# Information Retrieval Systems
# UNIT-2

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

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

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

  **TYPES OF CATALOGUES:**

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

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

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

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

**INDEXING:**

Indexing is an Important process in Information Retrieval Systems

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

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

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

**History of indexing:**

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

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

**Objectives of Indexing :**

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

•They can constrain there search by private index files

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

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

•They can constrain there search by private index files

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

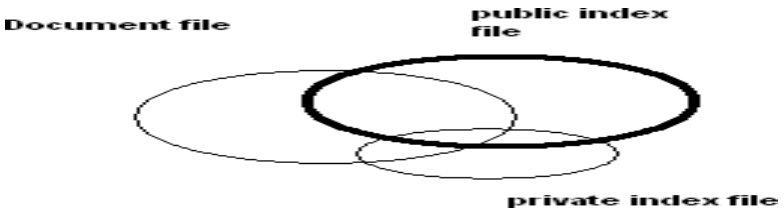



Fig:

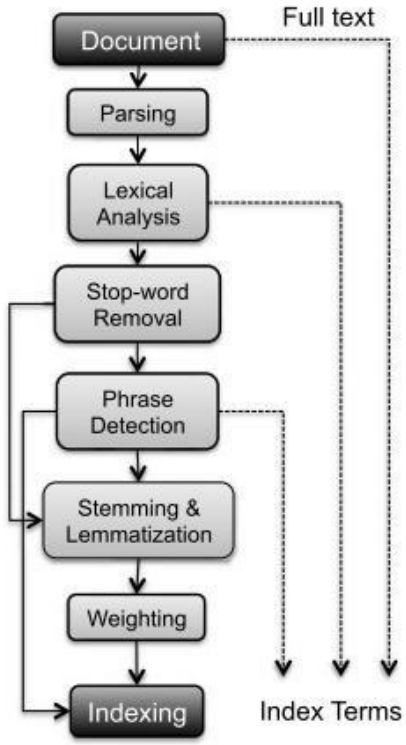
- 1.Decided the scope of indexing and the level of detail to be provided. Based on usage scenario of users.
- 2. Second decision is to link index terms together in a single index for a particular concept.

**Indexing process:**

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

TEXT PROCESSING

**Fig. 2.3** Text processing phases in an IR system



**Text process phases**

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

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

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

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

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

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

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

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

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

**SCOPE OF INDEXING**

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

## AUTOMATIC INDEXING

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

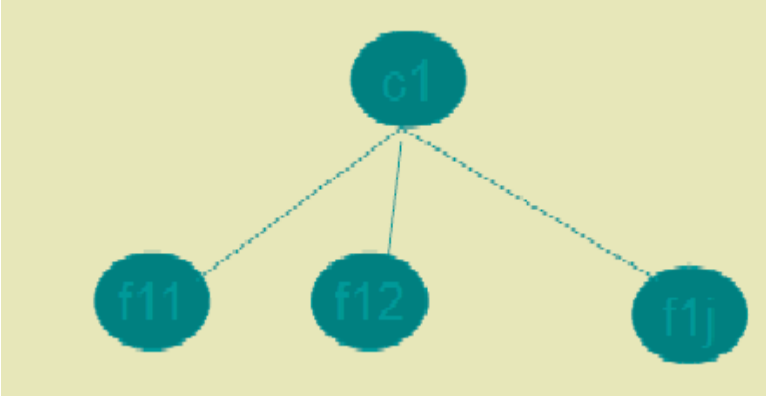
- Case Total document indexing.
- Automatic Indexing requires few seconds based on the processor and complexity of algorithms to generate indexes.'
- Index resulting form automated indexing fall into two classes , weighted and un weighted.
- Weighted indexing system: a attempt is made to place a value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

**Indexing By term**

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

document vector.

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




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

**INFORMATION EXTRACTION**

There are two processes associated with information extraction:

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

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

**DATA STRUCTURES**

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

**Introduction to Data Structures:**

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

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

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

- Two major data structures in anyIRS:

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

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

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

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

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

**Processing Tokens:**

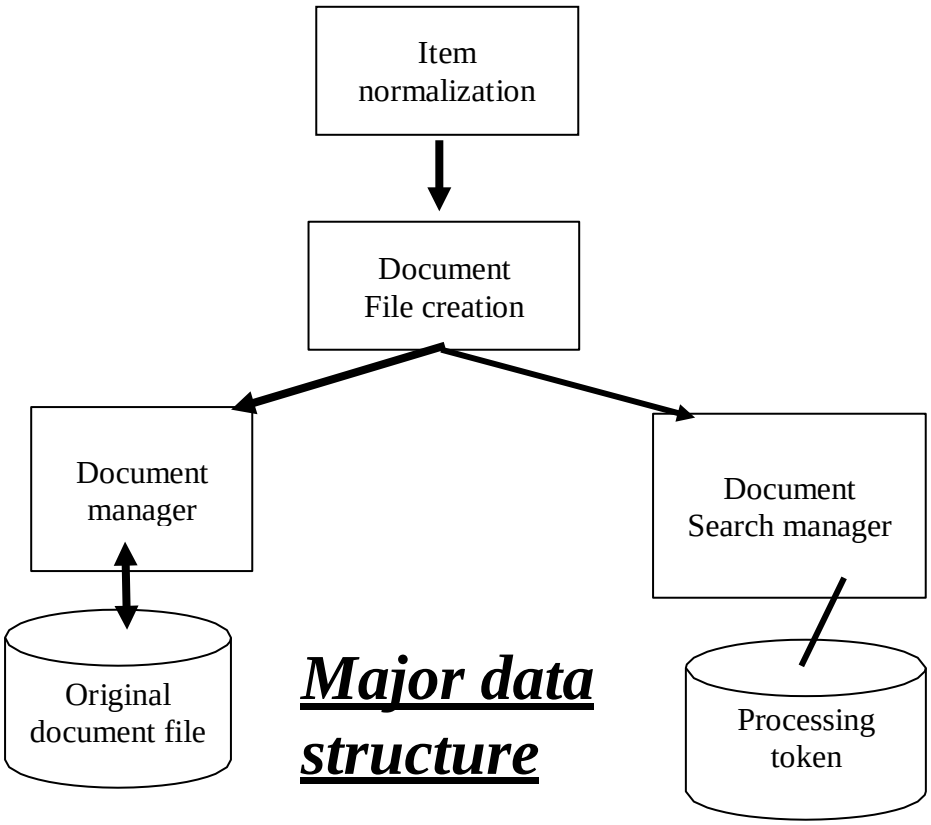
Identify the information that are used in the search process-processing tokens

Divide Input symbols into three classes

**Valid word symbols :** Alphabets, Numbers &special characters

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

**Special Processingsymbols**:-Hyphen(-)




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

Focus : on data structure that support search function

**STEMMING ALGORITHMS:**

Stemming is nothing but cutting &trimming

The concept of stemming is introduced in the 1960's

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

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

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

**Stemming variations:**

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

**Table lookup stemming:**

Uses Large Data structures

Ex: Retrieval ware

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

**Poster stemming Algorithm:**

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

Conditions are

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

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

  C(VC)M V

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

**Dictionary Stemming:**

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

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

**Uses 6 major Data files**

1. Dictionary of words

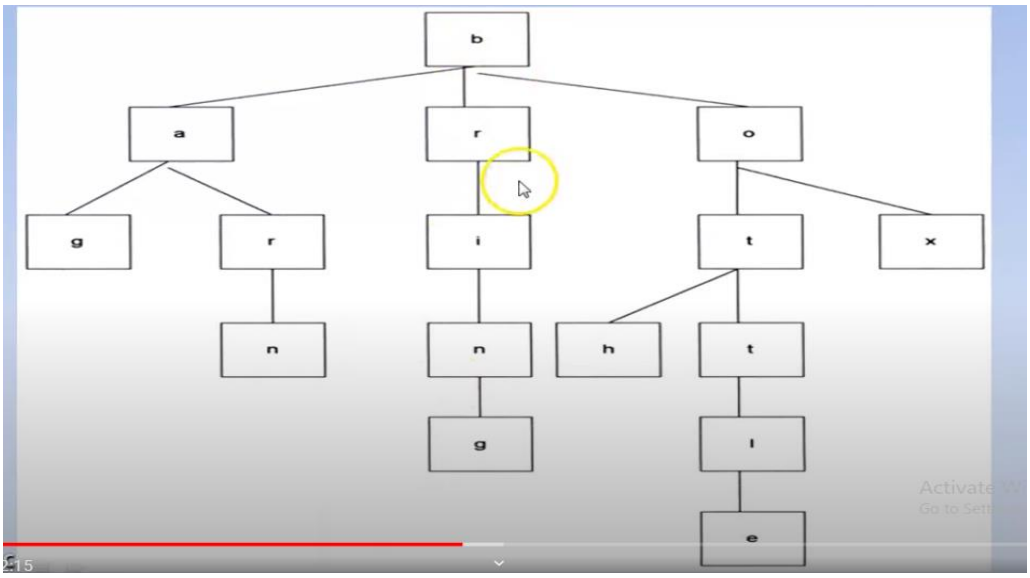
2. Supplement of words

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

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

**Successor Stemming:**

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




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

**Conclusion Stemming:**

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

**Inverted File Structure**

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

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

**Features of Inverted File:**

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

**Increases Precision:**

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

**Zoning used:**

It is logical sub setting Information has to store specific zone

**Ranking also used:**

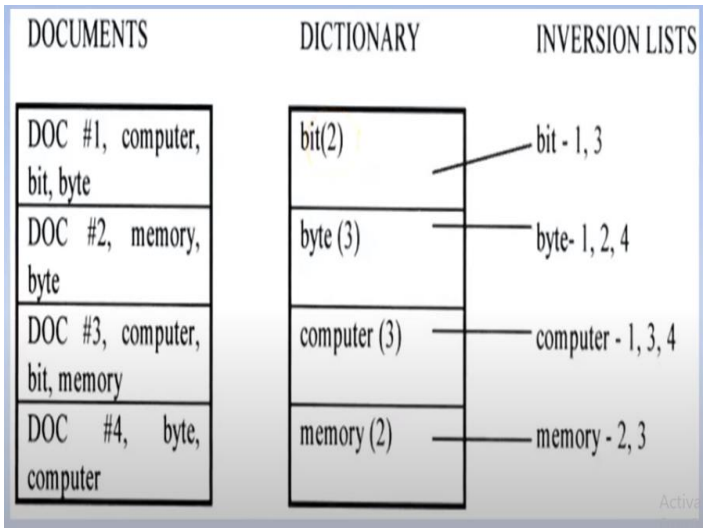
The public will be giving ranking based on particular file, Document, PDF, WORD…..etc

**Store concepts &relationships:**

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

**Natural Language Processing:**

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




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

# N-Gram Data Structures

**N-Gram Data Structures:**

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

**Specializations:**

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

**Special Data structure:**
It is one of the Data structure

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

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

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

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

It was Implemented a formula

MAX Segn =(λ)n

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

**Example:**

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

**Advantages:**

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

**PAT data structure:**

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

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

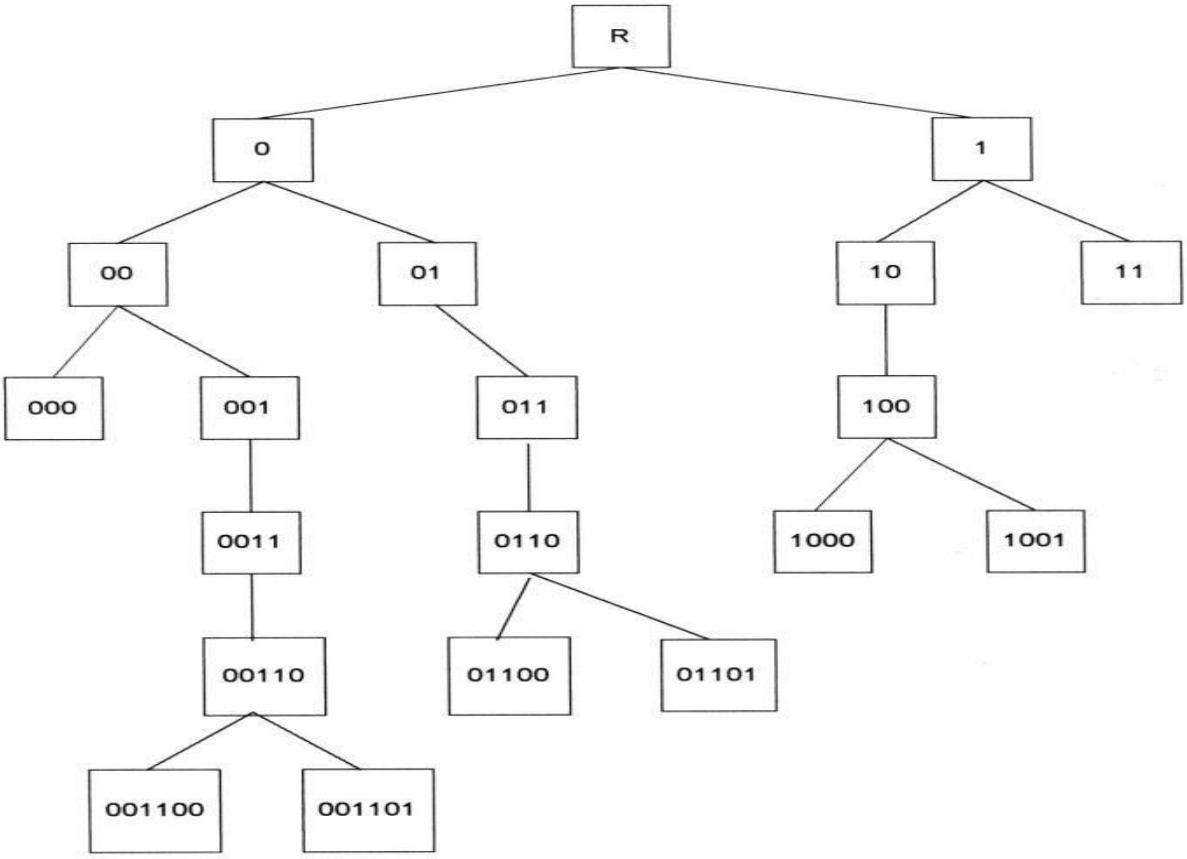
- Onomics for Warsaw is Complex
- Sistring=4
- Omics for Warsaw is Complex
- :
- :
- :
- :
- Sistring=9
- For Warsaw is Complex
- :
- :
- :
- :
- :
- :
- Sistring=25
- Ex

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

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

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

The full PAT binary tree is



**Signature file structure:**

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

**Application(s)/Advantage(s)**

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

**HYPERTEXT AND XML DATA STRUCTURES:**

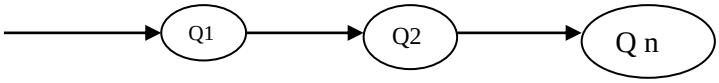
- The Hypertext Data structure is used Extensively in the Internet environment and requires an electronic media storage for the item
- Hypertext allows one item to reference another Item via an Embedded pointer Each separate Item is called a node and the reference pointer is called a link
- Each node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

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

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

**HIDDEN MARKOV MODELS:**

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




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

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

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

**Disadvantages of HMM:**

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

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