UNIT-I

(i) Introduction to Information Retrieval Systems:

(ii) Definition of information retrieval system.

(iii) objectives of inform^{n} RS.

(iv) functional overview.

(v) Relationship to DBMS

(vi) Digital libraries & Data warehouses

(vii) IRS capabilities:

(viii) Search capabilities.

(ix) Browse "

(x) Miscellaneous capabilities.

Introduction to IRS:

Definition of IRS:

"An Information Retrieval System is a system that is capable of storage, retrieval and maintenance of information."

Information in this context can be composed of following:

- text → including numeric and date data.
- images.
- audio
- video and other multi-media objects.

Of all the above data types, text is the only data type that supports full functional processing. The term term is used to represent the smallest complete textual unit that is processed and manipulated by the system.

The definition of item varies by how a specific source treats information. A complete document, such as a book, news paper or magazine could be an item. At other times each article may be defined as an item.

An information retrieval system consists of a s/w prog that facilitates a user in finding the information the user needs. The system may use standard computer h/w or specialized h/w to support the search subfunction.

The efficiency of an information system lies in its ability to minimize the overhead for user to find the needed information.

Overhead from a user's perspective is the time required to find the information needed, excluding the time for actually reading the relevant data. Thus search composition, search execution, and reading non-relevant items are all aspects of information retrieval overhead.

The first IRS originated with the need to organize information in central repositories like libraries. Not only librarians, professional searchers, etc. engage themselves in the activity of inform^{n} retrieval but nowadays hundred of millions of people engage in IR every day when they use web search engines.

The IR system assists the users in finding the information they require but it does not explicitly return the answers to the question.

IR also extends support to users in browsing or filtering document collection or processing set of retrieved documents. The system searches over billions of documents stored on

millions of computers. A spam filter, manual or automatic means provided by email prog for classifying the mails so that it can be placed directly into particular folders.

An IRS system has the ability to represent, store, organize, and access information items. A set of keywords are required to search. keywords are what people are searching for in search engines. These keywords summarize the description of the information.

III. Objectives of Information Retrieval systems (IRS):

The general objective of an IRS 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.

For example: Query generation, Query execution, Scanning results of query to select items to read, reading non-relevant items etc..

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

Measures associated with IRS:

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 4 segments.

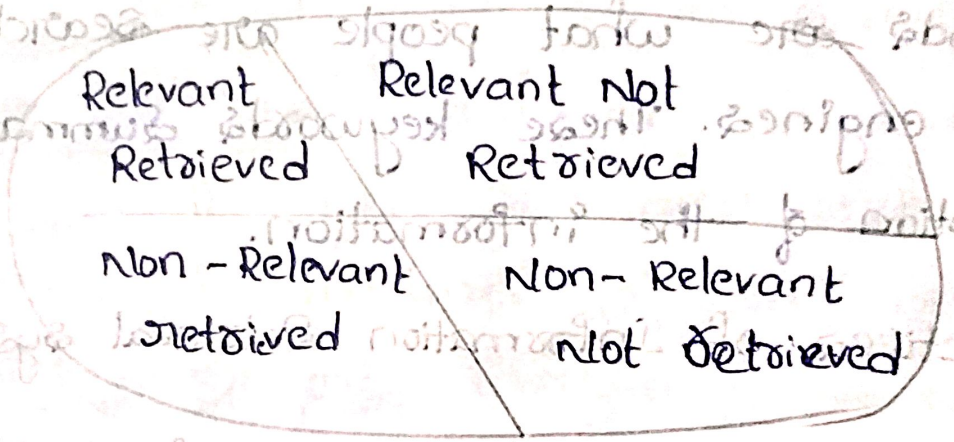


fig: Effects of search on total document space

» Relevant items are those documents that contain information that helps the searcher in answering his question.

» Non-relevant items are those items that do not provide any directly useful information.

There are two possibilities with respect to each item: It can be retrieved or not retrieved by the user's query.

The two measures precision & recall can be 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 inform^n retrieval overhead for a user associated with a particular search. If a search has a 85% precision, then 15% of the user effort is overhead reviewing non-relevant items.

Recall gauges how well a system processing a particular query is able to retrieve the relevant items that the user is intrested in seeing. Recall is a very useful factor of concept, but due to the denominator, is non-calculable in

operational systems.

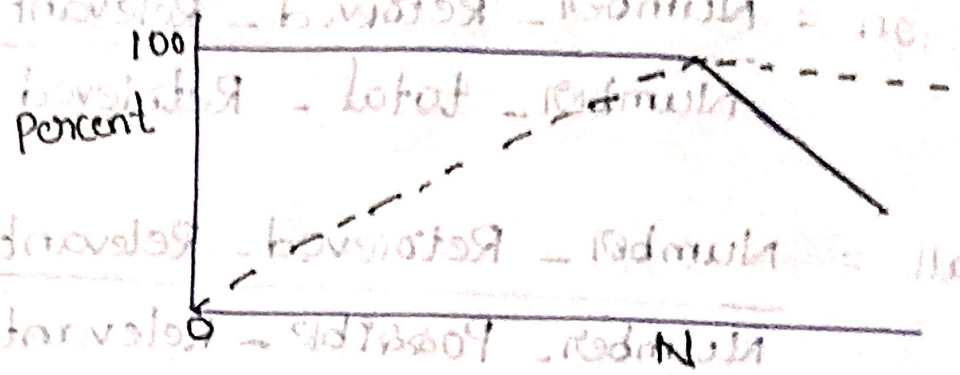


fig: Ideal precision and Recall

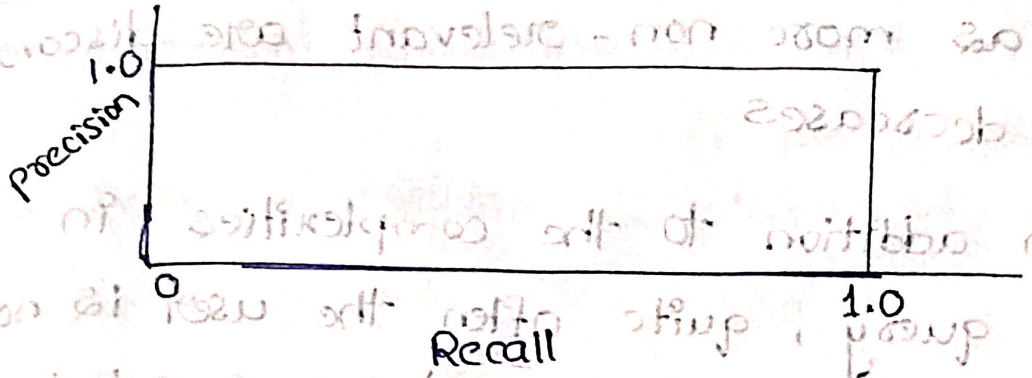
In the above fig shows the values of precision and Recall as the num of items retrieved increases, under an optimum query where every returned item is relevant. There are "N" relevant items in the database.

In the above fig; the basic properties of precision (solid line) and Recall (dashed line) can be observed.

precision starts off at 100% and maintains the value as long as relevant items are retrieved.

Recall starts off close to zero and increases as long as relevant items are retrieved until all possible relevant items have been retrieved.

Once all N relevant items have been retrieved, the only items being retrieved are non-relevant.



fig(i): Ideal Precision / Recall Graph

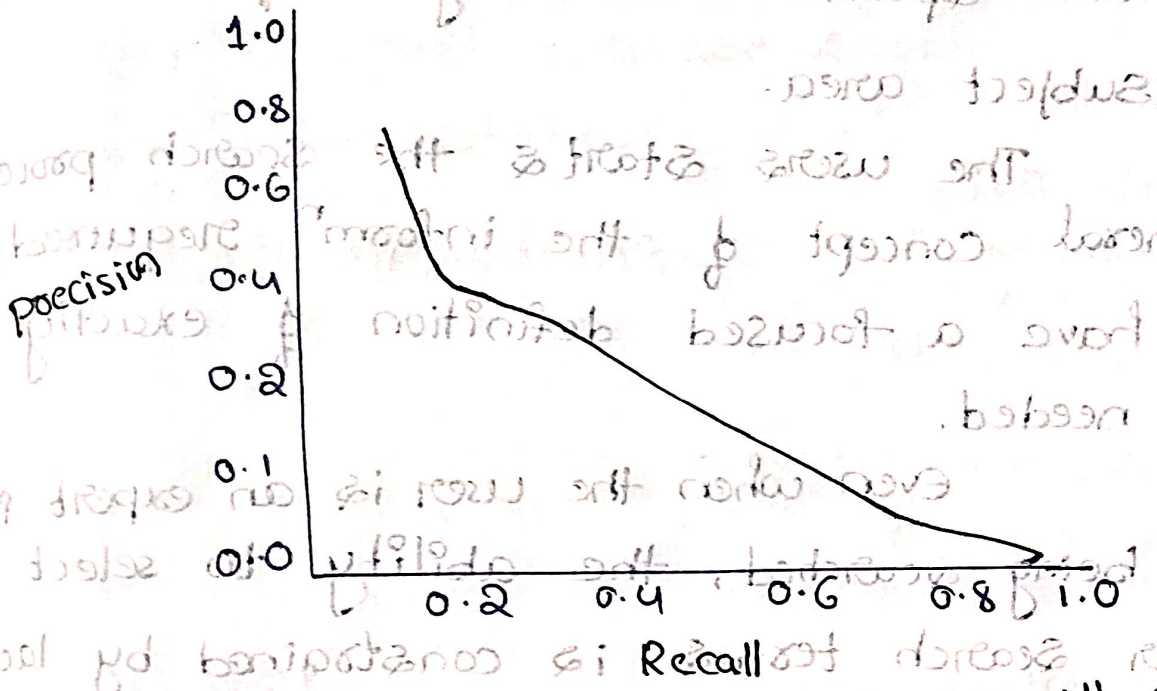


fig (ii): Achievable precision / Recall graph.

In the above fig (i) & (ii) shows the optimal and currently achievable relationships between precision and Recall.

fig (i) shows the precision stays at 100% (1.0). Recall continues to increase by moving to the right on the X-axis unit, until it also reaches the 100% (1.0).

fig (ii) shows Stops here, continuation stays at the same X-axis location (recall never changes) but decreases down the Y-axis until it gets close to

the X-axis as more non-relevant are discovered and precision decreases.

In addition to the complexities in generating a query, quite often the user is not an expert in the area that is being searched and lacks domain specific vocabulary unique to that particular subject area.

The users starts the search process with a general concept of the inform^{n} required, but not have a focused definition of exactly what is needed.

Even when the user is an expert in the area being searched, the ability to select the proper search terms is constrained by lack of knowledge of the author's vocabulary.

All writers have a vocabulary limited by their life experiences, environment where they were raised and ability to express themselves. The users search vocabulary does not match the authors vocabulary

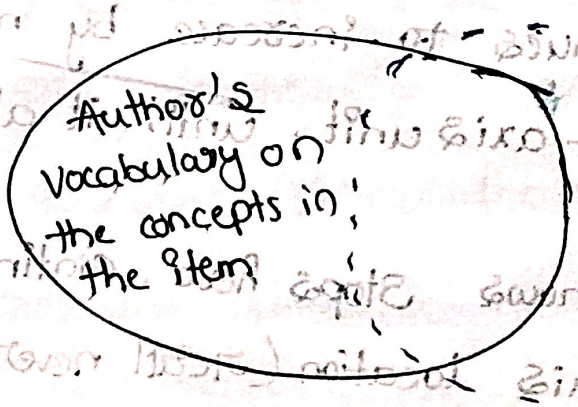


Fig: Vocabulary domains

Functional Overview:

A total information storage and Retrieval System is composed of 4 major functional process: They are.

i) Item Normalization
ii) Selective Dissemination of Information
iii) Document Database Search
iv) Index database search.

(i) Item Normalization :-

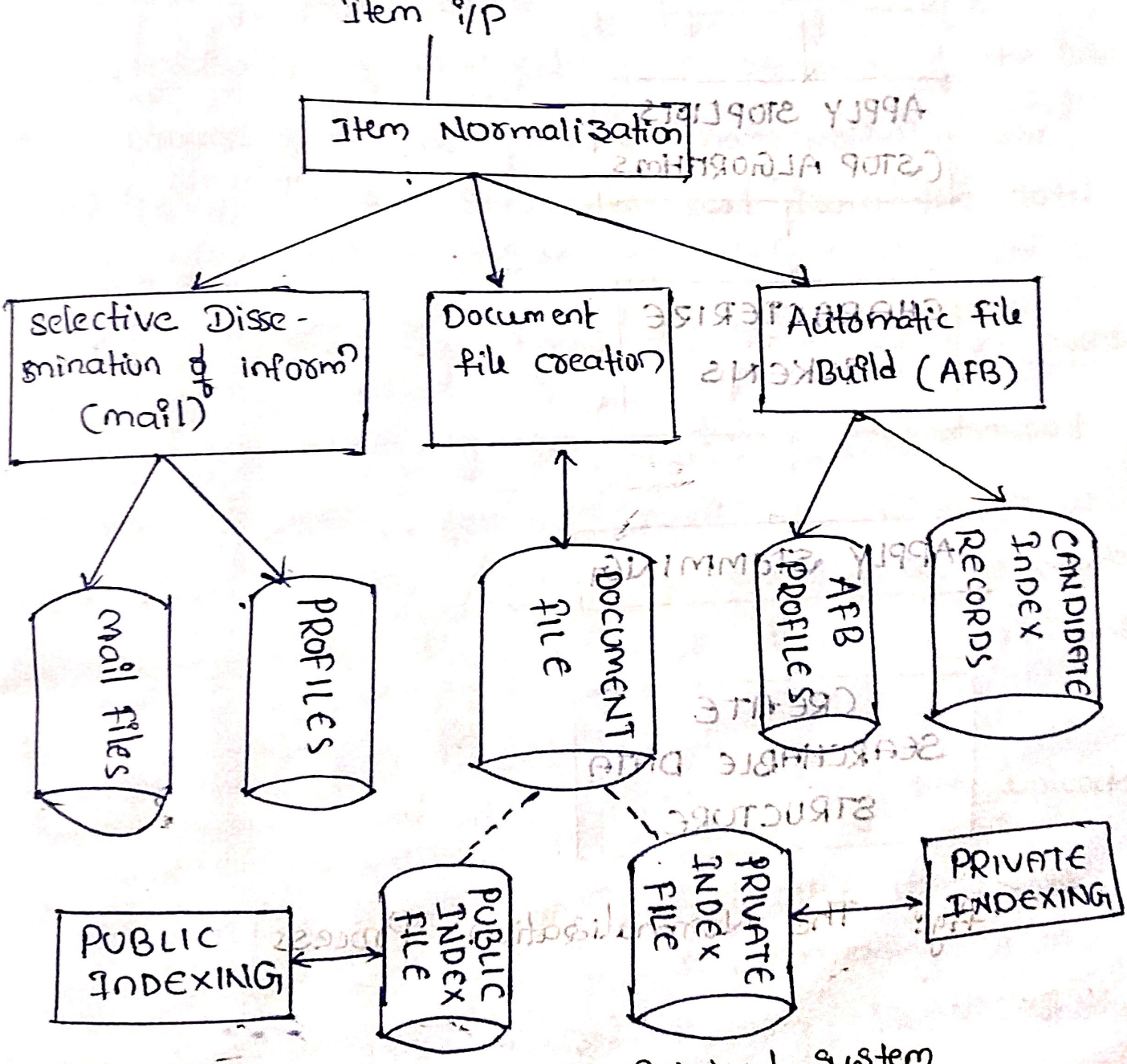


fig: Total Information Retrieval system

Item Normalization:

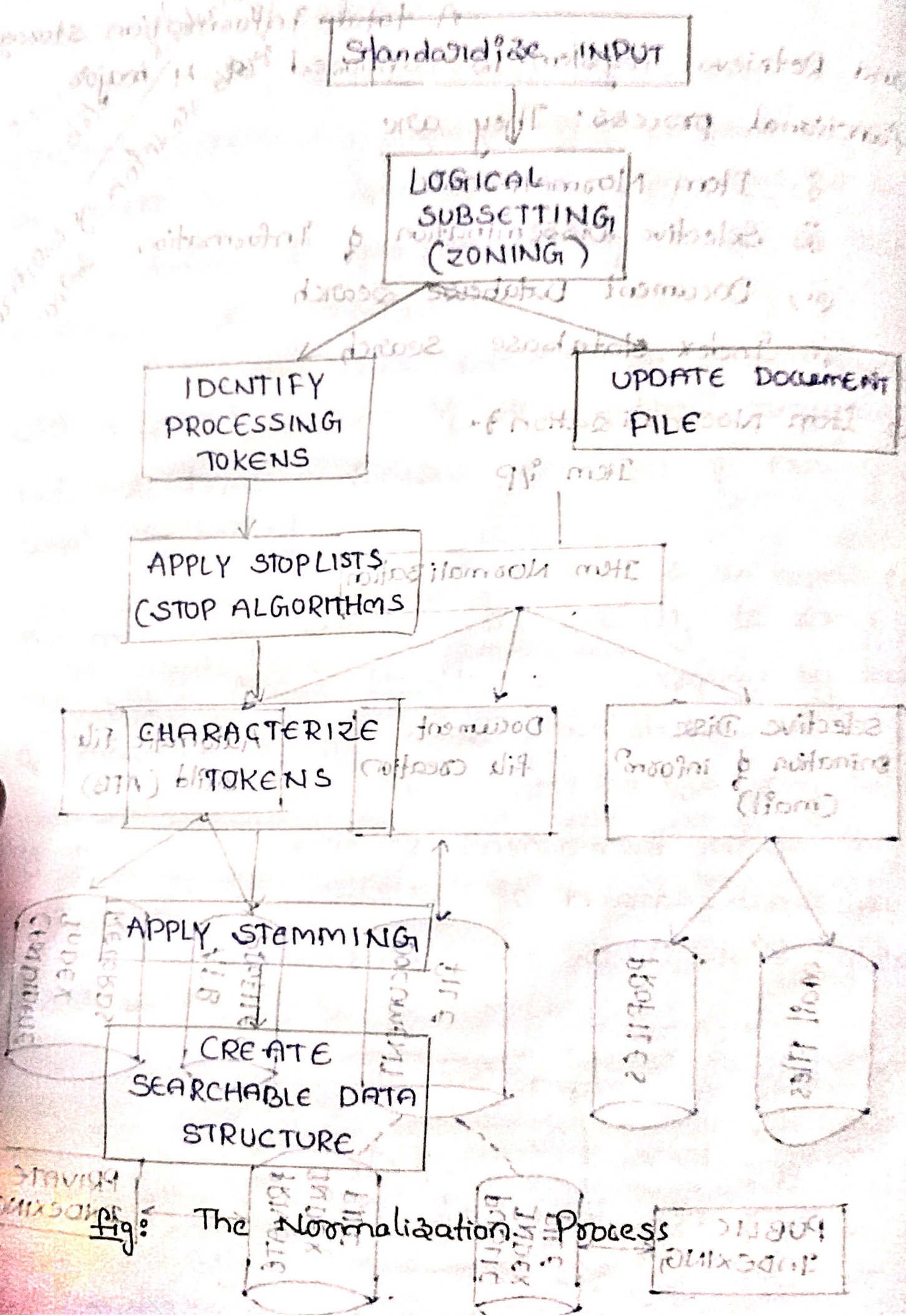


Fig: The Normalization Process

The first step in any integrated system is to normalize the incoming items to a standard format. In addition to translating multiple external formats that might be received into a single consistent data structure that can be manipulated by the "fun" process, item Normalization provides logical restructuring of the item.

Additional oper" during item normaliz- ation are needed to create a searchable data structure: Identification of processing tokens, characterization of the tokens, & stemming of the tokens. The processing tokens & their characterization are used to define the searchable text from the total received text.

The fig shows the normalization process.

→ Standardizing the i/p takes the different external formats of i/p data & performs the translation to the formats acceptable to the system.

A system may have a single format for all items & allow multiple formats.

for example: translation of foreign lang into unicode.

→ The next process is to parse the items into logical sub-divisions that have meaning to the user. This process, called "zoning", is visible to

the user and used to increase the precision of a search and optimize the display.

for example: A typical item is sub-divided into zones, which may overlap & can be hierarchial. such as, Title, Author, Abstract, main Text, conclusion, & References.

⇒ Next, a stop list/Algorithm is applied to the list of potential processing tokens. The objective of the stop fun' is to save system resources by eliminating from the set of searchable processing tokens those that have little value to the system. for example: any word found in almost every item would have no discrimination value during a search. ex: "The"; have no search value and are not a useful part of the user's query.

The majority of unique words are found to occur a few times. The rank-frequency law of Ziph is:

Frequency * Rank = constant.

Ziph → looking at the frequency of occurence of the unique words across a corpus of items.

where frequency is the number of items a word occurs and rank is the rank order of the word

⇒ The next step in finalizing on processing tokens is identification of any specific word characteristics. The characteristic is used in systems to assist in disambiguation of particular word.

⇒ Once the potential processing token has been identified & characterized, most systems apply "stemming algorithms" to normalize the tokens to a standard semantic representation. The amount of stemming that is applied can lead to retrieval of many non-relevant items.

⇒ Once the processing tokens have been finalized, based upon the stemming algorithm, they are used as updates to the searchable data structure. It is the internal representation (i.e, not visible to the user) of items that the user query searches. This structure contains the semantic concepts that represent the items in the database and limits what a user can find as a result of their search.

(ii) Selective Dissemination of Information:

These (mail) process provides the capability to dynamically compare newly received items in the Information System against standing statements of interest of users & deliver the item to those users whose stmt of interest

matches the contents of the item.

The mail process is composed of the search process, user statements of interest (Profiles) and user mail files. As each item is received, it is processed against every user's profile. 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.

User search profiles are different than ad hoc queries in that they contain significantly more search terms and cover a wider range of interests.

When the search statement is satisfied, the item is placed in the mail file associated with the profile. Items in mail files are typically viewed in time of receipt order & automatically deleted after a specified time period (ex: After 1 month) or upon command from the user during display. The dynamic asynchronous updating of mail files makes it difficult to present the results of dissemination in estimated order of likelihood of relevance to the user.

(iii) Document Database Search:-

It 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 & document database which contains all items that have been received, processed & stored by the system. It is the retrospective search source for the system.

If the user is online, the selective dissimination of information system delivers to the user items of interest as soon as they are processed into the system. Any search for information that has already been processed into the system can be considered a "retrospective" search for information. This does not preclude the search to have search statements constraining it to items received in the last few hours.

The Document Database can be very large, hundreds of millions of items or more. Typically items in the document database do not change (i.e., are not edited) once received. The document in the mail files are also in the document database, since they logically are i/p to both processes.

(iv) Index Database Search:

These process provide the capability to create indexes and search them. The user may search the index & retrieve the index and/or the document it references. The system also provides the capability to search the index & then search the items referenced by the index records that satisfied the index portion of the query. This is called a combined file search.

There are 2 classes of index files:

(i) Public index file

(ii) Private index file

(i) Public Index file: These are maintained by professional library services personnel & typically index every item in the Document database. There is a small num of public index files. These files have access lists that allow anyone to search or retrive data.

(ii) Private Index file:

Every user can have one or more private index files. leading to a very large num of files. Each private index file references only a small subset of the total num of items in the document Database. Private Index files typically have very limited access lists.

To assist the users in generating indexes, especially the professional indexers, the system provides a process called "Automatic file Build". Its also called as "Information Extraction".

The documents are processed for extraction of index inform" & the index term extraction process are stored in Automatic file Build (AFB) profiles

The capability to create private & public index files is frequently implemented via structured Database management system. This has introduced new challenges in developing the theory & algorithms that allow a single integrated perspective on the information in the system.

IV. Relationship to Database management systems:

There are 2 major categories of systems available to process items:

i) IRS ii) DBMS

An IRS is a sw that has the features & funs required to manipulate "information" files. Vs a DBMS that is optimized to handle "structured" data.

Information is fuzzy text. The term fuzzy is used to imply the results from the minimal standards & controls on the creators of the text items.

The author is trying to present concepts, ideas and abstractions along with supporting facts.

Structured data is well defined data typically represented by tables. In this, 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.

for example:

There is a no confusion blw the meaning of "employee name" & "employee salary" & what values to enter in a specific database record.

The user has to refine his search to locate additional items of interest. This process is called "iterative search."

The integration of DBMS's and IRS is very important. Commercial database companies have already integrated the two types of systems. One of the first commercial database to integrate the two systems into a single view is the INQUIRE DBMS. This has been available for over 15 years. A more current example is the ORACLE DBMS that now offers an embedded capability called CONVECTIS, which is an IRS that uses a comprehensive thesaurus which provides the basis to generate

themes for a particular item. The INFORMIX DBMS has the ability to link to Retrievalware to provide integration of structured data & information along with fun's associated with IRS.

VI. Digital Libraries and Data Warehouses:

Two other systems frequently described in the context of information retrieval are digital Libraries and Data Warehouses (or Data marts). There is a significant overlap between these & syster and an information Storage & Retrieval System.

All these systems are repositories of information & their primary goal is to satisfy user information needs. As the quantities of information grow exponentially, libraries have alwa were forced to make maximum use of electronic tools to facilitate the storage & retrieval process.

With the worldwide interneting of libraries & information sources (eg: publishers, new agencies) wid the internet, more focus has been on the concept of an electronic library.

Digital libraries (DL) is a metaphor for access to collections of electronic document through an It is an online database of digital objects that can include text, still images, audio, video, digital documents, or other digital media formats.

Information Storage & Retrieval technology has addressed a small subset of the issues associated with digital libraries. It has also ignored the issues of unique identification & tracking of information required by the legal aspects of copyright that restrict fun^n within a library environment. The conversation of existing hard copy text, images (pics, maps) & analog (eg: audio, video) data & the storage & retrieval of the digital version is a major concern to digital libraries which is not considered in Information system.

Data Warehouse:

A data warehouse consists of the data, an information directory that describes the contents & meaning of the data being stored, an i/p fun^n that captures data & moves it to the data warehouse, data search & manipulation tools that allow users, that means to access & analyze the warehouse data & a delivery mechanism to export data to other warehouses, data marts (small & large), & external systems.

Data warehouses are similar to information storage & retrieval systems in that they both have a need for search & retrieval of inform^n.

But a data warehouse is more focused on structured data & decision support technologies.

Information Retrieval System Capabilities:

1. Search Capabilities:

The objective of the search capability is to allow for a mapping between a user's specified need & the items in the inform" database. The search statement may apply to the complete item or contain additional parameters limiting it to a logical division of the item.

Based upon the algorithms used in a system many different fun" are associated with the system's understanding the search statement. The fun" define the relationship between the terms in the search statement and the interpretation (meaning) of a particular word.

translation

Eg: Boolean; Natural language, Proximity, contiguous word phrases, & fuzzy searches; Term masking, Numeric & Date range & Concept / Thesaurus expansion).

i) Boolean logic:

It allows a user to logically relate multiple concepts together to define what information is needed. The boolean functions apply to processing tokens identified anywhere within an item.

The Boolean operator are AND, OR, & NOT. These operations are implemented using set

intersection, set union & set difference procedure.

A special type of Boolean search is called "m" of "N" logic. i.e; any item that contains a subset of the terms.

For example: "Find any item containing any 2 of the following terms: "AA", BB, "CC". This can be expanded into a boolean search that performs an AND b/w all combination of two terms & OR the result together. (AA AND BB) OR (AA AND CC) OR (BB AND CC)

Here some search example & their meaning are given in below:

Search Statement	System operation
COMPUTER OR PROCESSOR NOT MAINFRAME	Select all items discussing computer and/or processors that do not discuss mainframes.
COMPUTER OR (PROCESSOR NOT MAINFRAME)	Select all items discussing computers and/or items that discuss processors & 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.

(ii) Proximity:

It is used to restrict the distance allowed within an item b/w two search items. The semantic concept is that the closer two terms are found in a text the more likely they are related in the discription of a particular concept.

If terms COMPUTER and DESIGN are found within a few words of each other then the item is more likely to be discussing the design of computers than if the words are paragraphs apart.

The typical format for proximity is:

TERM1 within "m" "units" of TERM2

The distance operator m is an integer number & units are: in characters, words, sentence, or paraphas.

Ex: Search statement

"Venetain" ADJ "Blind"

System operation would find items that mention a venetain Blind on a window but not items discussing a Blind venetain.

"United" within five words of "American"

would hit on "united states and American" interests, "United Airlines & American Airlines", Not on "United States of America & the American dream."

"Nuclear" within zero paragraphs of "clean-up"

would find items that have "nuclear" & "clean-up" in the same paragraph.

fig: Use of proximity.

iii) contiguous word phrases: [CWP]

A CWP is both a way of specifying a query term & a special search operator. A CWP is two or more words that are treated a single semantic unit.

Ex: "United states of America"

It is 4 words that specify a search term representing a single specific semantic concept (country).

A CWP also acts like a special search operator that is similar to proximity (Adjacent) operator. It is also called as literal strings.

iv) Fuzzy Searches:

It provides the capability to locate spellings of words that are similar to the entered search term. This function is primarily used to compensate for errors in spelling of words.

Example: A fuzzy search on the term "Computer" would automatically include the following words from the information database: "Computer", "Compiter", "conputer", "computter", "compute".

Fuzzy searching has its maximum utilization in systems that accept items that have been Optical Character Read (OCRed). In the OCR process a hardcopy items is scanned into binary image.

Term masking:

It is ability to expand a query term by masking a portion of the term and accepting as valid any processing tokens.

The value of term masking is much higher in systems that donot perform stemming.

- There are 2 types of search term masking:

(i) fixed length (ii) variable length.

Sometimes they are called fixed & variable length "don't care" functions.

Fixed length: It is a single position mask. It masks out any symbol in a particular position or the lack of that position in a word. It not allows any character in the masked position, but also ~~proo~~ accepts words where the position does not exist. This term masking is not frequently used and not critical to a system.

Variable length: Variable length dont cares allow 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 imbeded. These cases are called suffix search, prefix search & imbedded character string search respectively.

Ex: "*COMPUTER" - Sufix search - change word

"COMPUTER*" Prefix search.

"*COMPUTER*" Imbeded string search.

here * represents a variable length.

below fig shows both fixed & variable length

Search Statement	System operation
multi$national	matches "multi-national", "multy national", "multinational" but does not match multinational since it is two processing tokens.
computer	matches "minicomputer", "microcomputer" or "computer"
Comput*	matches "computers", "computing" "computes"
Comput	matches "microcomputers" "minicomputing", "computes"

fig: Term masking

(vi) Numeric and Database Ranges:

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

A user could enter inclusive to infinite ranges.

"125-425" or "4/2/93 - 5/2/95" for numbers & dates.

">125", "<=233", using greater than or Less than equal to

(vii) Concept / Theasurus Expansion:

The theasurus is typically a one-level or two-level expansion of a term to other terms that are similar in meaning.

A concept or class is a tree structure that expand each meaning of a word into potential concepts.

The below figs shows Theasurus & concept of class concepts.

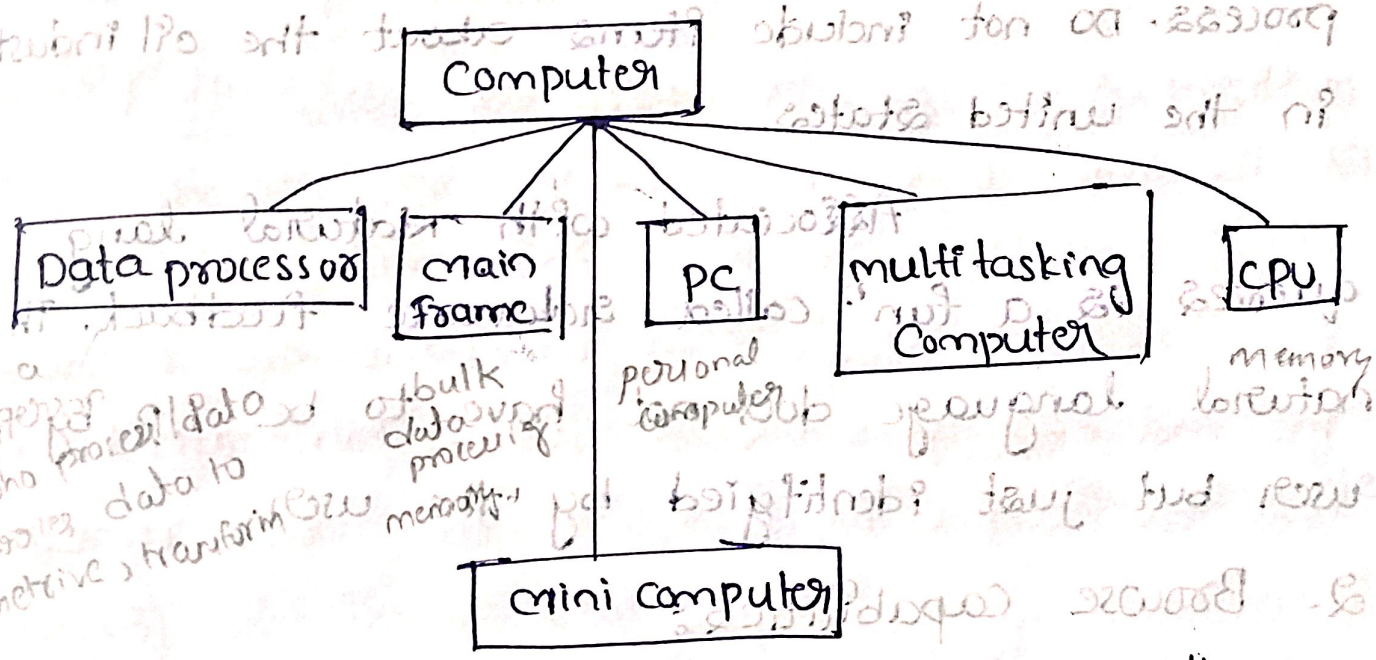


fig: Theasurus for term "computer".

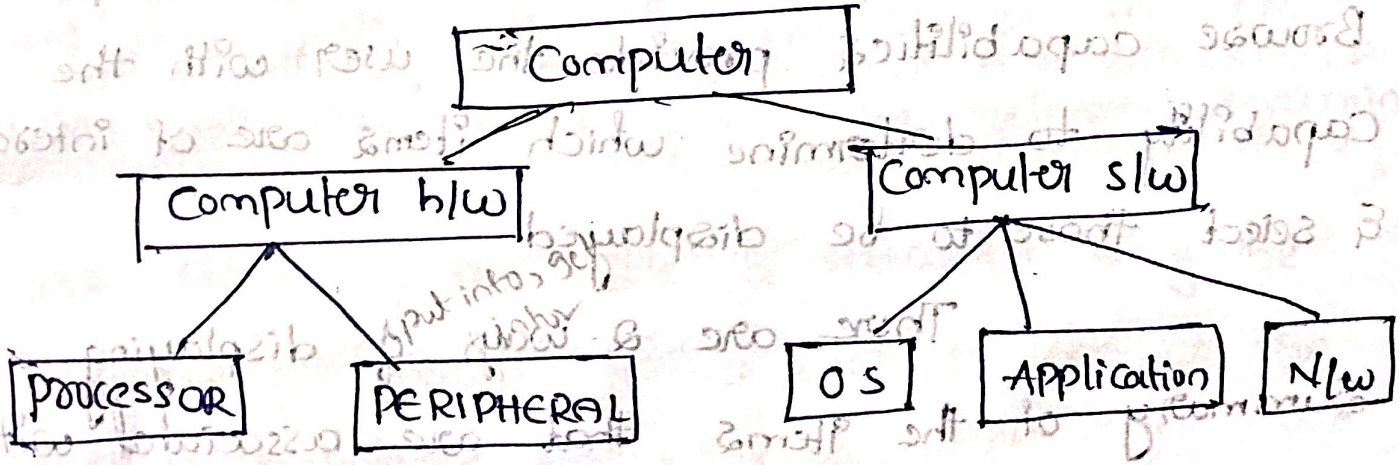


fig: Hierarchical concept class structure for "Computer".

(viii) Natural language Queries:

It allows a user to enter a prose statement that describes the information that the user wants to find.

Examples: find for me all the items that discuss oil reservers and current attempts to find new oil preserves. 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.

Associated with Natural lang queries is a fun^{n} called relevance feedback. The natural language does not have to be i/p by the user but just identified by the user.

2. Browse capabilities:

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

There are 2 ways of displaying a summary of the items that are associated with a query: line item status & data visualization.

From these summary displays, the user can select the specific item & zones within the items for display.

(i) Ranking:

Under Boolean systems, the status display is a count of the number of items found by the query. The relevance score is an estimate of search system on how closely the item satisfies the search statement.

(ii) Zoning:

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

Limited display screen sizes require selectability of what portions of an item a user needs to see to make the relevance determination.

Ex: display of the title & Abstract may be insufficient inform' for a user to predict the potential relevance of an item.

Limiting the display of each item to these two zones allows multiple items to be

displayed on a single display screen. This makes maximum use of the speed of the user's cognitive process in scanning the single image & understanding the potential relevance of the multiple items on the screen.

(iii) Highlighting:

Highlighting is another display aid is an indication of why an item was selected This indication, frequently highlighting, let the user quickly focus on the potentially relevant parts of the text to scan for item relevance. Highlighting has always been useful in Boolean systems to indicate the cause of the retrieval. This is because of the direct mapping b/w the terms in search & the terms in the item.

The highlighting may vary by introducing colors & intensities to indicate the relative importance of a particular word in the item, in the decision to retrieve the item.

3) Miscellaneous capabilities:

There are many additional fun^{n} that facilitate the user's ability to i/p queries, reducing the time it takes to generating

the queries. and reducing a priori the probability of entering a poor query.

In this miscellaneous capabilities provides 3 types

> Vocabulary Browse
> Iterative Search and Search History log
> Canned Query

Vocabulary Browse:

It provides the capability to display in alphabetical sorted order words from the document database. All unique words in the database are kept in sorted order along with a count of the number of unique items in which the word is found.

In below fig shows in vocabular browse if the user enters "comput"

TERM	OCCURRENCES
Compramise	53
Compulsive	22
Compulsory	4
Comput	
Computation	265
Compute	1245
Computen	1
Computer	10,800
Computerize	18
Computed	29

fig: Vocabular Browse list with entered term Comput

Vocabulary browse provides information on the exact words in the database. It helps the user determine the impact of using a fixed or variable length mask on a search term & potential mis-spellings.

The user can determine that entering the search term "compul*" in effect is searching for "compulsion" or "compulsive" or "compulsory." It provides insight on the impact of using terms in a search.

(ii) 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 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 i/s of refining the result of a previous search to focus on relevant items is called iterative search.

This also applies when a user uses relevance feedback to enhance a previous search.

During a login session, a user could execute many queries to locate the needed information. To facilitate locating previous searches as starting points for new searches, search history logs are available. The search history log is the capability to display all the previous searches that were executed during the current session. The query along with the search completion status showing no of hits is displayed.

(iii) Canned Query:

The capability to name a query & store it to be retrieved & executed during a later user session is called "canned or stored queries".

Users tend to have areas of interest within which they execute their searches on a regular basis. A canned query allows a user to create & refine a search that focuses on the user's general area of interest one time & then retrieve it to add additional search criteria to retrieve data that is currently needed.

Example: A user may be responsible for European investments. Rather than always having to create a query that limits the search to European geographic search terms & then the specific question req resol, a canned query can be created with all the needed geographic terms & used as starting point for additional query specification.