Unit-II

Cataloging and Indexing

History and objectives of indexing:-

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

→ Indexing (originally called cataloging) is the oldest technique for identifying the contents of items to assist in their retrieval.

→ The objective of cataloging is to give access points to a collection that are expected and most useful to the users information.

→ The basic information required on item is, what is the item and what it is about, has not changed over the centuries.

As early as the Hyman-89:-

* Third-millennium (in Babylon):-

The Libraries of cuneiform tablets were arranged by subject. (Hyman-89).

• Cuneiform tablets indicates the logo-symbolic script, used in Ancient.

* Upto 19th century there was little advancement in cataloging.

i.e only changes in the methods used to represent the basic information. [Norris-69].

* In the Late 1800's :- Subject indexing became hierarchical.

ex:- Dewey Decimal System (DDS)

*In 1963:-
The Library of "Congress initiated" a study on the computerization of bibliographic surrogates.

*From 1966-1968:-
The Library of congress ran its "MARC I pilot project".

MARC → (MAchine Readable Cataloging).

Here,
MARC standardizes
* the structure
* Contents &
* Coding of bibliographic records.

*In 1969:-
The system became Operational. [Avram-75]

*In 1965:- (DIALOG)
The earliest commercial cataloging system is DIALOG, which was developed by "Lockhead Corporation" for NASA.

*In 1978:-
It became (i.e, DIALOG) commercial with three government files of indexes to technical publications.

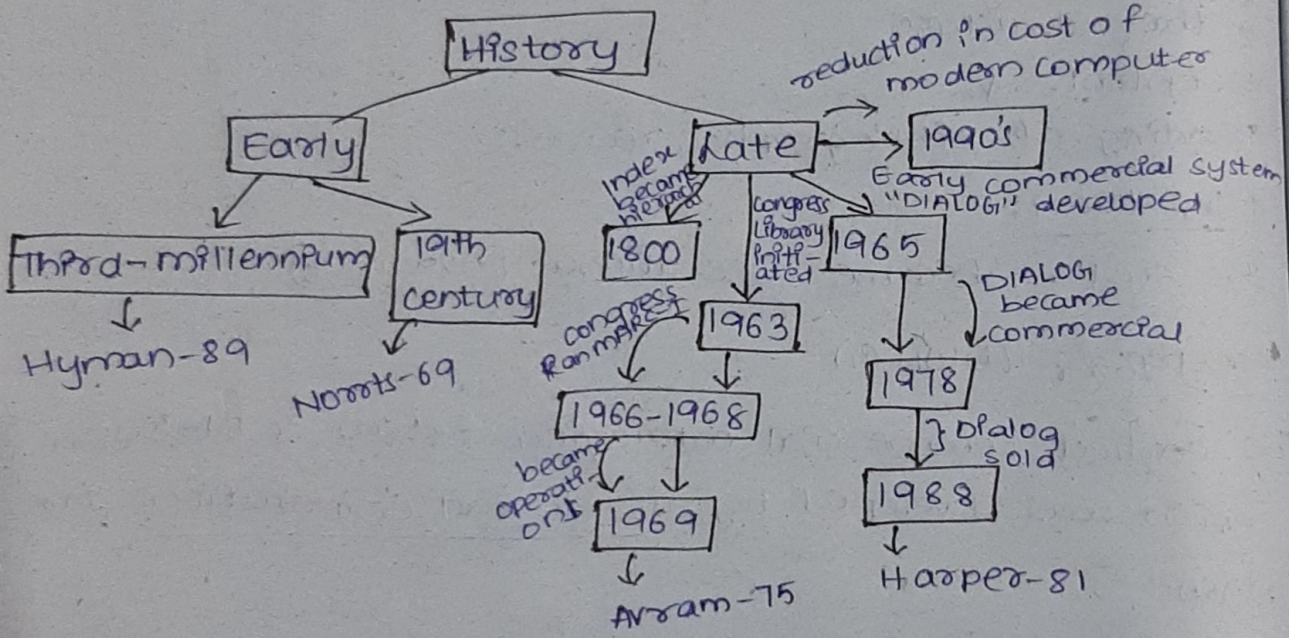
*In 1988:-
→ DIALOG sold to ~~KPA~~ Knight-Ridder, and it contained over 320 index databases used by over 91,000 subscribers in 86 countries. [Harper-81].

→ Indexing (Cataloging) until recent:-
The indexing was accomplished by creating a bibliographic citation in a structured file that references the original text.

- Citation means source information.
- These files contains citation information. "about item", "keywording the subjects of item" and "length of free text field used for abstract/summary".

*In 1990's :-

A significant Reduction in cost of processing power and memory in modern computer.



→ The indexing process is typically performed by professional indexers associated with library organizations.

Objectives :-

The objectives of indexing have changed with the evolution of IRS.

→ Availability of the full text of the item in searchable form alters the objectives historically used in determining guidelines for manual indexing.

→ In manual indexing environment, the use of controlled vocabulary makes the indexing "process slower", but it potentially simplifies the "search process".

→ The availability of items in "electronic board" changes the objectives of manual indexing.

→ The words used in an item do not always reflect the value of the concepts being presented.

Indexes < public file index
private file index

→ The public file indexer needs to consider the information needs of all users of library system.

Items overlap b/w full item indexing, public & private indexing of files.

→ users may use "public index files" as part of search criteria to increase recall.

→ They users can constrain there search by private index files.

→ The primary objective of representing the concepts with in an item to facilitate users finding relevant information.

→ The other objectives of indexing :-

ranking, item clustering.

* Indexing process :

→ Indexing is an important process in information Retrieval (IR) systems.

→ The indexing process is typically performed by professional indexers associated with library organizations

→ The indexing process depends on 3 thing/files.

i.e,
- Document file
- public Index file
- private Index file.

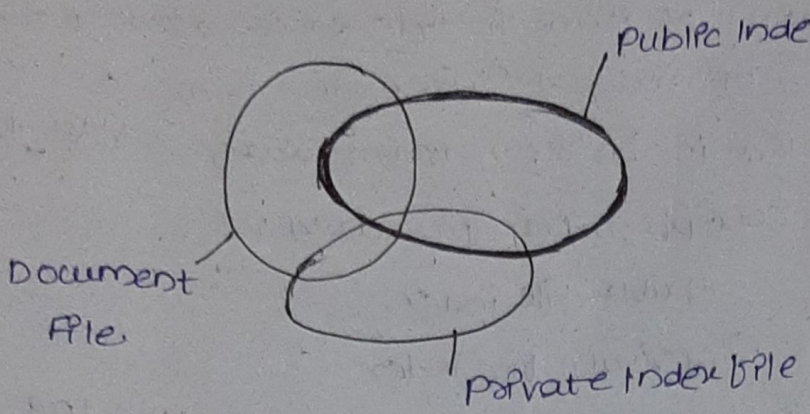


fig:- Indexing process.

⇒ The above diagram showing the items overlap b/w full item & indexing, public file indexing and private file indexing.

⇒ Linking index terms is needed when there are multiple independent concepts found with in an item.

Document file:-

The full text searchable data structure for items in "The Document file provides a new class of indexing called total document indexing".

Public Index files:-

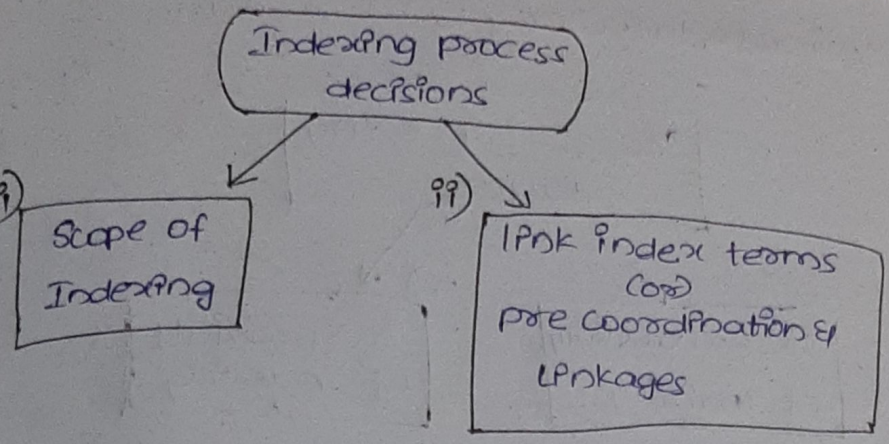
Users may use public index files as part of their search criteria to increase the recall.

Private Index files:-

Users may want to constraint the search by their private index file" to increase the precision of search/ Precision.

⇒ Indexing process takes the 2 types of decisions.

i.e



Manual Indexing:- The process of reliably & consistently determining the bibliographic terms that represents the concepts in an item is extremely difficult.

1) Scope of Indexing:-

→ The first decision is the scope of the indexing to define what level of detail the subject index will contain.

→ This is based upon usage scenarios of the end users.

→ When indexing is performed manually, the problems arise from two sources:

i.e "the author" and "the indexer".

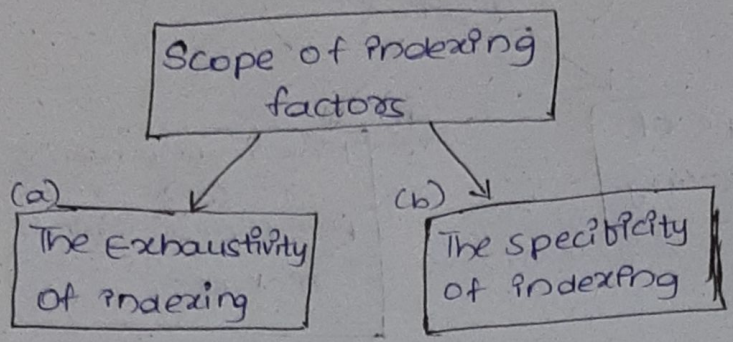
→ The vocabulary domain of the author may be different then that of indexer.

→ So the results is in different quality levels of indexing.

→ The indexer is not an expert on all areas & the indexer must determine when to stop the indexing process.

→ There are two factors involved in deciding on what

level to index the concepts in an item.



(a)
The Exhaustivity:

Exhaustivity of indexing is the extent to which different concepts in the item are indexed.

for ex:

If two sentences of a 10-page item on microprocessors discusses on-board caches - should this concept be indexed.

(b) The specificity:

Specificity relates to the preciseness of the index terms used in indexing.

for ex:

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.

⇒ High Exhaustivity and Specificity indexes almost every concept in the item using as many detailed terms as needed.

⇒ "Low Exhaustivity" has an adverse effect on both precision & recall.

⇒ "Low specificity" has an adverse effect on precision but no effect to a potential increase in recall.

⇒ Another decision on indexing is, what portion of an item should be indexed.

⇒ The simplest case to limit the indexing is to

* The title (or)
* The title & Abstract zones

⇒ This case leads to loss of both precision & recall

ii) pre coordination and Linkages (or) Link index terms:-

The another decision is the need to link index terms together in a "single index" for a particular concept

Linkages are used to correlate related attributes associated with concepts discussed in an item.

→ The process of creating "term linkages" at index creation time is called precoordination.

→ When index terms are not co-ordinated at index-time, the co-ordination occurs at search time is called Post Coordination.

→ Post coordination is implemented by "AND"ing index terms.

i.e. which only finds indexes that have all of search terms.

For ex:

The capability to link the source of a problem, the problem & who is affected by the problem may be desired.

• Linkage of Index Terms:

INDEX TERMS	Methodology
Oil, wells, mexico, CITGO, refineries, peru, BP, drilling	No linking of terms.
(Oil wells, mexico, drilling, CITGO) (U.S., oil refineries, peru, introduction)	linked (pre coordination)
(CITGO, drill, oil wells, mexico) (U.S., introduction, oil refineries, peru) (SUBJECT : CITGO; ACTON : drilling; OBJECT : oil, wells; MODIFIER : in mexico) (SUBJECT : U.S.; ACTION : introduces; OBJECT : oil refineries; MODIFIER : in peru)	linked (pre coordination) with position indicating role.

fig: Linkage of index terms.

The above fig shows the different types of linkages.

→ It assumes that an item discusses the drilling of oil wells in 'mexico' by CITGO & the introduction of oil refineries in 'peru' by the U.S.

→ when the linked capability is added, the system does not erroneously relate peru & mexico since they are not in the same set of linked items.

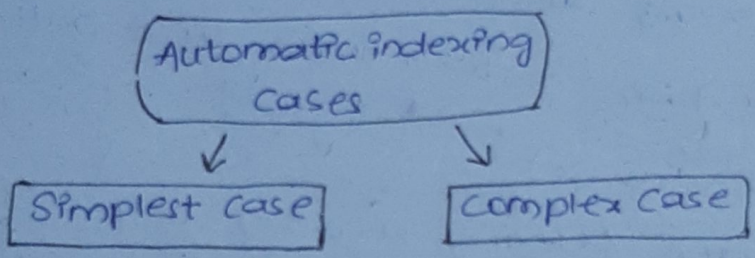
⇒ Positional role:

The U.S was introducing oil refineries in Peru, Bolivia and Argentina then the positional role technique would require three entries, only the difference is affected country position.

Here all 3 countries would be listed with three "MODIFIERS".

* Automatic Indexing

→ Automatic indexing is capability to automatically determine the index terms to be assigned to an item.



Simplest Case:

It deals with "total document indexing".

(or)

All words in the document are used as possible index terms.

Complex Case:

In this process the objective is emulate a "human indexer" and determine a limited no. of "index terms" for major concepts in the item.

Here we 1 advantage & disadvantage of human indexing.

Advantage: It has ability to "determine concept abstract and judge the value of a concept.

Disadvantage: Cost, processing time & consistency over the automatic indexing.

Human Indexing Vs Automatic Indexing:-

→ The processing time of an item by a human indexer varies significantly based upon the "Indexer's knowledge" of the concepts being indexed with the exhaustivity & specificity guidelines & the amount & accuracy of processing via Automatic File Build.

→ Usually it take at least "five minutes" per item.

i.e. Human iteration with computer.

eg:- typing speeds, cursor positioning, correcting spelling errors, taking breaks b/w activities.

for short items it take 5min time pattern.

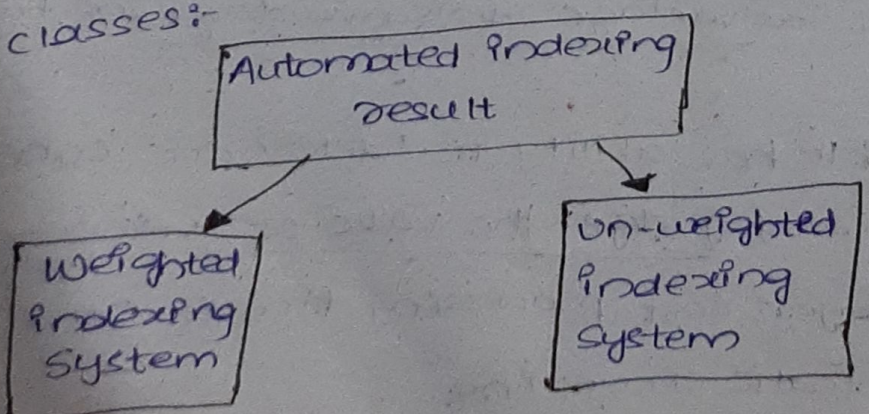
eg:- (300-500 words)

→ Automatic indexing requires only a "few seconds" (or) less of computer time based upon the size of the processor and the complexity of the algorithms to generate the index.

⇒ If indexing is being performed automatically by an algorithm, there is consistency in the index term selection process.

Human indexers typically generate different indexing for the same document.

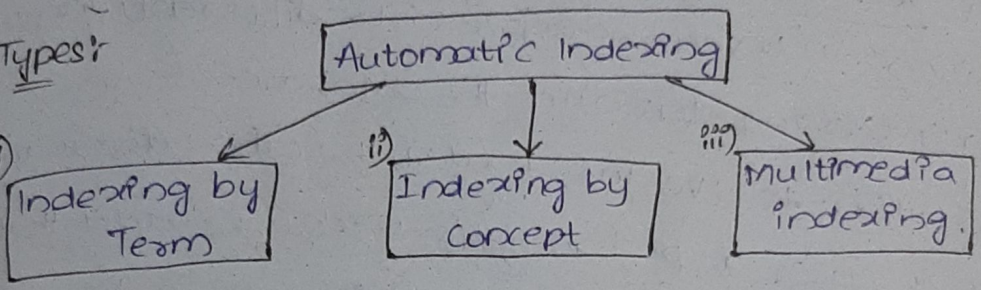
⇒ Indexes resulting from automated indexing fall into two classes:-



Weighted Indexing System:- Weight is based on the function associated with the frequency of occurence of item. In a weighted indexing system an item is made to place a value on the index terms that represents concept in the document.

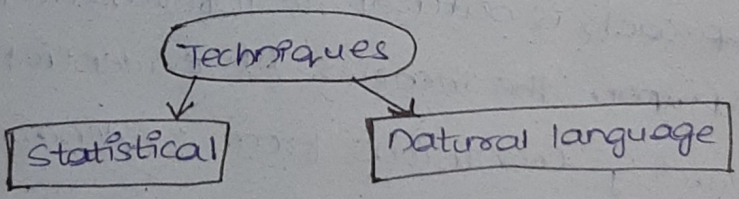
Unweighted indexing System:- In an un-weighted indexing system, the existence of index term in a document & some times its "word" location(s) are kept as part of the searchable datastructure.

Types:-



i) Indexing by Term:-

The terms of the "original item" are used as a basis of the index process. There are 2 major techniques used for creation of the index.



*Statistical Techniques:-

⇒ These techniques can be based upon "vector models" and "probabilistic models" with a special case being "Bayesian models".

⇒ They are classified as "statistical" because their calculation of weights use statistical information such as frequency of "occurence of words" & their distributions in the "searchable DS".

⇒Often weighted systems are discussed as vectorized information systems.

This association comes from the "SMART system" at cornell university created by Dr. Gerald Salton. (i.e Salton → 73, Salton → 83).

→ The system emphasizes weights as a foundation for "information detection" & stores these weights in a "vector form".

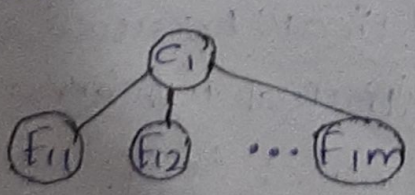
Each "vector" represents a document and each "position" in a vector represents different unique word (processing token) in the database.

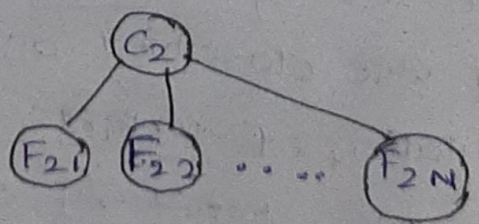
⇒In addition to a vector model, the other dominant approach uses a probabilistic model. The model that has been most successful in this area is the Bayesian Approach.

Bayesian Approach:-

This approach is natural to information system & is based upon the theories of evidential reasoning (i.e. drawing conclusions from evidences).

A Bayesian n/w is a directed acyclic graph in which each "node" represents a random variable and the "arcs" blw "nodes" represents a probabilistic dependence blw the node and its parents.





Here
Nodes C_1, C_2 represents
"the item contain concept C_i".

F nodes represent "The item has feature F_{ij}"
eg: words.

The goal is to calculate the probability of 'C_i' given F_{ij}'. To perform that calculation two sets of probabilities are needed:

1. prior probability "$P(C_i)$" ⇒ that item relevant to concept C_i.
2. Conditional probability $P(F_{ij}/C_i)$ ⇒ that the features F_{ij} where $j = 1, n$ are in an item contains topic C_i.

* Natural language processing / Technique:
→ It process the items at the
* morphological,
* Lexical,
* Semantic,
* Syntactic &
* discourse levels.

→ Each level uses information from the previous level to perform its additional analysis.
→ The discourse level is abstracting information beyond the sentence.

ii) Indexing by Concept:-

The basis for concept indexing is that, there are many ways to "express the same idea" & increase the retrieval performance comes from single representation.

→ Here Indexing by term treats each of these occurrences as a "different index" & then uses the Saurus (or) other query expansion techniques to expand the query.

→ Concept indexing determines a canonical set of concepts based upon a test set of terms & uses them as basis for indexing all items.

iii) Multimedia Indexing:-

Indexing associated with multimedia differs from the previous discussions of indexing.

Indexing video (or) images can be accomplished at the raw data level, (eg: aggregation of raw pixels) & the feature level distinguish primitive attributes. Such as "color" & "luminance" & at the semantic level meaningful objects are recognized.

ex:- An example is "processing of video".

→ The system will periodically collect a frame of video i/p for processing.

It might compare that present frame to the last frame captured, to determine the difference b/w the frames.

→ If the difference is below threshold it will discard the frame.

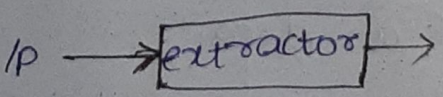
⇒ processing of image:-

To processing an image, semantic level indexing requires for the pattern recognization of objects with in the images.

ex: photobook (pentland-94)
IBM's QBIC.

Audio:
⇒ If you consider an analog audio i/p : the system will convert the audio to digital format.
⇒ In multimedia indexing two mechanisms are used "positional" and "temporal".
→ The first mechanism is used to represent "a linear sequential composition".
→ The second mechanism is based upon 'time'. because the modalities are executing concurrently.
modalities means mode in which something exists.
⇒ The typical video source off television is inherently a multimedia source.
It contains video, audio & potentially closed captioning.

Information Extraction:
IE is the task of automatically extracting structured information from unstructured (or) semi-structured documents & other sources and store them in a database.



*Information Extraction:
There are two processes associated with the information extraction.
i.e., 1. Determination of facts to go into structured fields in a database.

2. Extraction of text, that can be used to summarize an item.

⇒ In first case, only a subset of the important facts in an item may be "identified and extracted" where as in summarization all of the major concepts in the item should be represented in the summary.

⇒ The process of extracting facts to go into indexes is called Automatic File Build (AFB). Its goal is to process incoming ~~inter~~ items & extract "index terms" that will go into a structural database.

⇒ An IRS goal is to provide an indepth representation of the total content of an item.

⇒ An information Extraction system, only analyzes those portions of a document that potentially contain information relevant to the extraction criteria.

⇒ The term "slot" is used to define a particular category of information to be extracted.

Here

→ Slots are organized into "templates" (or) "Semantic frames".

→ Information Extraction requires multiple levels of analysis of text of an item.

→ It must understand the words & their context.

→ The processing is very similar to the Natural language processing, that described under Indexing.

Metrics to Compare Information Extraction:-

The previously defined measures of precision & recall are applied with slight modifications to their meaning.

Recall:

Recall Refers to how much information was extracted from an item versus' how much should have been extracted from the item.

precision:

precision refers to how information was extracted accurately "versus" total information extracted.

Additional metrics:-

The additional metrics use are over generation & fallout.

Over Generation:

Over Generation measures the amount of irrelevant information that is extracted.

Fallout:

Fallout measures how much a system assigns "incorrect slot fillers" as no. of potential incorrect slot fillers increases.

⇒ Another related information technology is document summarization.

The goal of document summarization is to extract a summary of an item maintaining the most important ideas that significantly reducing the size.

⇒ ~~The term "slot" is~~

Data Structure

* Introduction to Data structure:

Datastructure:

→ A datastructure is a specialized format for organizing, processing, retrieving & storing data.

→ There are usually "two" major datastructure in any information system.

* One data structure stores & manages the received items in their Normalized form.
This process is called the "document manager".

* The other major data structure contains the processing tokens & associated data to support search.
This process is called the "document search manager".

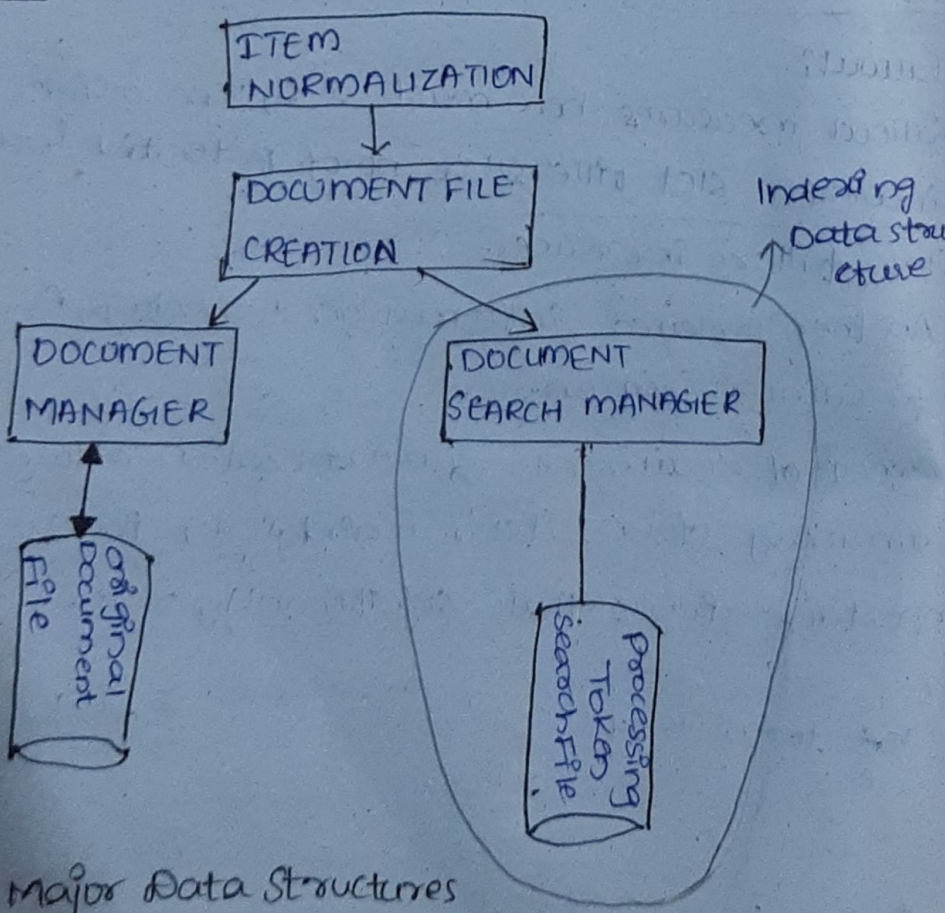


Fig:- Major Data Structures

→ The above fig expand the Document file creation.
i.e document manager &
document search manager.
The result of search are references to the items that satisfy the search statement which are passed to the document manager for retrieval.
→ The most common data structure encounters in both "database" and information systems is the "Inverted file system". →(4th chapter)

Inverted File System:-
It minimizes secondary storage access, when multiple search terms are applied across the total database.
→ A various of the Searchable data structure is the "N-gram structure".
The N-gram structure that breaks processing token into smaller string units & uses the token fragments for search.
→ A "special data structure" that is becoming common place because of its use on the internet is "Hypertext".

Various Data structures are:-

* Stemming algorithms
* Inverted Structures
* N-Gram data structures
* PAT data Structures.

*Signature data structures
*Hyper text data structures.

Stemming Algorithms:-

Stemming means cutting and trimming.
i.e. un-necessary things will be trim/stem.
→ The concept of stemming is introduced in the 1960's.
→ The main goal of stemming is to improve performance & it require less system resources by reducing the no. of unique words that the system contain.
→ Stemming algorithms are used to improve the efficiency of the information system & to improve recall.

Stemming:-

→ Stemming is the process of producing morphological variants of a root/base word.
→ stemming programs are commonly referred as stemming algorithms (or) stemmers.
i.e. reduce the word forms to proper ones.
→ The concept of stemming has been applied to the information systems from 1960's.
→ The original goal of stemming was to improve performance by reduce the no. of resources that system contain.

Introduction to the stemming process:-

stemming algorithms are used to improve the efficiency of the information system.

The another major use of stemming is to improve the recall.

→ stemming is a technique used to extract the base form of the words by removing affixes from them.

It is just like cutting down the branches of a tree to its stems.

for ex:- ① The stem of words

eating, eats, eaten is **eat**

② The stem "comput" could associated with computable, computability, computation, computational, computed, computing, computer, computerse, computerize to one compressed word

i.e. **comput**

③ Stemming/stem word "calculat"

"calculate, calculates, calculation, calculations, calculating."

→ The most common stemming algorithm removes suffixes & prefixes sometimes recursively to derive the final stem.

⇒ The other Techniques such as

"Table/Dictionary look-up n stemmers and

"Successor Stemming/nstemmers provide ~~attentive~~

alternatives to the stemming algorithm

Porter stemming Algorithm:-

It is one of the most popular stemming method proposed in 1980.

It is based upon set of conditions of the stem, Suffix and prefix & associated actions given the condition.

Some examples of stem conditions are:

1) The measure 'm' of a stem is a function of sequences of vowels (a, e, i, o, u, y) followed by a consonant.

i.e. a, e, i, o, u (V), y (C)

If 'V' is a sequence of vowels and 'C' is a sequence of consonants.

Then 'm' is: $C(VC)^m V$

Where, the initial 'C' and final 'V' are optional 'm' is the no. of 'VC' repeats.

Measure	Example
m=0	C C V V tree (This word doesn't contain any combination of VC).
m=1	C C V V C Trees, C C V V C C V TROUBLE
m=2	C V C C V C V compute

2) *<X>

Stem ends with letter X.

3) *V*

Stem contains a vowel.

∴ Based on 'm' value the porter stemmer will work.

4) *d

Stem ends in "double consonant"

5) *o

Stem ends with "consonant-vowel-consonant" sequence.

(CVC)

⇒ 2nd 'C' should not be w, x, y.

where the final consonant is not

w, x, (or) y

ex:- WIL, HOP

CVC CVC

Steps in porter stemmer

The porter stemmer: step 1(a)

Conditions:-

① SSES → SS

② IES → I

③ SS → SS

④ S → ε

① SSES → SS:-

If we have a word/term that is ending with "SSES" is simply replaced by "SS".

ex:- caresses

i.e caresses after applying the porter stemmer the word became caress

i.e caresses → caress

SS ↑

② IES→I :-

If we have a coord(or) term that is ending with "IE" is replaced by "I".

ex:- Ponies → Poni

ties → ti

③ SS→SS :-

If we have a word/term that is ending with "SS" is remains same i.e, "SS"

ex:- caress → caress

④ S→ϵ :- (i.e. null) :-

If the stem is ending with single 's' then it will replaced by Null 'ϵ'.

ex:- cats → catϵ

Step 1(b)

Condition's :-

① (m>1) EED → EE

② (*v*)ED → ϵ

③ (*v*)ING → ϵ

Step 1(b) (stemming with Conditions)

① (m>1)EED → EE (condition1)

If the measurement of stem is '>'1 (i.e. combination of 'vc' is appearing >1) then it is ending with "EED" will be replaced by "EE".

ex:- "Agreed"

Check 'vc' combination i.e, Agreed (v c c v v c)

∴ m=2 & our condition is m>1

2>1

i.e 2>1 is satisfied.

∴ EED is replaced by EE.

i.e Agreed → Agree

∴ Condn verified : Agreed → Agree.

ex:- "Feed"

Check 'vc' combination i.e. Feed (c v v c) ∴ m=1

our condition is m>1 but we have only '1' VC combination 1>1 × not satisfied.

∴ So the word is remains same.

i.e Feed → Feed

∴ condition not verified : Feed → Feed.

(2) (*v*)ED → ε (condition 2)

If our stem containing vowel followed by "ED" then it will be replaced by 'ε'.

ex:- "plastered"

The above stem containing vowels(V) (i.e plastered (c c v c c v c v c) & ε 'a' and 'e') followed by "ED" is replaced by null, so the word will become "plaster".

i.e plastered → plaster

condition verified : plastered → plaster.

ex: "bled"

In the above word check 'vowels', but in this word non-of them is vowel so it is remain

Same.

i.e bled there is no vowel

i.e bled → bled

Condition not verified : bled → bled.

③ (*V*) ING → E (Condition 3):-

If our stem containing vowel followed by "ING" then it will be replaced by 'E'.

ex:- "Motoring"

→ Find out "ING" is given word Motoring after find-out "ING" the remaining stem contains vowel 'v'

i.e Motor ing (c v c v c i i g)

So the "ING" is replaced by null.

i.e. Motoring → Motor

Condition verified : Motoring → Motor.

ex:- SING

→ findout "ING" in given word i.e SING and the remaining stem contains only 's' and it is "not vowel" i.e S ING (c v c i g)

This rule will not be applied here and the word will be remain same.

i.e. SING → SING

Condition not verified : SING → SING

Step 1(c) (clean up)

Conditions

① AT → ATE

② BL → BLE

③ (*d & ! (*L or *S or *Z)) → single letter.

(9) (m=1 & *o) → E

These rules are ran if second (or) Third conditions in "1b" apply.

(1) AT → ATE

Here our word is conflated → conflate

check the word "conflated" with 2nd condition in 1b i.e (*v*) ED → E

so the word became conflat(ed) → conflat

the remaining stem conflat contains 'v' vowel and "ED" is replaced by 'ε' i.e conflated → conflat (ε)

then convert the word "conflat" into proper word.

i.e conflate

∴ conflat → conflate [∵ AT → ATE]

(2) BL → BLE :-

ex: Troubling → Trouble

Here take a word is "Troubling".

i.e Troubling → Trouble

check the word "Troubling" with "3rd" condition in 1b.

i.e (*v*) ING → ε

So the word became Troubling → Troubl

Findout the 'ING' in given word and then check remain stem i.e "Troubl" contain vowel (v) and then replace "ING" by "ε"

i.e Troubling → Troubl (ε)

then convert "troubl" into proper word

i.e Trouble.

∴ Troubl → Trouble [∵ BL → BLE]

③ (*d & !(*L or *S or *Z)) → single letter.

The stem is ending with double consonant (*d) and that consonant is not *L (or) S (or) Z then we convert it into single letter.

For Example(1): "Hopping."

Here the word hopping contains "ING" so verify by this Hopping with 3rd condition step 1b(step)

Hopping

→ find out "ing" and the remaining word 'Hopp' then check "v" vowel in remaining stem.

i.e C V C C
Hopp so in hopp vocoel is there and "ING" will be replaced by ε.

Hopping → Hopp

Here the word "Hopp" Stem is ending with double consonants C V C C
Hopp and that consonant is not L (or) S (or) Z then we convert it into single letter.

i.e pp → p

Hopp → single letter
↳ (Hop)

condition verified: Hopping) → hop

ex:- tanned

C V C C V C
i.e check V & C → in above word tanned Here the term is ending with "Ed" & check the remaining stem contains "v" vowel followed by "ED". So it will be replaced by null "ε".

tanned → tann
E

Here the stem is ending with double consonants i.e "nn" and then convert double consonant into single letter / consonant.

i.e tann → tan

∴ Condition verified : tanned → tan.

ex: falling → fall

fall it is ending with double consonant i.e 'l'. But our condition is not "L" (or) "S" (or) "Z". So it will not converted into single letter, so the word is remain same.

∴ condition not verified : fall(ing) → fall.

(4) (m=1 & *o) → E

Here the measure 'm' is 1 (i.e vc combination) and the stem is ending with CVC combination is replaced by single letter.

i.e. CVC → E
C C [not w|x|y]

ex: Filing → file

Filing
CVC m=1

Filing → File

the word is ending with "ing" so the remaining Stem is in cvc combination is replaced by 'E'.

i.e fil is in CVC combination.

i.e Fil → File.

∴ Condition verified : bfling → file.

ex:- Fail → Fail.

Fail
c v v c

The word 'Fail' is not ending with 'CVC' combination. so the condition is not validated and the word will be remain same.

∴ Condition not verified : fail → fail.

Step 1C and 2:-

The porter stemmer : Step1C and 2

Step3C:-

Condition:

Y Elimination (*V*)Y → I :

Our aim is to take out 'Y' from word and the remaining stem contain at least one vowel if this condition is satisfies then 'y' is replaced by I.

For ex:- Happy → Happi

Happy
↳ find out 'Y' in given word
↳ check is vowel is present in remaining word or not if 'v' is there then it will replaced by 'i'.

f v f f f
Happy

so 'y' is replaced by 'i'.

∴ Condition verified : Happy → Happi

ex: SKY → Sky

SKY
! !
c c ↳ I

If we take out 'y' from stem the remaining stem is 'sk' which is not containing vowel. so the rule will not be applied here and the word is remain same

∴ Condition not verified ∴ SKY → SKY.

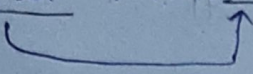
Step 2: Derivational Morphology-I.

conditions:

*(m>0) ATIONAL → ATE

ex:

Relational → Relate



first check the measure 'm' from given word.

How many vc combinations are there. 'm'.

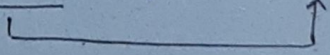
Relational
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ m=4
C V C V C V V C V C

our condition is (m>0)

Here m=4 i.e 4>0 condition satisfied the "ATIONAL" is replaced by "ATE"

*(m>0) IZATION → IZE

ex: Generalization → Generalize



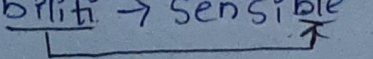
check 'm' value and verify condition then after replace "IZATON" with "IZE".

i.e. Generalization
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
C V C V C V C V C V V C

∴ m=5 our condition is m>0
5>0 satisfied.

*(m>0) BILITI → BLE

ex: Sensibiliti → Sensible



Check 'm' value and verify condition then after replace "BILITI" with "BLE"

i.e Sensibiliti
C V C C V C V C V

∴ m=4 our condition is m>0
i.e. 4>0 satisfied.

The porter stemmer

Step 3 and 4:

Step 3 :- Derivational Morphology - II

Conditions:

(*) (m>0) ICATE → Ic

ex:- triplicate → triplic

i.e. check 'm' value and if it is satisfies condition then "ICATE" is replaced by "Ic".

triplicat
C C V C C V C V C

m=3 our condition m>0 i.e. 3>0 satisfies.

(*) (m>0) FUL → ε

ex:- hopeful → hope
ε

Check 'm' value and verify it with condition

hopeful
C V C V C V C
m=3>0 ✓ satisfied

FUL is replaced by null i.e. 'ε'

∴ hopeful → hope
ε

(*) (m>0) NESS → ϵ

ex: goodness → good
↳ϵ

i.e. check 'm' condition for given word after verify the condition 'ness' is replaced by 'ϵ' null.

goodness
c v v c c v c c
m=2 >0.

Step4: Derivational Morphology - III

conditions:-

(*) (m>0) ANCE → ϵ

ex: allowance → allow
↳ϵ

m=2 >0 ✓

(*) (m>0) ENT → ϵ

ex: dependent → depend
↳null 'ϵ'.

m=3 >0 ✓

(*) (m>0) IVE → ϵ

ex: effective → effect
↳ϵ

m=3 >0

The porter Stemmer:-

Step(5) (clean up):-

Step 5(a):-

Conditions

(*) (m>1) ϵ → ϵ

(*) (m=1 & ! *o) NESS → ϵ

(m>1) ϵ → ϵ

ex:- C V C V C V C
Porbat e → porbat
↳ E

m=2 > 1 then replace E by Null'

(m=1 & ! *O) NESS → E
↳ stem is not ending with "CVC"

ex:- C V V C C V C C
Goodness → Good
↳ E

Here the stem is not ending with CVC so it replaced by E. "NESS" replaced by 'E'.

Step 5(b):-

Condition

(m > 1 & *d & *L) → single letter

Check the condition of 'm' and the stem is ending with 'd' i.e. double consonant and 'L' is replaced by single letter.

ex(1):- Controll → Control

C V C C C V C C
Controll

m=2 > 1 ∴ Condition verified.

The stem is ending with double consonants i.e 'LL' is replaced by single letter, 'L'. Condition verified.

Controll → Control
↳ single L

ex(2):- roll → roll
C V C C

Check the condition of 'm' in this only one combination of 'VC' is there and it is not satisfies i.e m=1 the condition. and the stem remain same. Condition not verified: Roll → Roll

Dictionary look-up Stemmers:-

→ A simple stemmer look-up the inflected form in a look-up table.

→ The "original term" (or) stemmed version of the term is looked up in a dictionary and its replaced by stem.

→ This technique has been implemented in the "INQUERY" and "Retrieval ware systems".

→ The "INQUERY" system uses a stemming technique called "kstem".

K-stem is a morphological analyzer that conflates word variants to a root form.

ex:-

"Memorial" and "memorize" reduced to "memory". But "memorial" and "memorize" are not synonyms and they have very different meanings.