Unit-I

Definition of IRS.

IRS stands for Information Retrieval Systems.

Information: Data, Facts

Information is communicated (or) received knowledge conveying a particular facts.

Retrieval: To get and bring back i.e recover from storage.

System: A set of Principles or Procedures (or) an Organized Scheme.

* IRS has a capability of Representing, storing and maintaining the information such as audio, video files, text files, Documents etc.
* Goal of IRS is to provide the information needed to satisfy the user's question.
* The meaning of the term Information retrieval can be 'very broad'.
* The term IR (Information retrieval) was coined by kelvin Mooere in 1950. It Gained popularity in the research communication from 1961 onwards.
* It deals with the representation, storage, organization and access information items.
* The first IRS's originated with the need to organize information in central repositories.

e.g.: libraries.

* Information Retrieval (IR) is the formal study of efficient and effective ways to extract the right bit of information from a collection i.e. web is a special ~~case~~ case for Retrieving the information.

* The information in this context can be composed of
  - ⇒ text (including numerical data)
  - ⇒ images
  - ⇒ audio
  - ⇒ video and other multimedia objects.

* An Information Retrieval system consists of a "software program" that facilitates a user in finding the information that 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 a searchable media.

  eg: transcription of audio to text.

Objectives of information Retrieval System.

The General objective of an information retrieval system is to minimize the overhead of a user locating needed information.

overhead can be expressed as the time, a user spends in all of the steps leading to reading an item containing the needed information.

e.g.: * Query Generation
* Query Execution
* Scanning results of query to select items to read.
* reading non-relevant items.

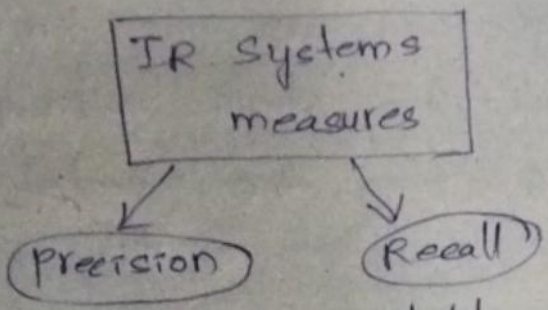
→ In Information retrieval the term "relevant" item is used to represent an item containing the needed information.

Let us see relevant in user's perspective and system perspective.

From User's perspective "relevant" and "needed" are synonyms.

From system perspective, information could be relevant to a search statement.
i.e. (matching the criteria of the search statement)

→ To minimize the overhead of a user locating needed information, there are two major measures commonly associated with information Systems are precision and Recall.

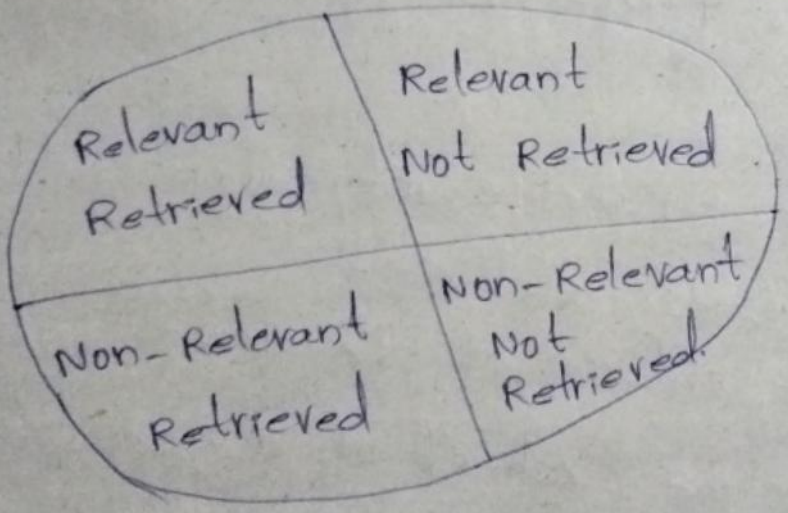


Precision: It is the ability to retrieve top-ranked documents that are mostly relevant.

Recall: The ability of the search to find all of the relevant items in the carpus.

carpus means collection of information stored in the system.

→ When a user decides to issue a search for information on a topic, the total database is logically divided into four ~~slateme~~ segments shown in fig.



lets see what is Relevant item & Non-relevant item.

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 2 possibilities with respect to each item:
It can be retrieved (or) not retrieved by the user's Query.
The precision and recall are defined as:

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

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

where,
* Number_Possible_Relevant are the number of relevant items in the database.
* Number_Total_Retrieved is the total number of items retrieved from the Query.
* Number_Retrieved_Relevant is the number of items retrieved that are relevant to the user's search need.

⇒ Precision measures one aspect of information retrieved overhead for a user associated with a particular search.
⇒ If a search has a 80% precision, then 15% of the user effort is overhead or, rerrewing non-relevant items.

Recall gauges how well a system processing a particular query is able to retrieve the relevant items that the user is interested. Let us see precision & Recall Graphs.

1. Precision & Recall

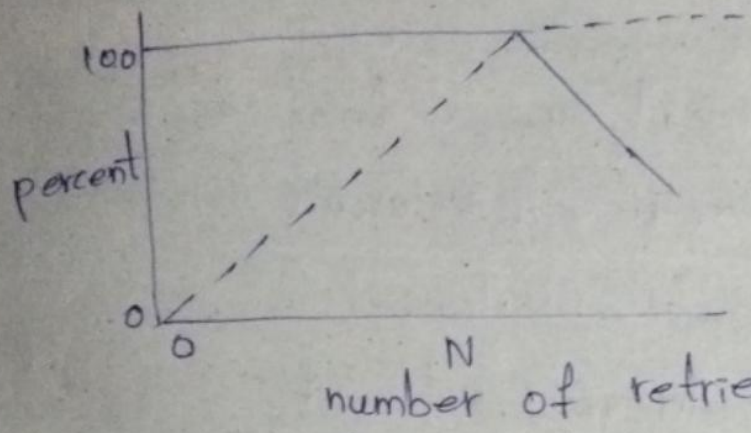


fig (a) precision & Recall graph

where N → relevant items in db

If we see the basic properties precision & recall.

In above graph,

* Solid line indicates precision
* dashed line indicates Recall.

Here, precision starts at 100 % and maintain that value as long as relevant items are retrieved

Recall starts off close to zero and increase as long as relevant items are retrieved.

2. Ideal precision / Recall Graph

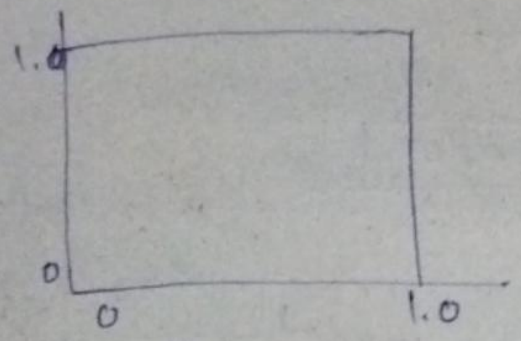


fig: b> Ideal precision / Recall graph.

3. Achievable precision / Recall Graph.

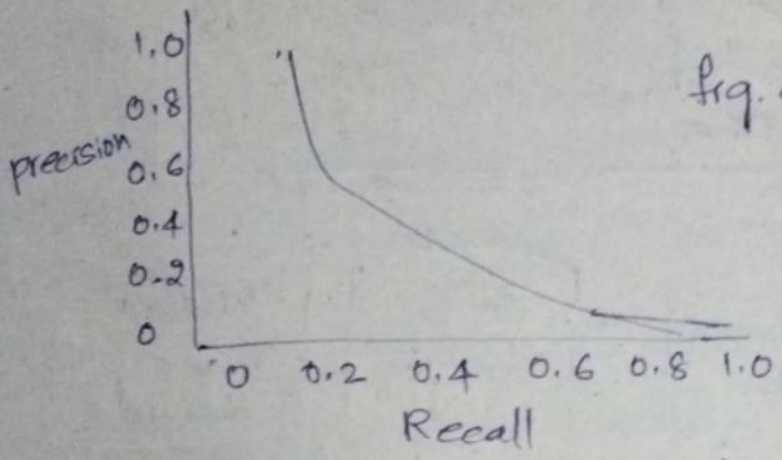


fig. c> Achievable precision/ Recall graph.

Here b & c shows the optimal & currently achievable relationships between precision and recall.

Functional Overview

Item Normalization	Selective Dissemination of information (i.e. mail)	Archival Document Database search and an index.

Functional overview of IRS

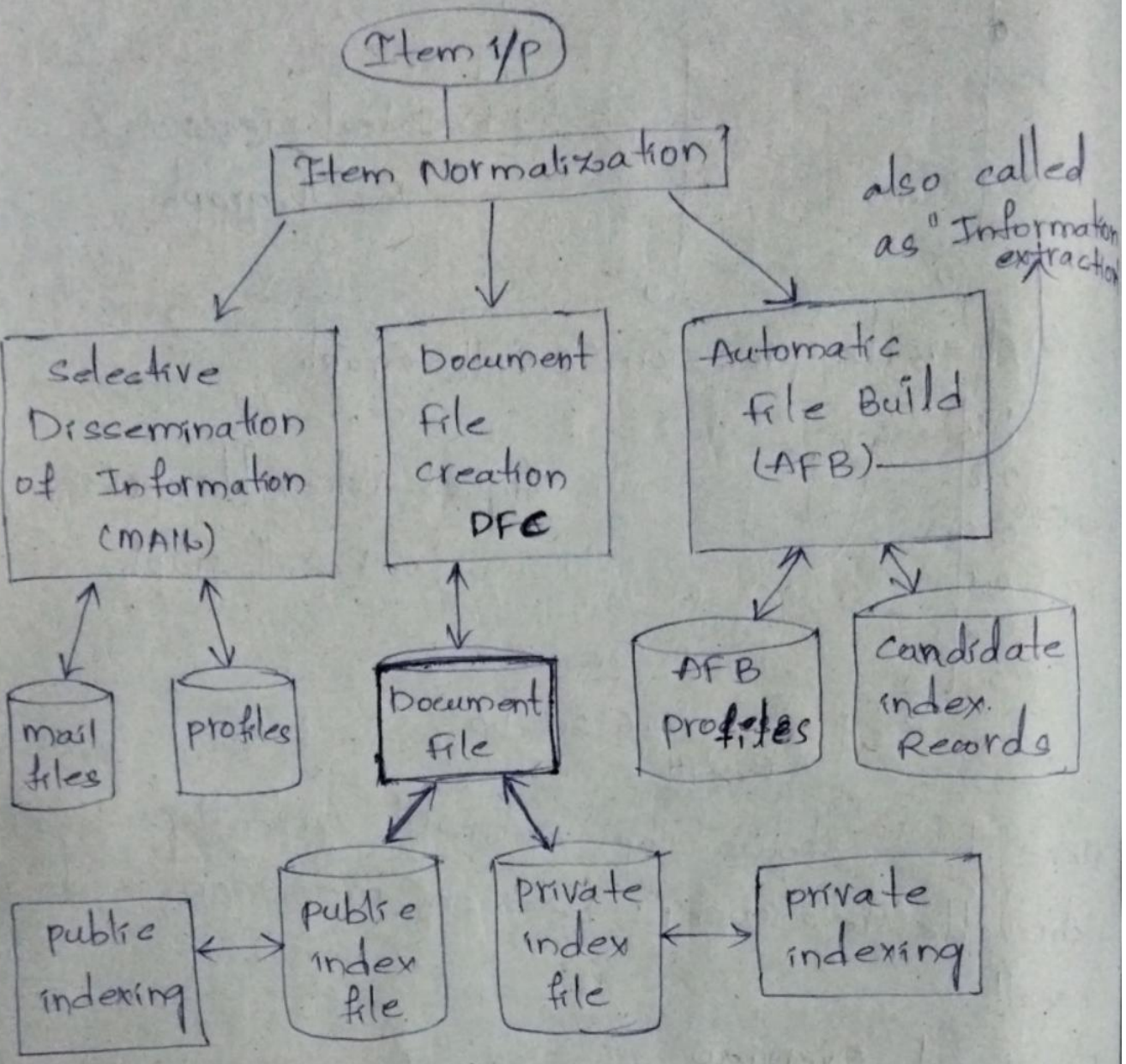


fig: Total IRS

Indexing: It is originally called as cataloging.

→ Indexing refers to organisation of data according to specific scheme (or) plan.

Functional Overview:

A total information storage and Retrieval system is composed of 4 major functional process.

i) Item Normalization

ii) Selective Dessimination of information (i.e. mail)

iii) Archival Document Database search and index.

(iv) Index Database is searching along with the "Automatic file Build Process" that supports index files.

1. Item Normalization:

→ The first step in any integrated system is to normalize the incoming items to a standard format.

Item

→ The term "item" is used to represent the "smallest complete unit" that is processed and manipulated by the system.

→ The Definition of item varies how a specific source treats information.

e.g.: A complete document, such as a book news paper (or) magazine could be an item at other items each "chapter" (or) "article" may be defined as an item

⇒ IRS 4 functional Process:

First you have to give your "item as input" here, item is nothing but any thing either it may be document / image / video etc.

These items are given as i/p then those items are Normalize.

1.e. Item Normalization.

Item Normalization.

→ The first step in any integrated system is to Normalize the incoming items into a standard format.

Normalization: It is a process of bringing (or) retrieving something to a normal condition (or) state.

→ Item normalization provides logical restructuring of the item.

→ The following operations are performed during item Normalization:

* Identification of processing tokens (e.g. words)
* characterization of tokens and stemming (e.g. removing word endings) of the tokens.

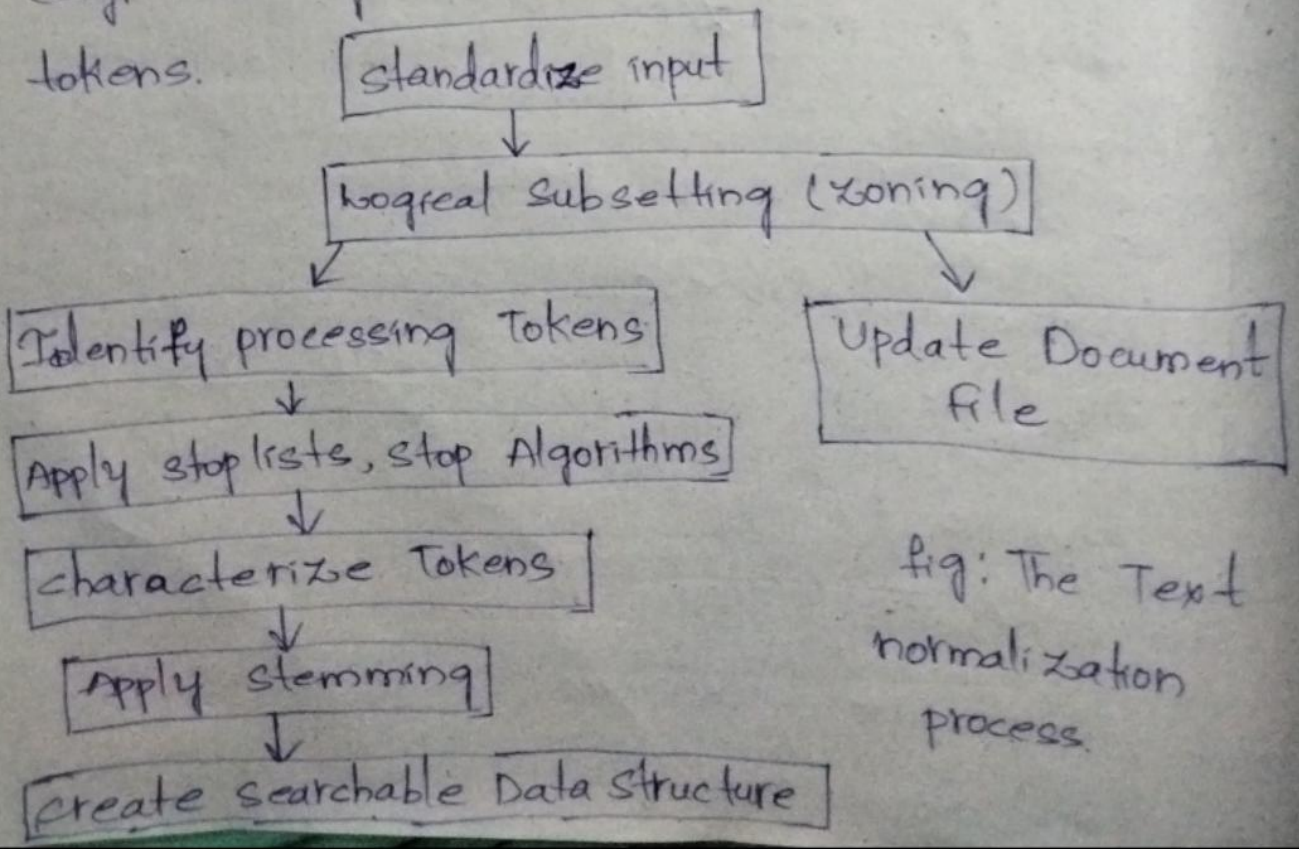


fig: The Text normalization process.

Standardize input:

→ standardizing the input takes the different external formats of input data and performs the translation to the formats acceptable to the system.

→ A system may have a single format for all (or) allow multiple formats.

For example:

→ standardization could be translation of "foreign languages" into "unicode".

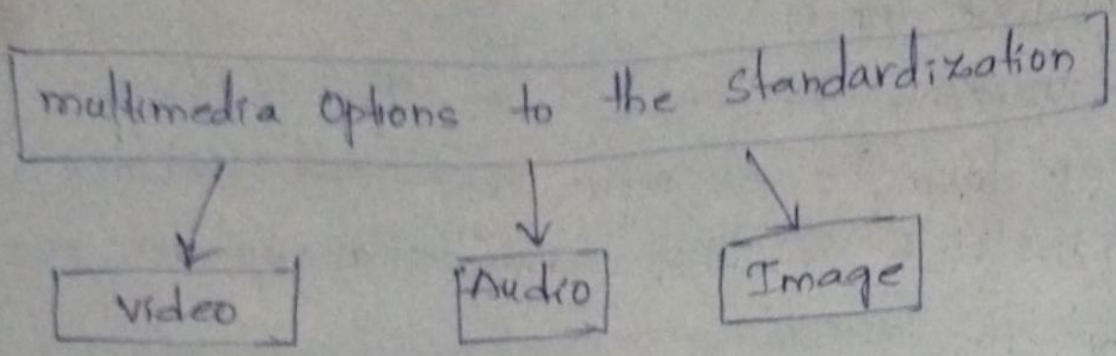
→ The Unicode is an evolving international standard based upon 16 bits (2 bytes) that will be able to represent all languages.

→ Every language has a different internal binary encoding for the characters in the language.

The "standard encoding" that covers English, French, Spanish etc is "Iso latin"

→ multimedia is an extra dimension to the normalisation process.

→ There are a lot of options to the standards that are being applied to the Normalization.



Video:

→ If the input is "video", the "digital standards" being applied to MPEG-2, MPEG-1, AVI (or) Real media.

→ MPEG - Motion Preture Expert Group.

→ MPEG is the most Universal standards for higher Quality video where Real media is the most common standards for lower quality video being used on internet.

Audio:

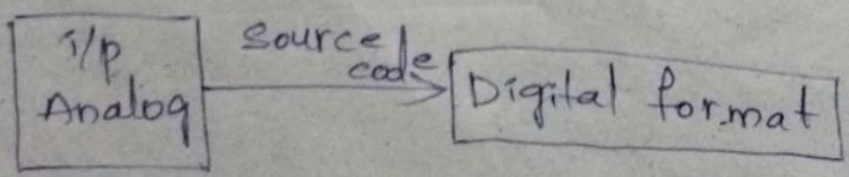
Audio standards are typically WAV (or) Real media (Real Audio).

Images:

Images vary from JPEG to Bmp.

finally,

In multimedia, standardization i/p analog source code is encoded into a digital format.

i.e. 

Logical subsetting (Zoning):

→ This process is used to "parse" the item into "logical subdivisions" that have meaning to the user, this process is called "Zoning."

→ A typical item is sub-divided into Zones, which may overlap and can be hierarchichal such as
- Title
- Author
- Abstract
- main text
- conclusion
- References.

→ Sometimes Zoning allowing search to be restricted to a "specific zone."

for e.g.:
If the user is interested in articles discussing "Einstein" then the search should not include Bibliography, which could include references to articles written by "Einstein."

⇒ Once a search is complete, the user wants to efficiently review the results to locate the "needed information."

⇒ The Zoning is divided into two process
* identify processing tokens.
* Update document file.

*Identify processing token:

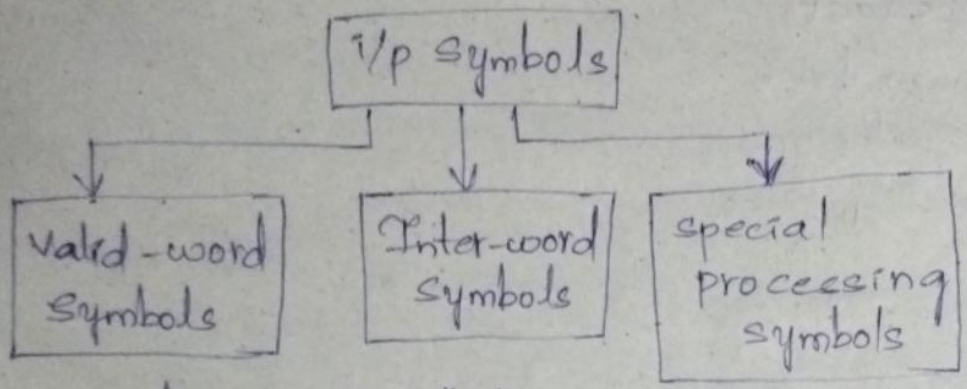
The first step in identification of a processing token consists of determining ~~the word~~ "a word".

⇒ Systems determining words by dividing input symbols into three classes.

i.e. i) Valid word symbols.

ii) inter-word symbols

iii) Special processing symbols.



eg: words e.g.: " ", — e.g.: '-' (hyphon)

semicolons, blanks

Word:

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

Inter-word Symbols:

In many systems, the inter-word symbols are non-searchable and should be carefully selected.

e.g.: semicolons, blanks.

Special Processing Symbols:

There are ~~many~~ some symbols that may required in special processing.

e.g.: - (hyphen)

A hyphen may use in many ways. i.e. at the end of a line, it is used to indicate continuation of a word.
In other places, it links the independent words.

APPLY STOPLISTS (stop Algorithms):

Next,
A stop list/algorithm is applied to the list of potential processing tokens.
The objective of the stop function is to save system resources by elimination.
→ The Best Technique to elminate the majority of these words is via a stop algorithm ~~various~~ versus trying to list them individually.
Examples: (E.g. of stop algorithms are:)
* stop all numbers greater than '999999'
* stop any processing token that has numbers and characters intermixed.

characterize Tokens:
The Next step in finalizing on processing tokens is identification of specific word characteristics.
→ The characteristic is used in system is assist disambiguation of a particular word.
eg: for a word such as "plane", the system understands that it could mean "level (or) flat"

as an "adjective", "aircraft or fact" as a noun (or) "The act of smoothing" as a verb

Apply stemming.

Once the token has been identified and characterized, most of the systems are applying "stemming algorithms".

The stemming algorithms are used to normalize the "token to a standard semantic representation".

e.g.: The system must keep singular, plural, past tense etc (i.e. keep the stem of word) eliminating the ends.

Create Searchable Data structure

Once the Processing tokens have been finalized. Based upon the stemming algorithm and they are used as updates to the searchable data structure.

→ The searchable data structure is the internal representation (i.e. not visible to the user) of items that the user searches.

→ The structure contains semantic concepts that represent the items in the database and it limits what a user can find as result of their search.

2) Selective Dissemination of Information:

SDI is a concept that was introduced in information science by Hans peter Luhn in 1958.

The selective Dissemination of Information (mail) process provides the capability to dynamically compare newly received items in the information system against standing statements of users.

→ The mail process is composed of the

* search process
* User statements of interest (profiles) &
* User mail files.

Search Process

As each item is received, it is processed against every user's profile.

profiles

A profile contains a typically board search statements along with list of user mail files.

These mail files will receive the document if the search statement in the profile is satisfied

⇒ These profiles define all the areas in which a user is interested.

User mail files:

when a search statement is satisfied, the item is placed in the 'mail files' associated with the profile.

→ Items in mail files are typically viewed in time of receipt order and they are automatically deleted after a specified time period. (e.g. after a month)

→ selective Dissemination of information has not yet been applied to multimedia sources.

3) Document dataBase Search:

→ The document database search process provides the capability for a query to search against all items received by the system.

→ The Document database search process is composed of the

* Search process
* user Entered Queries (typically adhoc queries)
* the document database.

→ The document database which contains all items that have been * received
* processed &
* stored by the system.

→ The selective Dissemination of Information system and document database search both are used in "restrospective Search".

restrospective search means searching ~~the~~ in past & time invalid.

It can be search in two ways.

prospective search and retrospective search.

prospective search

~~It~~ can searching in future and it is time valid.

Restrospective Search

A search for information that has already been processed into the system can be considered as a "retrospective" search for information.

The document database can be very large, it consist of hundreds of millions of items (or) more.

Typically items in the document database do not change once received (i.e are noted edited).

Index database search

The index database search process provides the capability to "create indexes" and search them.

→ Index files are "structured database" whose records can optionally reference the items in the document database.

→ The user may search the index and retrieve

→ The system the index (or) the document it references.

→ The system also provides the capability to search the index and then search the items referenced by "index record" that is satisfied by index portion of query.

→ There are 2 classes of index files

1. public index files
2. private index files.

1. public index files:

→ public index files are maintained by professional library services.

→ They are typically index every item in the document database.

→ There is a small number of public index files.

2. private index files → These files have access lists (i.e. list of users and their privileges) that allow anyone to search (or) retrieve data.

2. private index files:

→ Every user can have one (or) more private index files that leading to a very large number of files.

→ private index files references only small subset of the total number of items in the

document database.
→ Private index files typically have very limited access.

4) Automatic File Build (AFB)

The automatic file Build is also called as Information Extraction.
⇒ To assist the users in generating indexes (i.e. especially the professional indexes) the system provides a processes carried automatic file build.
→ The AFBP has capability to processes selected incoming documents and the automatically determine potential indexing for the item.
→ The rules that govern, by which documents are processed for extraction of index information and index term extraction process are stored in "AFB profile".
→ when an item is ~~produced~~ processed it results in creation of "candidate index records"

Relationship to Database management Systems

There are two major categories of systems available to process items. They are

* Information Retrieval System (IRS)
* DataBase management System (DBMS)

→ A "confusion" can be arise when the slw systems supporting each of these applications and they get confused with the data they are "manipulating."

→ The Information Retrieval System is software that has the "features and functions" required to manipulate information items.

Vs

→ A database that is optimized to handle "structured data".

→ The information is "fuzzy text" (i.e. not clear (or) ~~dee~~ blurred), Here, the term "fuzzy" is used to represent the "result" from the minimal standards (or) controls on the creators text items.

i.e. the author is trying to "present the concepts, ideas and abstractions" along with supporting facts.

Vs

→ The structured data is well defined data typically represented by tables.

There is a semantic description associated with each "attribute" with in a table that well defines the "attribute".

for example:
→ There is no confusion between the meaning of "Employee name" (or) "Employee Salary" to enter the values in specific database.

on other hand,
if two different people generate an abstract for the same item, but they can be different.
i.e. One abstract may generally discuss the most important topic in an item.
Another abstract, using a different vocabulary that may specifies the details of many topics.

→ with structured data, user enters a specific request and the result they get is desired information.
i.e. the results an frequently tabulated and presented in a "report format."

3) A search of "Information" items has a "high probability" of net finding all the items i.e. a user is looking for.

→ The IRS gives the capabilities that assist the user in finding the relevant items such as "relevance feedback."

→ The Results from an Information System

search are presented in "relevance ranked order".

→ The confusion comes when DBMS software is used to store "information"

This is easy to implement, but the system lacks the ranking and relevance feedback features are critical.

Distinguish between IRS and DBMS.

IRS	DBMS
1) It does not offer an advance "DMF" usually data modeling in IRS is restricted to classification of objects.	1) It offer an Advanced Data Modeling facility (DMF) It includes the DDL & DML for modeling and manipulating data.
2) In IRS, Defining the data integration constraints validation mechanisms are less developed.	2) A major strength of the DDL of DBMS is the capability to define the data integrity constraints
3) In most of time, IRS provides 'imprecise' semantics	3) DBMS provides 'precise' semantics
4) IRS is characterized by unstructured data format	4) DBMS has structured data format.

IRS	DBMS
5) Query specification is incomplete in IRS.	5) Query specification is complete in DBMS
6) Query language is near to natural language in IRS	6) Query language is artificial in DBMS.

From practical standpoint, the integration of DBMS & IRS is very important.

Digital libraries and Data Warehouse

Digital libraries

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

1) Digital libraries
2) Data Warehouse (or) Data marts

1) Digital libraries:

They also called as Online library, internet library and Digital Repositories.

→ Dl is a metaphore for access to collections of electronic documents through a network.

→ A Digital library is a library in which collections are stored in digital formats (i.e. as opposed to print microform, or other media) and accessible by computers.

→ An electronic (or) digital library is a type of information retrieval system (IRS)

→ It provides the Architecture to

* model
* map
* Integrate and
* transform scattered information in drgital documents.

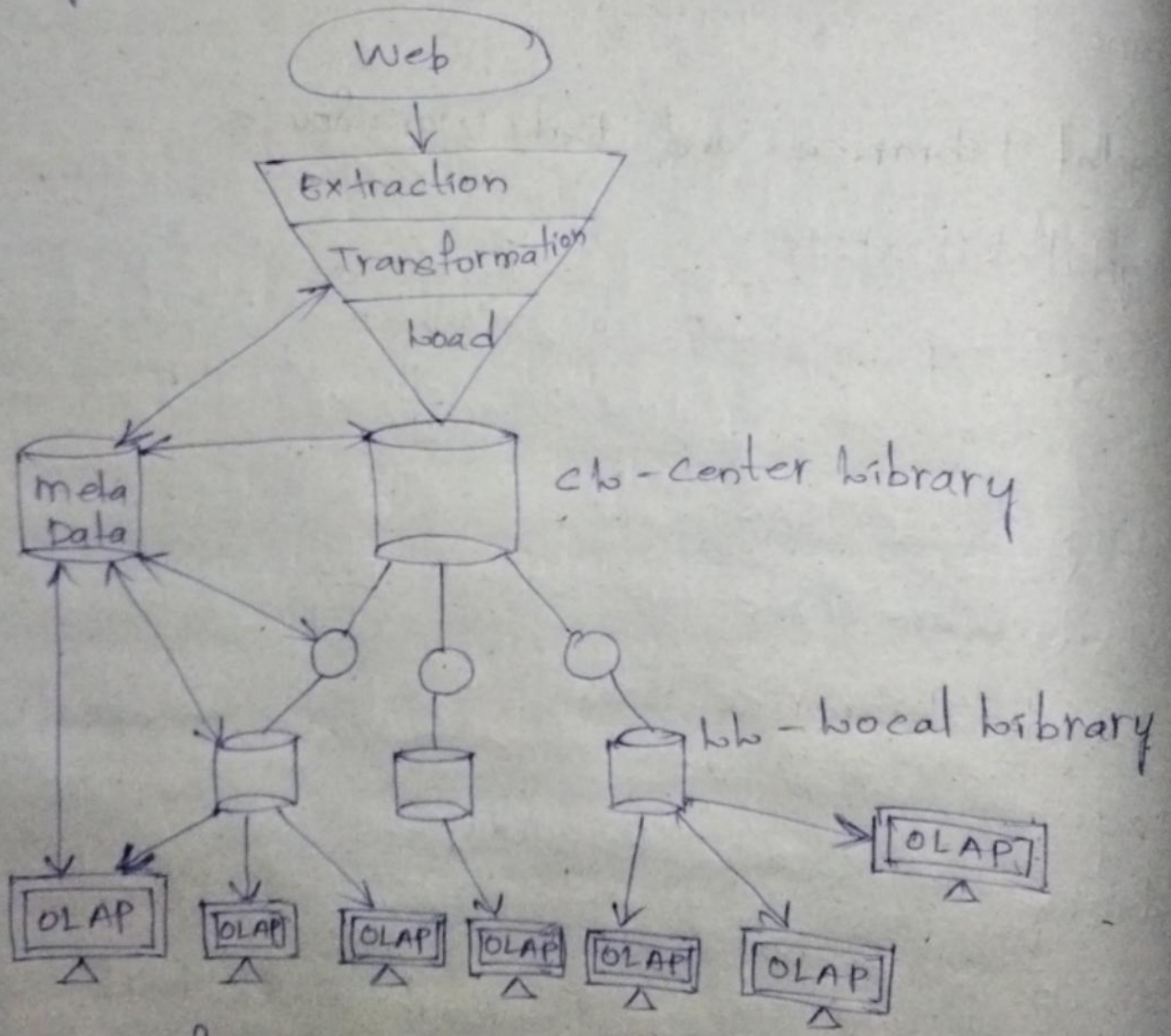
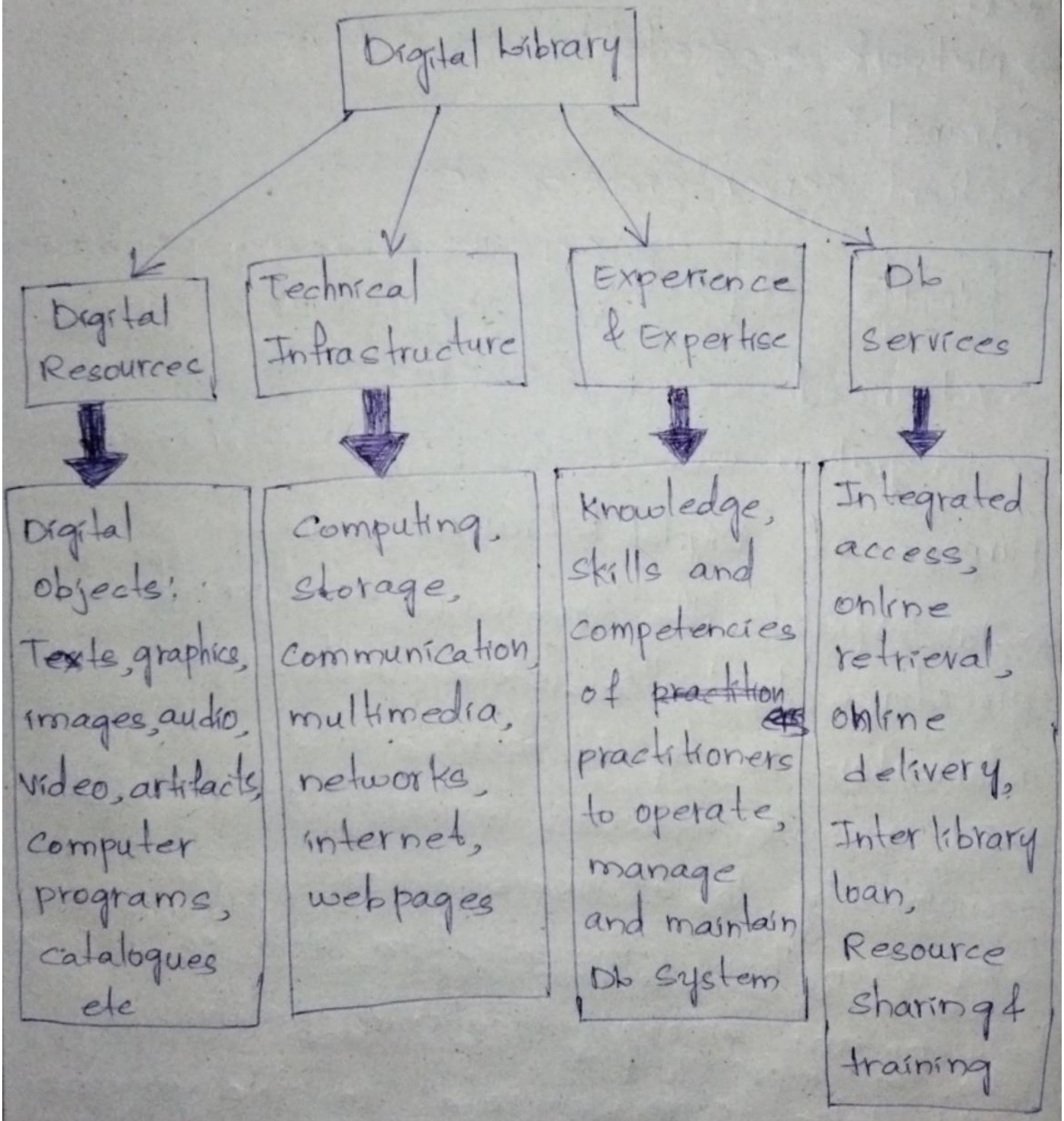


fig. Drgital library

→ The term Db's was first popularized by the NASA Drgital libraries initiative in 1994.

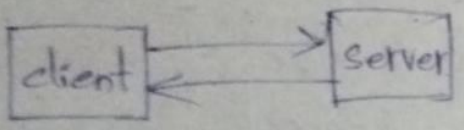
→ Digital library is not a single Entity. It requires technology that link the resources of many collections.

→ The links between the digital libraries and their resources are ~~tan~~ transparent to users shown in following figure.



Functions of Digital Library:

→ Access to large amounts of information to users whenever they need the informat
→ Access to primary information sources.
→ Support the multimedia content along with text.
→ Network accessibility on "internet" and "intranet."
→ client server architecture



sending & receiving information between client & server.
→ Advanced search and retrieval.
→ Integration with other digital libraries.

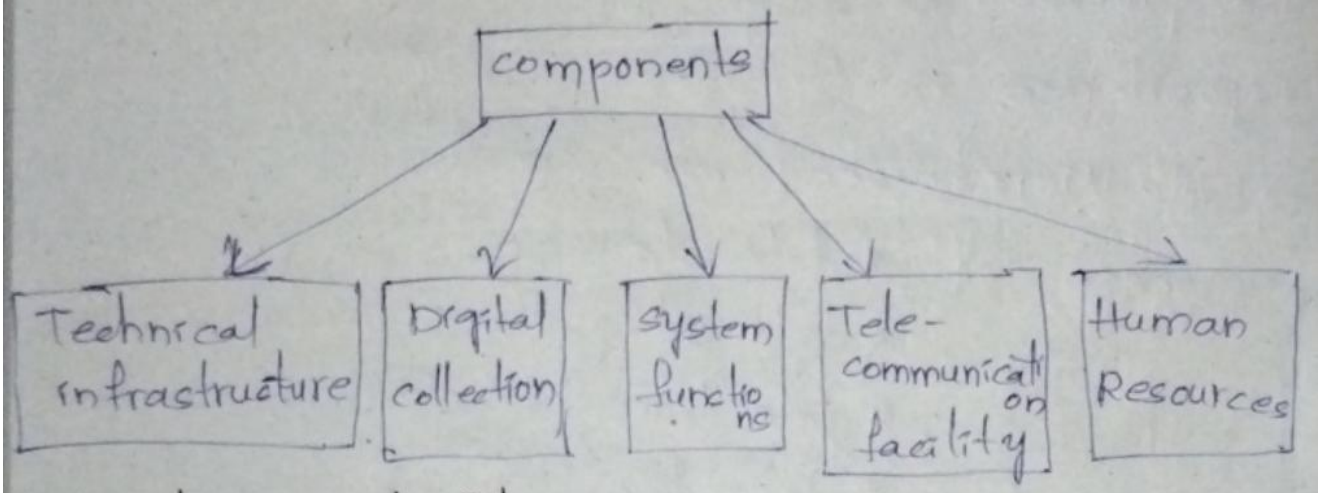
purpose of Digital library:

→ To expedite, the systematic development of procedures to collect, store and organize information in digital form.
→ It promote efficient delivery of information economically to all users.
→ It strengthen communication and collaboration between and among educational institutions.
→ It take "leadership role" in the generation and dissemination of knowledge.

Components of Digital Library.

The components of Digital library are:

* Infrastructure
* Digital Collection
* System functions
* Telecommunication facility
* Human Resources.



Advantages of DLs:

-> No physical Boundary
-> multiple Accesses.
-> Universal accessibility
-> Round the clock Availability
-> Added values
-> Enhanced IR

Limitations

* Lack of screening (or) validation
* Lack of preservation of "Best in class"
* Lack of preservation of "fixed copy".
  i.e. for the record and duplicating scientific research.

2) Data Ware house (or) Data marts

A ~~database~~ Data Ware house is a repository of information collected from multiple sources, stored under a "unified schema" and that usually resides at a single sites.

(or)

The process of collecting information from repositories.

→ Data Ware houses are constructed via a process of * data cleaning
* data integration
* data transformation
* data loading and
* periodic data refreshing.

→ The term data ware house comes from the commercial sector then Academic Sources.

→ Its goal is ~~to~~ to provide the critical information to decision makers so that they can answer the future direction queries.

→ ~~frequently, a data warehouse consists of the data, an implementation information~~

→ frequently, a datawarehouse is focused slowly on structured database.

→ A data ware house consist of the data, an information directory that describes the contents and meaning of data being stored.
→ An i/p function that captures data and moves it into the datawaare house, data search and manipulation tools that allow user to access and analyze the warehouse data and delivery mechanism to export data to other warehouses, datamarts (small ware houses (or) subset of large ware houses) and external systems.
→ Data Ware houses are similar to information storage and retrieval systems in that they both have a need for search and retrieval of information.
→ But the data warehouse is more focused on "structured data" and decision support technologies.
for e.g.:
The typical frame work for construction and use of a data ware house for all Electronics.

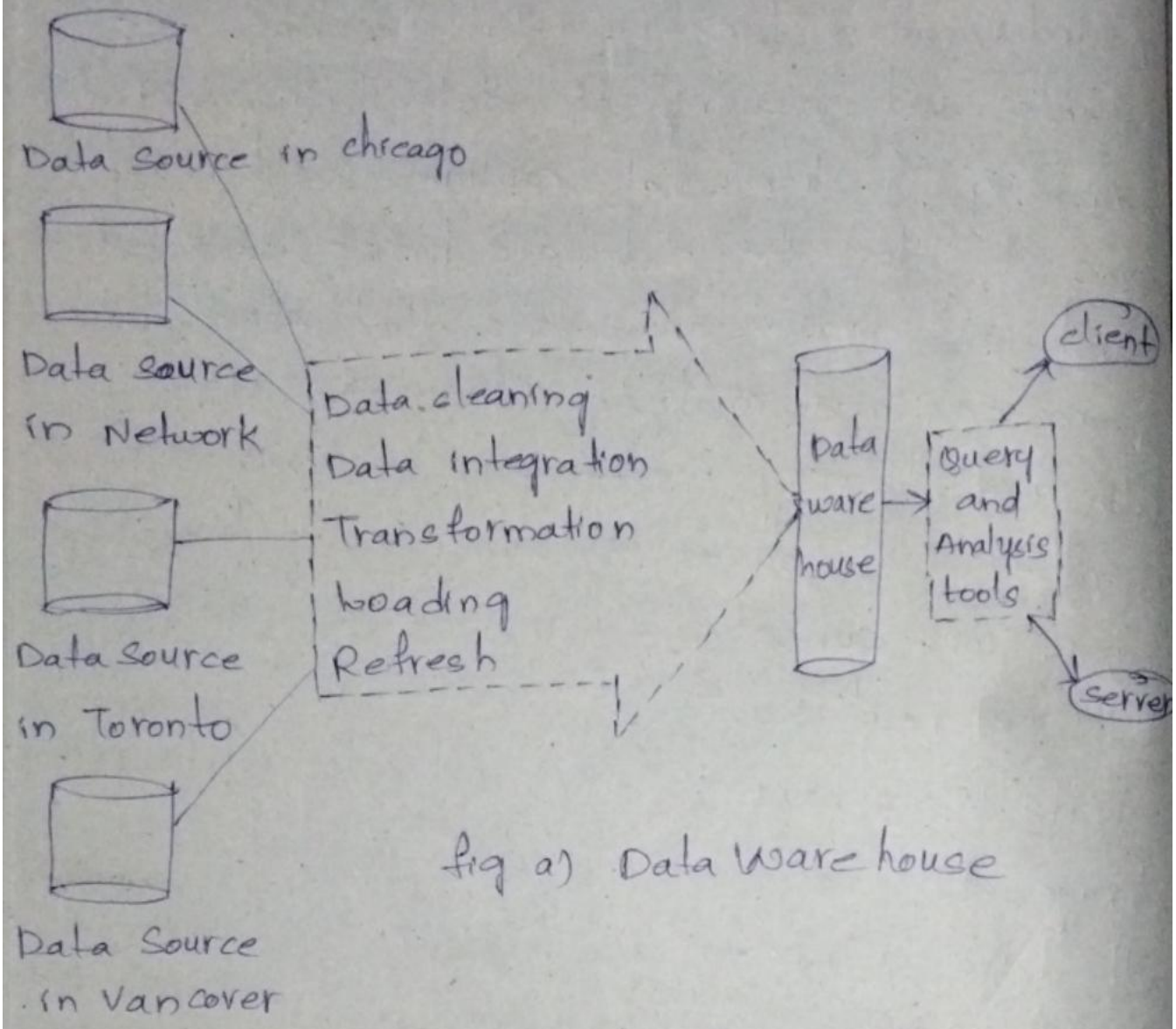


fig a) Data ware house

In addition to the Normal search process, a complete system provides a flexible set of analytical tools "mine" the data.

Data mining:

DM is a search process that automatically analyses data and extract the relationships.

→ DM is originally called "KDD"

KDD → Knowledge Discovery in Database

KDD is the process of discovering valueable information from a colleation of data (or) It is the process of converting raw data into useful information.

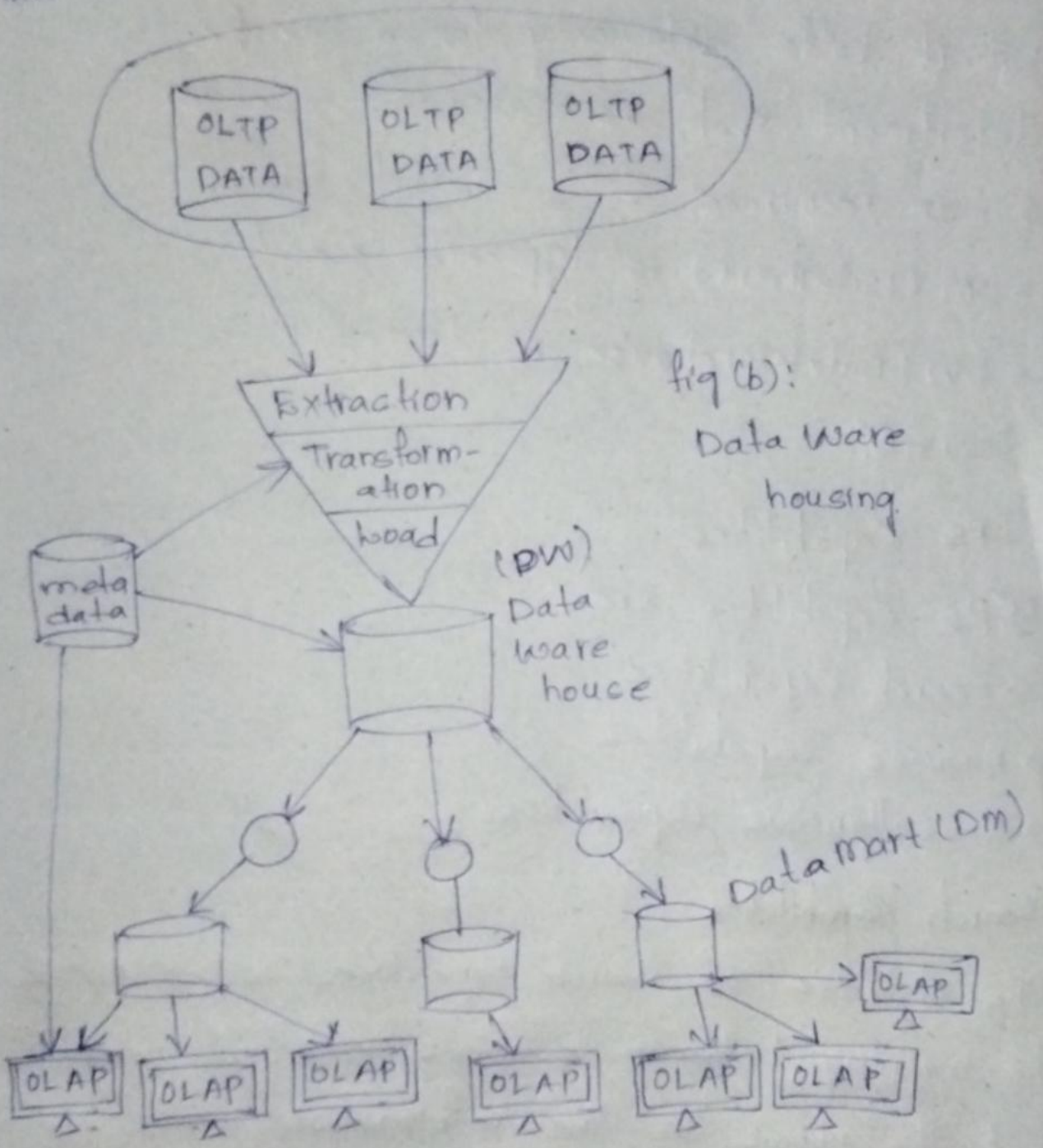


fig (b): Data Ware housing

from fig (b):

OLAP → online analytical processing

* It is based on multidemensional data model
* It is an approach to answer multidimensional data model

OLTP -> Online Transaction Processing

* It is based on captures and maintains transaction data in data base.

* Data mart is similar to Data Warehouse, but it holds data only for a specific department such as sales, finance (or) human resources.

-> DM (Datamart) is typically less than "100 GB."

-> DW (Data Warehouse is typically larger than "100 GB."

IRS Capabilities

IRS Capabilities are

-> Search Capabilities

-> Browse Capabilities

-> miscellaneous Capabilities.

Search Capabilities

The objective of Search Capabilities is to provide the mapping between a user's specified need and the items in the information database that will answer that need.

-> It addresses the both boolean logics and Natural language Querys.

-> It ~~going~~ It going to develop a concept

called weighting concept. Weighting is a process of holding locations and Ranking to a Relevant process.

IRS provides different search capabilities. They are

Search Capabilities

- Boolean logic
- Contiguous word Phrases
- Fuzzy searches
- Term Masking
- Numeric and Data Ranges
- Natural language Queries
- multimedia Query
- Concept / Thesaurus Expansion
- proximity

Boolean logic

Boolean logic allows a user to logically relate multiple concepts together to define what information is needed. The typical Boolean operators are AND, OR, and NOT. These operations are implemented using set intersection, set union and set difference procedures.

→ placing portions of the search statement in parantheses are used to specify the order of Boolean operations (i.e. nesting function).

→ If paranthesis are used, the system follows a default precedence ordering of operations.

(e.g.: typically NOT then AND then OR).

A special type of Boolean search is called "m of n" logic. for example, "find any item containing any two of the following terms: "AA", "BB", "CC". This can be expanded into a Boolean search resulte together ((AA AND BB) or (AA AND CC) or (BB AND CC)).

Proximity.

proximity is used to restrict the distance allowed within an item between two search terms. The semantic concepts is that the closer two terms are

Search Statement	System Operation
COMPUTER OR PROCESSOR → NOT MAINFRAME	select all items discussing Computers and/or Processors that do not discuss Mainframes
→ COMPUTER OR (PROCESSOR NOT MAINFRAME)	select all items discussing Computers and/or items that discuss Processors and do not discuss Mainframes.
→ COMPUTER AND NOT PROCESSOR OR MAINFRAME	select all items that discuss computers and not processors or mainframes in the item.

fig: Use of Boolean Operators

Proximity

The typical Proximity is used to restrict the distance allowed within an item between two search terms.

The typical format for proximity is:

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

The distance operator "m" is an integer number and units are in characters, words, sentences, or Paragraphs.

→ Sometimes the proximity relationship contains a direction operator indicating the direction (before or after)

→ A special case of the Proximity operator is the Adjacent (ADJ) operator that normally has a distance operator of one and a forward only direction (i.e. in WAIS).

→ Another special case is where the distance is set to zero, items are represented in the same semantic unit.

SEARCH STATEMENT

"Venetian" ADJ "Blind"

SYSTEM OPERATION

would find items that mention a Venetian Blind on a window but not items discussing a Blind Venetian.

SEARCH STATEMENT	SYSTEM OPERATION
"United" within five words of "American"	would hit on "United states and American interests", "United Airlines and American Airlines" not on "United states of America and the American dream."
"Nuclear" within zero paragraphs of "clean-up"	would find items that have "nuclear" and "clean-up" in the same paragraph.

fig use of Proximity

Contiguous Word Phrases.

A Contiguous Word Phrase (CWP) is both ways in

→ a way of specifying a query term

→ a special search operator

A Contiguous Word Phrase is two or more words that are treated as a single semantic unit. An example of a CWP is "United States of America".

* Two or more words represent in single semantic unit called query term.

e.g: United states of America

* ADJ is special search operator. ADJ performs as same as proximity

e.g: "Veritan" ADJ "Blind"

"United" ADJ "states" ADJ "of" ADJ "America"

In proximity, we use single ~~adjacent~~ ADJ operator and in CWP, we use multiple ADJ operator.
→ Contiguous Word Phrase maintained by Wide Area Information System (WAIS).

Fuzzy Search
Fuzzy searches provide the capability to locate spellings of words that are similar to the entered search term.
This search is mainly used to compensate for errors in spelling of words.
A fuzzy search on word "computer" include the following words from the information database:
→ computer
→ compiter
→ conputer
→ computter
→ compute.

Fuzzy searching has its maximum utilization in systems that accept items that have been Optical Character Read (OCR).
→ The OCR process is a pattern recognition process that segments the scanned image into a meaningful subregions.
→ Fuzzy search spends more time on search.

Term masking

Term masking is the ability to expand a query term by masking a portion of the term.

There are two types of search term masking: fixed length and varrable length.

→ fixed length masking is a single position mask. It masks out any symbol in a particular position.

→ varrable length "don't cares" allows masking of any number of characters within a processing token. The masking may be in the front, at the end, at both front and end, or imbedded

"*computer" Suffix search

"computer*" prefix search

"*computer*" imbedded string search

Search statement	System Operation.
multi $ national	matches "multi-national", "multinational", "multinational", but does not match "multi national" since it is two processing tokens
computer	matches, "minicomputer" "microcomputer" or "computer"

Search statement	System operation
comput*	matches "computers", "computing", "computes"
comput	matches "microcomputers", "minicomputing", "compute"

fig: Term masking

Numeric and Date Ranges

Term masking is useful when applied to words but does not work for finding ranges of numbers or numeric ~~date~~ dates. To find numbers larger than "125", using a term "125*" will not find any number except those that begin with the digits "125."

→ Indicating the ranges between 2 Ranges is called Numeric.

→ Indicating the date between two dates is called Date Ranges.

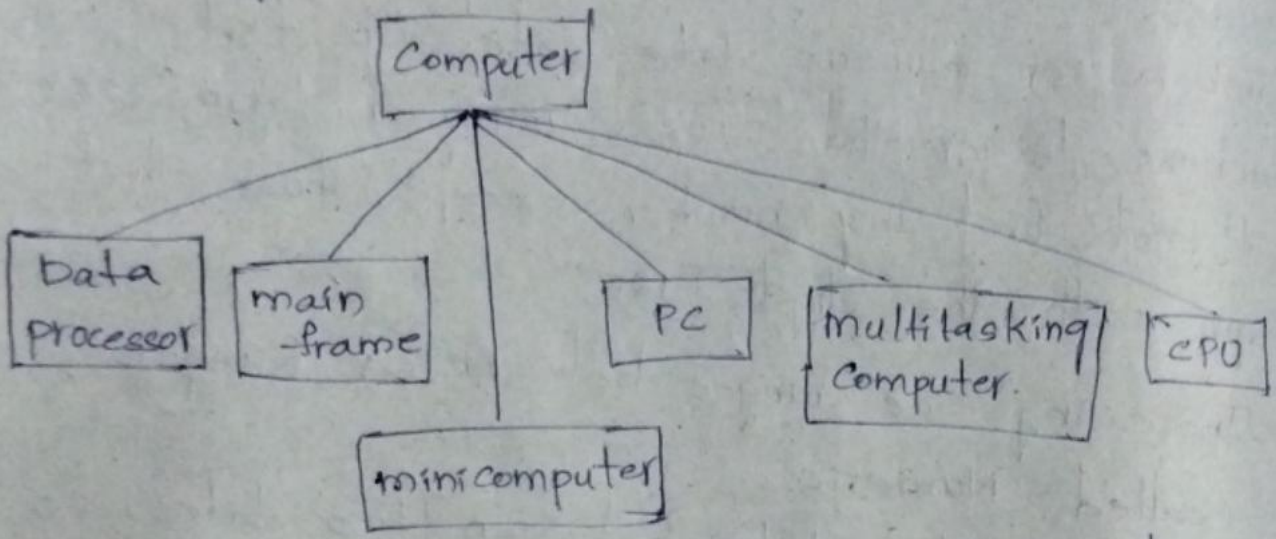
A user could enter inclusive (e.g.: "125-425" or "4/2/93-5/2/95" for numbers and dates) to infinite ranges ("> 125," "<= 233", representing "Greater Than" or "Less Than or Equal") as part of a query

Concept/Thesaurus Expansion

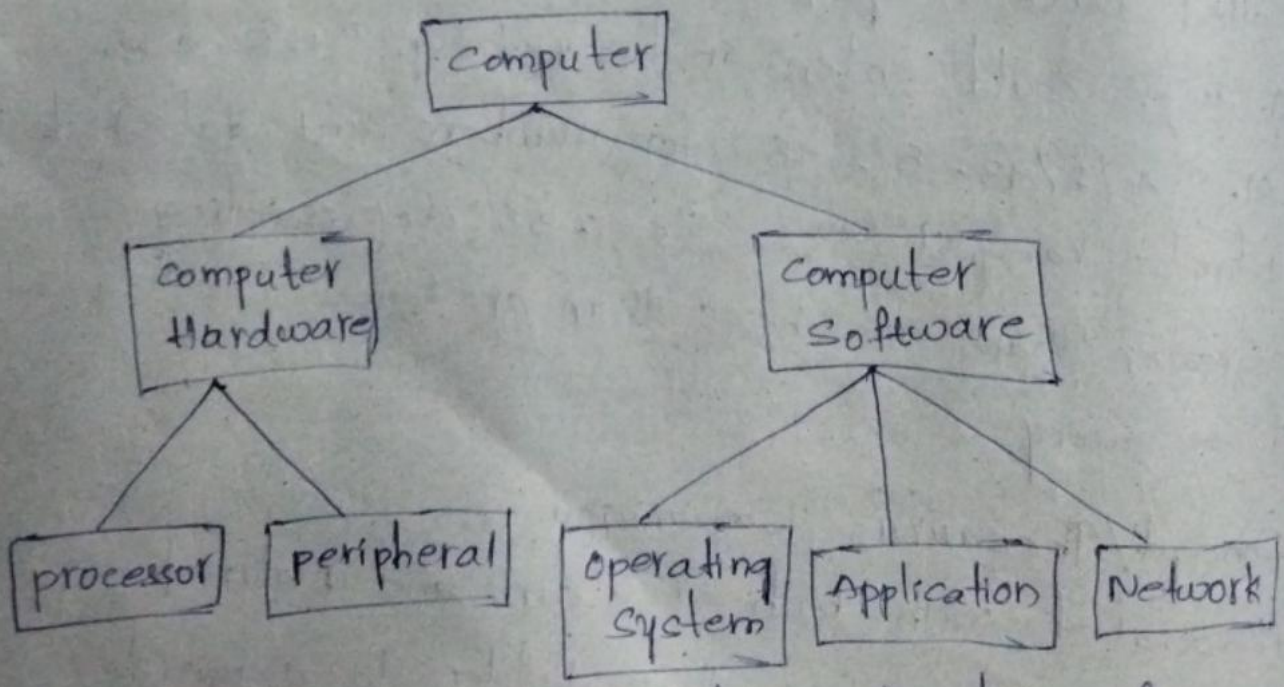
Associated with both Boolean and Natural language queries is the ability to expand

the search terms via Thesaurus or Concept class database reference tool. A Thesaurus is typically a one-level or two-level expansion of a term to other terms that are similar in meaning.

Thesauri are either semantic or based upon statistics. A semantic thesaurus is a listing of words and then other words that are semantically similar.



~~Thesaurus~~ Thesaurus for term "Computer"



Hierarchical Concept class structure for "Computer"

Natural Language Queries

Natural Language Queries allow a user to enter a prose statement that describes the information that the user wants to find.

An example of a Natural Language Query is:

Find for me all the items that discuss oil reserves and current attempts to find new oil reserves.

Include any items that discuss the international financial aspects of the oil production process.

Do not include items about the oil industry in the United States.

Multimedia Queries

The user interface becomes far more complex with the introduction of the availability of multimedia items.

The still image could be used to search images that are part of an item. They also could be used to locate a specific scene in a video product.

The ability to search for audio as a match makes less sense as a user specification.

Browse Capabilities:

Once the search is complete, Browse capabilities provide the user with the capability to determine which items are of interest and select those to be displayed.

Browse capabilities can assist the user in focussing on items ~~that~~ that have the highest likelihood in meeting his need.

→Ranking

It is used to determining different tokens of query

→ It is based on two things.

i) Predict Relevant values and

ii) status summary

→ It is based on ~~characterised~~ characterization item and database.

This allows the user to determine at what point to stop reviewing items because of reduced likelihood of relevance.

The relevance score is an estimate of the system and normalized to a value between 0.0 and 1.0

Zoning:

* process of dividing the standard ~~zones~~ input into several zones.
* when the user displays a particular item, the objective of minimization of overhead still applies.
* The user wants to display the complete item for detailed review.

e.g.: display of the title and Abstract and author and reference for sufficient information for user to predict the potential relevance of an item.

* In book, items are divided into uniform sized passages.

Highlighting

It is used to display and in an indication of why an item was selected.

* Different strengths of highlighting indicates how strongly the highlighted word participated in the selection of item.

Highlighting ~~has~~ is always been useful in Boolean systems to indicate the cause of retrieval.

The term being highlighted that ~~earn~~ caused

particular item to be returned may not have direct or obvious mapping to any of search terms entered.

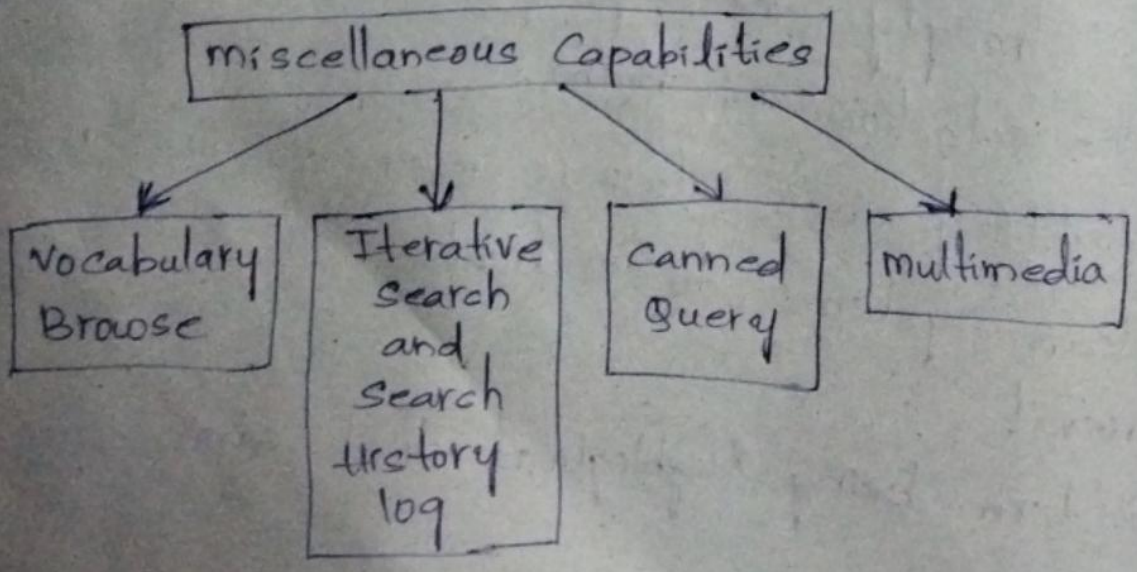
The highlighting may vary by introducing colors and intensities to indicate the relative importance decision to retrieve the item.

Misscellaneous Capabilities.

* Vocabulary Browse
* Iterative Search and Search History log
* Canned Query
* Multimedia

There are many additional functions that facilitate the User's ability to input queries, and reducing the time, it ~~forces~~ ~~to~~ takes to generate the queries.

The Different types of miscellaneous capabilities:



Vocabulary Browse:

→ Vocabulary Browse provides the capability to display in alphabetical sorted order words from the document database.

→ All Unique words (processing tokens) in the database are kept in sorted order along with a count to the number of unique items in which the word is formed.

→ The User can enter a word/word fragment and the system will begin to display the dictionary around the entered text.

For example: In vocabulary Browse, if the user enters a word "Compute", then system display result to that word "compute".

~~step~~ step1:

The system "indicates" what word ~~fragment~~ the user entered and then alphabetically displays other words found in the database.

step 2:

The user can continue scrolling in either direction for reviewing addition terms in the database.

step3:

Vocabulary browse provides information on the exact words in the database.

Example 1:

→ Here, the user enter a search term "comput" in effect of searching "compulsion" (or) "compulsive" (or) "Compulsory".

Here, comput* ← indicates fixed / varrable length masking.

Example 2

It also shows that some one probably entered the word "comput" when then really meant "compute".

i.e. has a result of vocabulary browsing, a term may be seen to exist in large number of documents.

	TERM	OCCURENCES
comp*	Compromise	53
	comptroller	18
compul*	compulsion	5
	compulsive	22
	compulsory	4

	TERM	OCCURENCE
comput	computation	265
	compute	1245
	computer	1
	computen	101800
	computerize	18
	computes	29

fig: Vocabulary Browse hirst with entered term "Comput".

Iterative Search and Search History log.

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

The search history log is the capability to display all the previous searches that were executed during the current session.

Canned Query

The capability to name a query and store it to be retrieved and executed during a later user session is called canned or stored queries. Canned query features also allow for variables to be inserted into the query and bound to specific values at execution time.

Multimedia

Once a list of potential items that satisfy the query are discovered, the techniques for displaying them when they are multimedia introduces new challenges.

→ It used to use graphical display to show a higher level view of information.

→ But this has the disadvantage of using more than one line per hit and reducing the number of hits that a user can select from a single screen.

⇒ If the source is audio, then other problems associated with the human linear processing of audio becomes major issues.