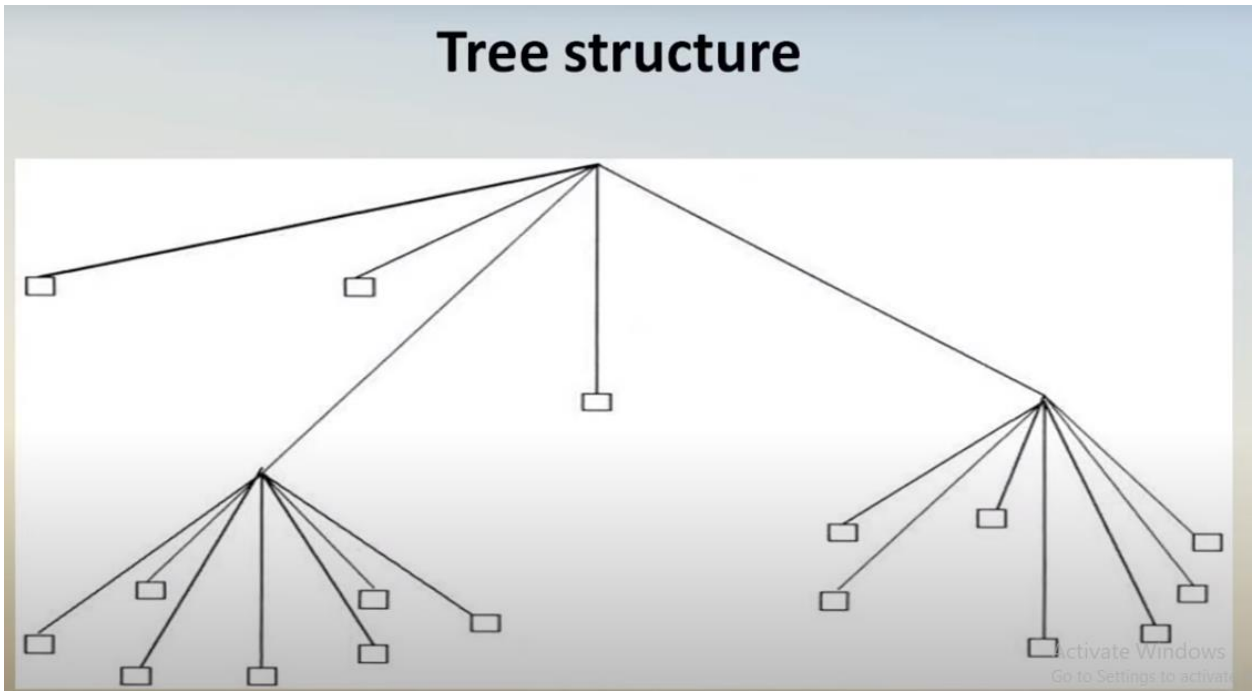
- The Investigating how to display the results of Searches Effectively, the structured Data from DBMS
- The Results of link analysis Correlating Data
- The goals of display the result in search is divided into 2 Major Classes
    - Document Clustering
    - Search Statement Analysis

- Visualization tools helps to display the Clusters in their Size &Topic to provide users to Navigate to items of Interest
- The Displaying the total set of terms
- Including additional terms from relevance feedback thesaurus Expansion Along with Documents retrieved and Indicate the Importance of the terms to the retrieval and Ranking process various information Visualization Technologies are as follows
    - Tree Structure
    - Cone-Tree
    - Perspective Wall
    - Tree Maps
    - Envision System
    - Document content Analysis &Retrieval System(DCARS)
    - City Space

Tree Structure:

A Tree Structure is useful in representing Information that ranges overtime

- The Constituents of a larger Unit(ex:- Organization Structures, Mechanical device Definitions
- The Aggregates from the higher to lower level(Eg: Hierarchical clustering of Documents)



CONE-TREE:

The Cone-Tree is a 3-Dimensional Representation of Data

Where one node of the tree is represented at the Apex

And all Information Subordinate to it is arranged in a Circular Structure at its base

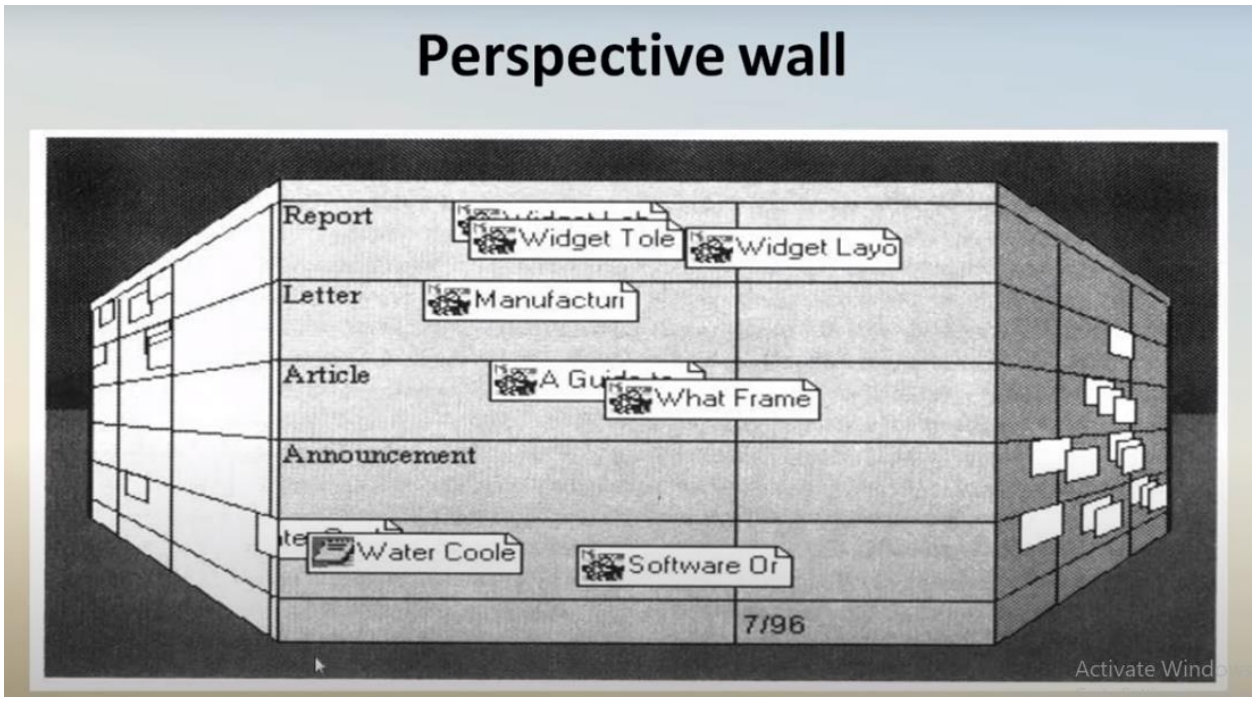
Any Child node may also be the Parent of another Cone

Selecting a particular node, rotates it to the front of the display

Perspective Wall:-

The perspective wall divides the Information into three Visual Areas

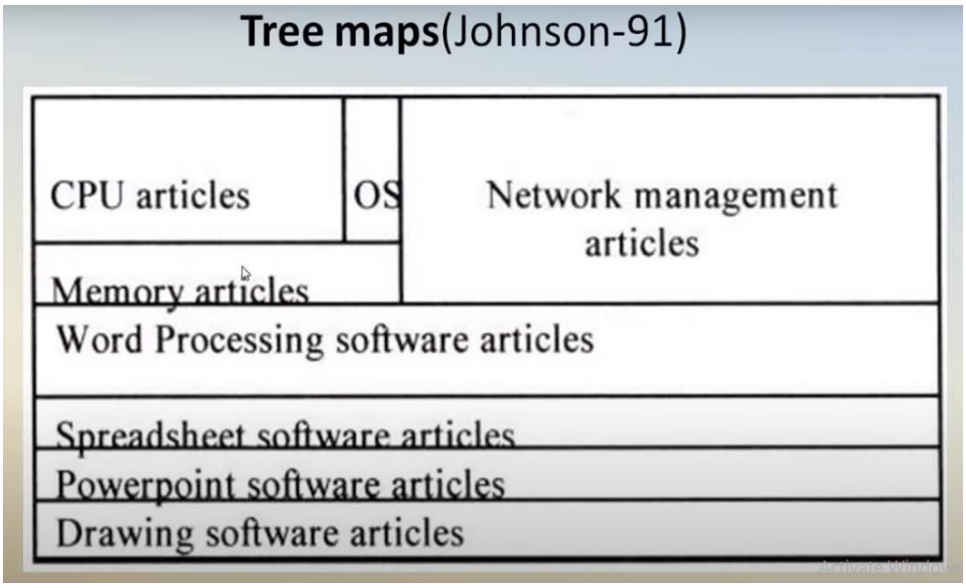
This allows the user to keep all of the Information in Perspective while Focusing on Particular Area



Tree Maps:-

This Technique Makes Maximum use of the display Screen Space by using Rectangular boxes that are Recursively subdivided and based on Parent-Child relationships between the Data

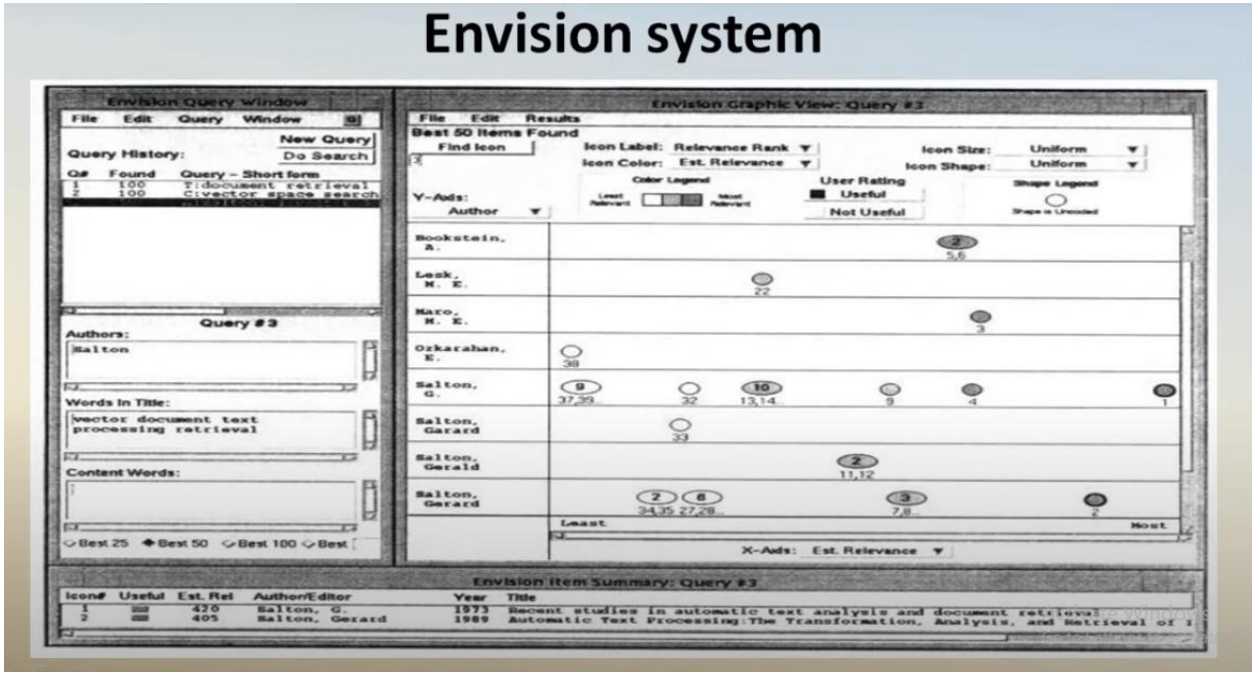
- The CPU
- OS
- Memory and
- The Network Management articles are all related to a general category of computer operating Systems
- The Computer Applications which are shown in the rest of the Figure



Envision System:-

The Envision System not only displays the relevance rank and estimated relevance of each Item found by a query, but also Simultaneously Presents other query Information

The design is Intentionally graphical and simple using two Dimensional Visualization

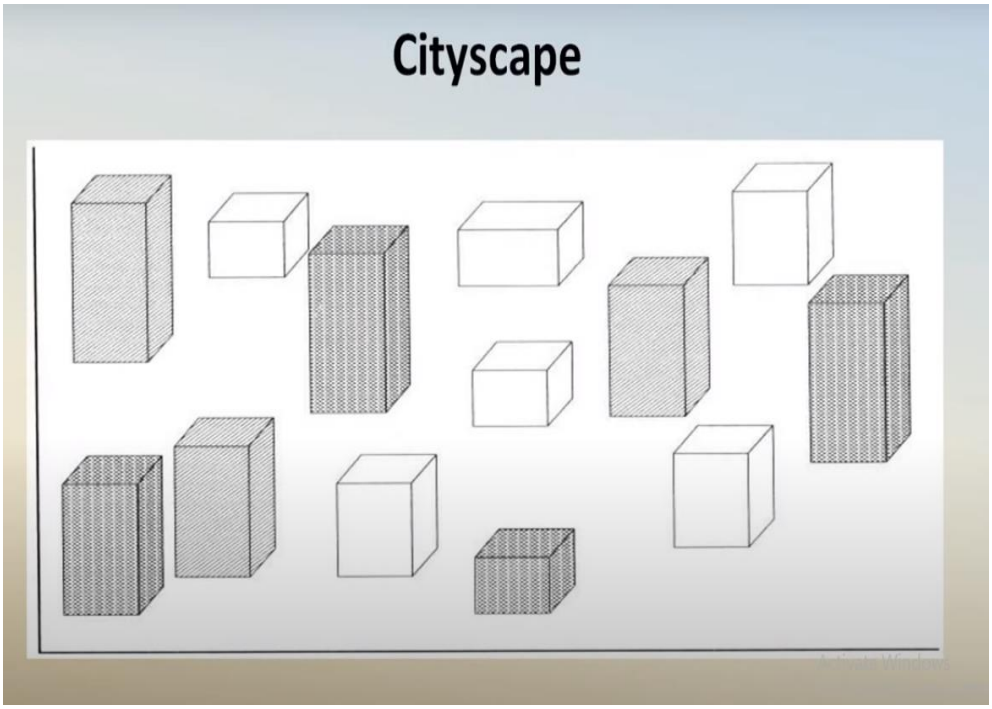


Document content Analysis &Retrieval System (DCARS):-

- The Document Content Analysis and Retrieval System DCARS being developed by Calspan Advanced Technology Center
- Their System is designed to Augment the Retrieval ware Search Product
- They display the query results as a histogram with the items as rows and each terms Contribution to the Selection Indicated by the Width of a Title bar on the row
- DCARS Provides a friendly user Interface that Indicates why a particular Item was Found but it is much harder to use the Information in determining how to modify Search Statements to Improve them

CITY SCAPE:

- This Representation is widely used for both hierarchical and Network related Information is the Cityscape which uses the Metaphor of Movement within a city
- In lieu of using hills as in the terrain approach Sky Scrapers Represent the theme Areas



UNIT-V

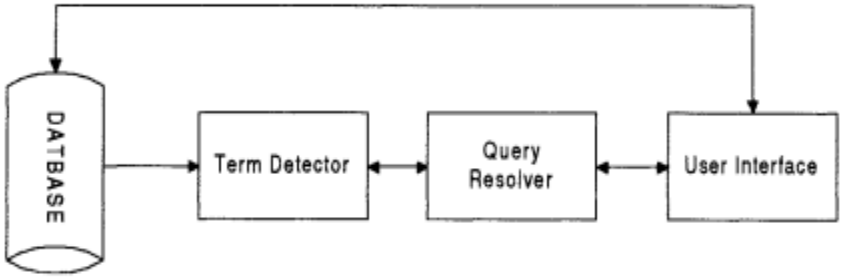
Introduction to Text Search Techniques:

The Basic Concept of a Text Scanning System is the ability for one or more users to enter Queries with the text of the Items to be Searched Sequentially and Compared to the Query Terms

When all of the text has been accessed the query is complete

One Advantage of this type Architecture is that as Soon as Item is Identified as satisfying a query the results can be presented to the user for retrieval

Text Streaming Search System Architecture



- The database contains the full text of the Items
- The Term Detector is the Special Hardware/Software that contains all of the search terms and in Some Systems the logic between the terms
- The outputs to the query resolver the detected terms
- The query Resolver performs two major functions accepting Search Statements from the users and Extracting the logic and search terms to pass to the detector
- It also accepts results from the detector and determines which queries are satisfied by the Item and possible the relevance weight Associated with Hit
- The query Resolver passes Information to the User Interface, allowing it to Continually update the search Status &on Request Retrieve any items that satisfy the user search statement
- The text streaming process is focused on finding at least one or all occurrences of a pattern of text in a Text Stream
- It is assumed that the same Alphabet is used in both Search terms and text being Streamed

Software Text Search Algorithms:

In Software Streaming Techniques, the items to be Searched is read into memory and then the algorithm is applied

There are three Major Algorithms Associated with Software Text Search

- Brute Force Approach
- Boyer-Moore
- Knuth-Morris-Pratt

Brute Force Approach:

This approach is the Simplest String Matching Algorithm

- The idea is to try and match the search String against the Input text
- If as soon as a Mis-Match is detected in the Comparison process, shift the Input text one position &Start the comparison process over
- The Expected number of comparisons when Searching an Input text string of N Characters for a pattern of M Character is

$$N_c = c/c\text{-}1(1\text{-}1/c^m)(n\text{-}m+1)+O(1)$$

Where is the expected number of Comparisons and C is the Size of the Alphabet for the text.T

The brute-force Pattern Matching Algorithm Compares

The pattern P with the text T for each possible shift of Relative to T until either

Brute-Force pattern Matching runs in time O(nm)

O=Order

Order of n*m

-T= aaa…….ah

-P=aaah

It is an algorithm is for Brute force Approach

Boyer Moore Algorithm:

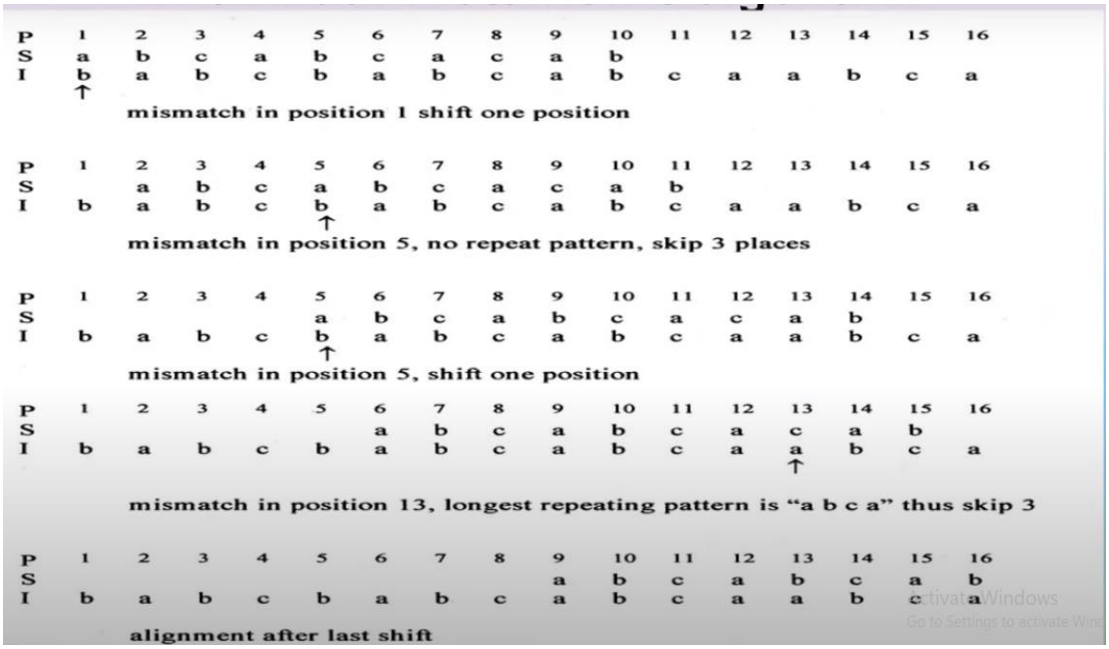
- It is a String Algorithm
- It is Significantly Enhanced as the Comparison process Started at the end of the search pattern processing right to Left versus the start of the search pattern
- The advantage is that large Jumps are Mis-Matched Character in the Input Stream the search Pattern which occurs frequently(one position to another position)
- The original Boyer-Moore Algorithm is developed for additional text search Techniques
- It was originally designed to support scanning for a single Search String
- It was Expanded to handle Multiple Search Strings on a Single Pass
- It enhanced and simplified versions of the Boyer-Moore Algorithm have been developed by May Researchers (Mollier-Nielson-84, Iyengar-80)
- The Boyer-Moore's Pattern Matching Algorithm is based on two Heuristics
- Looking-glass heuristic: Compare P with a Sub-Sequence of T Moving backwards
- Character-Jump Heuristic: when a Mis-Match occur at T[i]=C
- If P Contains C, Shift P to Align the last occurrence of C in P with T[i] Else Shift P to align P[O] with T[i+1]
- It is an Boyer-Moore Algorithm

Knuth-Morris Pratt Algorithm:

The Knuth Morris Pratt Algorithm made a major Improvement in Previous Algorithms

- It is an text search Algorithm
- Even in the worst Case it works well
- Unlike the previous Algorithms it does not Depend on the length of Input String
- It can also work as long Strings
- The basic Concept Behind the algorithm is that whenever a Mis Match is detected, the previous Matched Characters define the Number of Characters that can be Skipped in the Input Stream Prior to Process Again
- Starting The Comparison
- Position: 1 2 3 4 5 6 7 8
- Input Stream a b d a d e f g
- It is going to Identified as Position
  When the Mis-Match Occurs in Position 4 with a "F" in the pattern and a "b" in the Input Stream this Algorithm allows the Comparison to Jump at least the three Positions associated with recognized "abd"
    - Since the Mis Match on the position could be the beginning of the search strings four Positions can not be Skipped

- To Know the Number of Positions to Jump based up on Mis Match in the Search Pattern the Search Pattern is Pre-Processed to define a number of characters to be Jumped for Each Position
  P=Position
  S=Search Position
  I=Input Stream



Hardware Text Search Systems:

The Software text Search is applicable for many Situations but faced Some Restrictions to handle Many Search Terms Simultaneously against the same text and Limits due to I/O Speeds

- One approach is to have specialized hardware Machine to perform the searches &Pass the results to the Main Computer which supported the user Interface &Retrieval of Hits
- Since the searcher is hardware based Scalability is achieved by Increasing the number of hardware Search devices
- The only Limit on speed is the time it takes to flow the text off of Secondary Storage(Disk drives) to the Searchers
- By having one search Machine Per Disk, the maximum time it takes to search a Database of any size will be the time to search one disk
- In some systems, The Disks were formatted to optimize the data flow off the drives
- Another Major Advantage of Using a hardware Text Search unit is in the Elimination of the Index that represents the Document Database
- Typically the Indexes are 70% the size of the Actual Items

- Other Advantages are that new items can be Searched as soon as received by the System rather than waiting for the Index to be Created and the search Speed is Deterministic
- Even through it may be slower than using an Index the predictability of how long it will take to stream the data provides the user with an Extract Search Time
- As hits as Discovered they can Immediately be made Available to the user versus waiting for the total search to complete as in Index Searches
- One of the earliest hardware text string Search units was the Rapid Search Machine developed by General Electric
- A More Sophisticated Search units was developed by operating Systems INC Called the Associative File Processor(AFP)
- It is Capable of Searching Against Multiple Queries at the Same Time
- In this regard we are going to discuss two types of Hardware Text Search Systems

    - GESCAN TEXT ARRAY PROCESSOR
    - FAST DATA FINDER

GESCAN TEXT ARRAY PROCESSOR:

In Some of the Systems, the Boolean Logic between terms is resolved in the term Detector Hardware

Eg:- GESCAN MACHINE

- The GESCAN System uses a Text Array Processor that Simultaneously Matches Many terms &Conditions against a given Text Stream

FAST DATA FINDER(FDF):-

The Fast Data Finder(FDP) is the most recent Specialized hardware text search unit Still in Use in Many Organizations

- It was Developed to search Text and has been used to Search English &Foreign Languages
- The early Fast Data Finders Considered of an Array of Programmable Text Processing cells Connected in Series a Pipline Hardware Search Processor

MULTIMEDIA INFORMATION RETRIEVAL

Spoken Language Audio Retrieval:

The Speech Retrieval refers to the task of retrieving the specific pieces of Spoken Audio Data from Large Collection that pertain to a Query requested by a user Before discussing Methods for Speech retrieval it is Important to define the different types of speech retrieval tasks and the methods in which potential solutions to these tasks will be evaluated

- The speech Information Retrieval Applications the basic Scenario assumes that a user will provide a query and the system will return a list of rank ordered Documents
- It will be treat as voice of query
- It will be assigning as Ranking based Algorithm, How many no of Users given particular Query through voice or Text, no of Users are demanding particular Query , visited as site of Query related Information
- The query is generally assumed to be in the form of a string of Text-based words
- When the task requires the retrieval of Utterance Retrieval
- In this case, The purpose of SUR(Spoken Utterance Retrieval)utterance related to the query even when Multiple Utterances related to the Query Exist within a single longer Document Applications include browsing broadcast news,Voicemail,Teleconferences&Lectures

Non-Speech Audio Retrieval:

Non-Speech Audio one of the Categories of Audio Interface, can be described as Audio Cues that accompany Specific Events

- Audio Cues means performing of object action hearing as Audio
- These cues for centuries have given human feedback on Important Information
- With the best examples like Fire Alarm is modern day, aircraft Audio cues have given Significant Amounts of Information regarding the Action of objects
- In addition to Audio cues, Music like Instrumental audio or Music is considered Non-speech audio
- In this research Music is mentioned with regard to Industry standards &Developments, Therefore this project will be directed specifically towards Audio cues
- The Non-Speech Audio Input is the use of Non-Speech sounds Such as humming or hissing for Entering Data or Controlling the User Interface
- It will be needs for user interface ,user must be alert and Attention of the sounds like Notification, Receiving Mail ,system Updates

- He had to know indications based on performing as actions
- He had to know those system Instructions based on Sounds

Graph Retrieval:

The Graph based Information Retrieval System whose query can be Expressed as a graph of Topics/subtopics, Documents are ranked with respect to a query up on relationships among Documents, relationships among Topics/Sub topics& Relationship between query terms &Documents

- The relationship between one object to another object by using as Bar chart, Pie-chart like Pictorial Representation
- User should understandable very easy way
- The system is evaluated and compared with two Information retrieval Systems on two standard text collections
- It is also Ranking based Algorithm
- Example like No of products sales, No of products refunded, Students No of pass percentage of students &No of Fail Percentage of Students it will be done as Result Analysis
- The results show that the proposed approach outperforms the other systems
- The goal of Information retrieval is to effectively retrieve Documents relevant to users queries
- The graph based approach to information retrieval, its computation is fast &Scalable and its Structure is flexible to Incorporate
- Many Performance Enhancement Techniques
- It most useful for Result Analysis, user will understandable very easily

Imagery Retrieval:

The Image Retrieval System is used for retrieving Images related to the user request from the Database

In the Presented Image Retrieval System

- An Image Retrieval System is a computer System used for browsing, Searching &Retrieving Images from a large Database of Digital Images
- The most Traditional and common methods of Image retrieval Utilize Some method of Adding Meta Data, Such Captioning, Keywords, Title or descriptions to the Images So, that retrieval can be Performed over the Annotation Words
- The Manual Image Annotation is time-Consuming Laborious and Expensive to address this, There has been a large Amount of research done on Automatic Image Annotation
- Additionally, the Increase in social web applications and the Semantic Web have Inspired , The development of several web-based Image Annotation Tools
- The need to retrieve a desired Image from a collection and to efficiently access the Information is Shared by Many groups Including Journalists, Engineers. Designers, Artists, Advertising Agencies
- Image needs and uses across users in these groups vary considerably users many require Access to Image based on Primitive Features Such as Color, Texture or Shape or users may Require access to Images based on Abstract Concepts and Symbolic Imagery
- Content Based Image Retrieval(CBIR)Technology is now beginning to Move out of the Laboratory into the marketplace
- The technology still lacks Maturity and is not yet being used in a Significant Scale

Video Retrieval:

The video is an Electronic medium for the recording Copying, Playback, Broadcasting and Display of Moving Visual Media

- Video was first developed for Mechanical Television Systems, which were Quickly replaced by CRT(Cathod-Ray-Tube)Systems which in turn were replaced by Flat Panel display
- It is Interactive web based Application which takes Video Frame from users and retrieve the information from the Database
- The Database consists of various Video data like still Video Frames, Audio &Video
- During Recent years, Methods have been developed for retrieval of videos based on their Visual features Color, Texture, Shape, Motion &Spatial-Temporal Composition are the most Common Visual Features used in Similarity Match
- Additionally ,the increase in Social Media web applications and the Semantic web
- The videos is shared by many groups designers, as well as Social Media Platforms like Youtube,…etc