III UNIT

AUTOMATIC INDEXING:

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

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

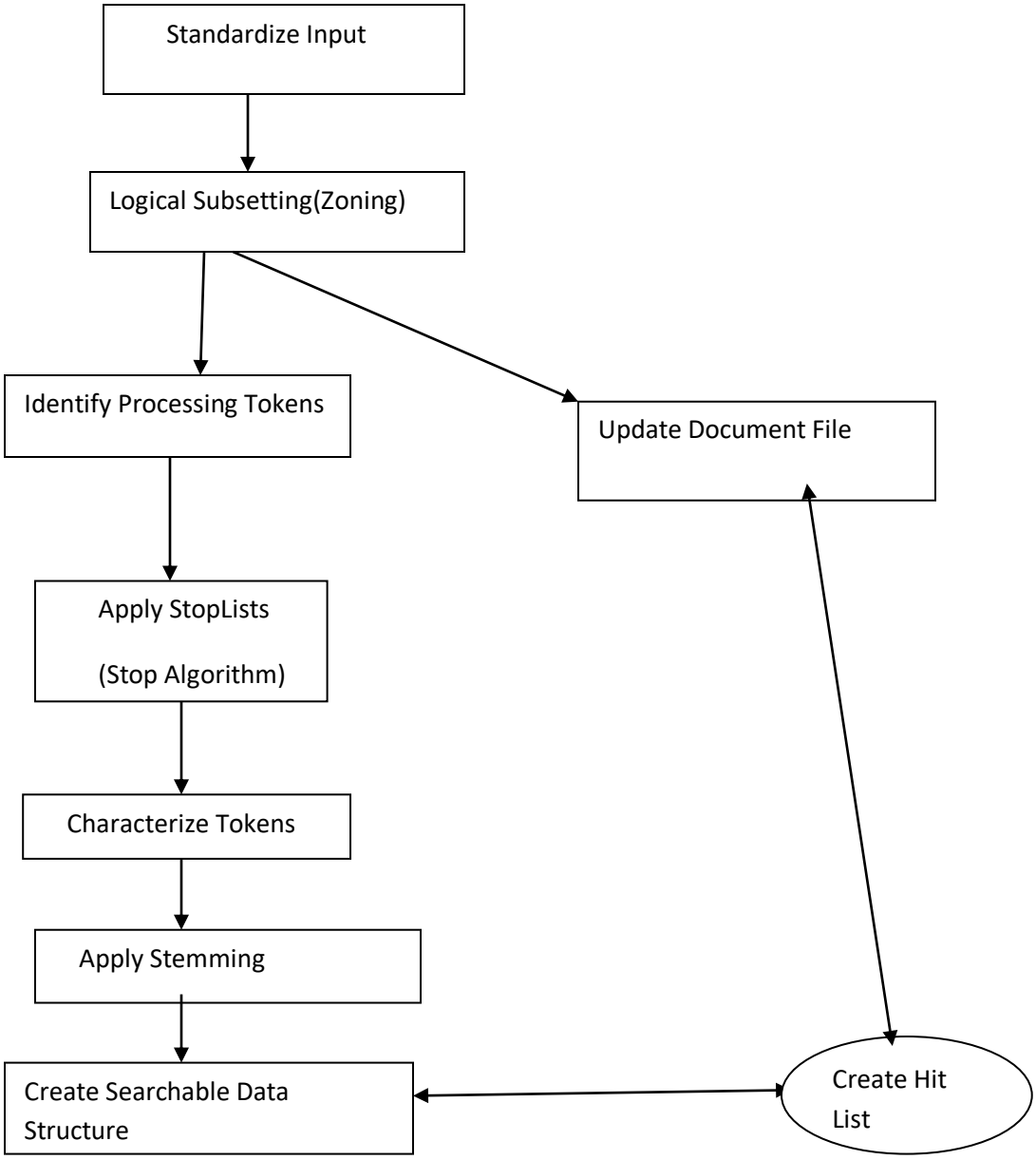
Indexing By term

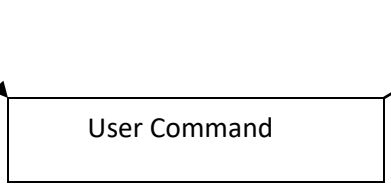
- Terms (vocabulary) of the original item are used as basis of index process.
- There are two major techniques for creation of index statistical and natural language.
- Statistical can be based upon vector models and probabilistic models with a special case being Bayesian model(accounting for uncertainty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

Types of Classes in Automatic Indexing:

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

Data flow in Information Processing System





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

Standardize Input:

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

Logical Sub-setting (Zoning) :

- Parse the item into logical sub-divisions that have meaning to user Title, Author, Abstract, Main Text, Conclusion, References, Country, Keyword
- Visible to the user and used to increase the precision of a search and optimize the display The zoning information is passed to the processing token identification operation to store the information, allowing searches to be restricted to a specific zone

Identify Processing Tokens :

- Identify the information that are used in the search process – Processing Tokens (Better than Words)
- The first step is to determine a word
- Dividing input symbols into three classes
- **Valid word symbols**: alphabetic Characters, Numbers
- **Inter-word symbols:** blanks, periods, semicolons (non-searchable)
- **Special processing symbols:** hyphen (-)

Stop Algorithm:

- Save system resources by eliminating from the set of searchable processing tokens those have little value to the search Whose frequency and/or semantic use make them of no use as searchable token

Characterize Tokens :

- Identify any specific word characteristics Word sense disambiguation Part of speech tagging
- Uppercase – proper names, acronyms, and organization Numbers and dates

Stemming Algorithm :

Stemming is cutting &Trimming of words

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

Create Searchable Data Structure:

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

List of the classes of Automatic Indexing:

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

Statistical Indexing:

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

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

There are two types of frequencies

- Document Frequency
- Term frequency

Document Frequency:

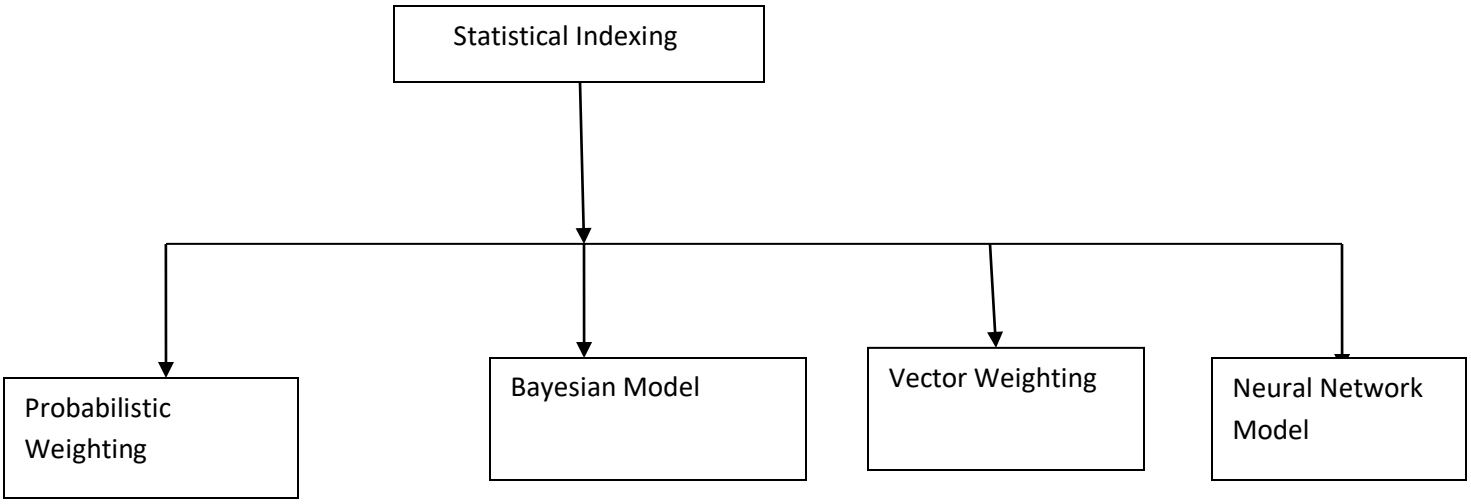
it is identifies as particular documents items is existing or not

Term Frequency:

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

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

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



Probabilistic Weighing:

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

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

HYPOTHESIS:

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

PLAUSIBLE COROLLARY:

The most promising source of techniques for estimating the probabilities of

usefulness for output ranking in IR Standard Probability Theory & Statistics

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

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

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

Bayesian model:

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

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

Vector Weighting:

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

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

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

Neural Network Model:

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

Natural Language:

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

Eg: "Happy or sad Good or Bad"

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

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

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

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

Concept Indexing:

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

Hypertext Linkage Indexing:

The Hypertext is a Data structures are generated Manually

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

Introduction of Clustering:

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

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

Term Clustering:

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

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

Documents Clustering:

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

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

Define the Domain for Clustering

Thesaurus: The Domain may be medical or education

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

Determine the attributes of the objects to be clustered

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

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

Reduce Errors Association

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

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

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

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

Guide lines on the characteristics of the Classes in Clustering

A well-defined semantic definition should exit for each class

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

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

Additional Decisions for Theasurus

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

Word Relationships:-

Equivalence, Hierarchical, Non-Hierarchical

Parts-Wholes:-Aggregation and composition

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

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

Thesaurus Generation:

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

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

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

Thesaurus Generation(Manual Clustering):

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

EX:

KWOC

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

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

KWOC

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

KWIC

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

KWAC

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

Figure 6.1 Example of KWOC, KWIC and KWAC

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

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

Automatic Term Clustering:

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

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

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

Item Clustering:

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

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

Ex: 10 is greater than are equal
Item &Item relationship Matrix

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

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

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

HIERARCHY OF CLUSTERS

Hierarchy Clustering:

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

Hierarchical Agglomerative clustering(HAC):-

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

Hierarchical divisive clustering:-

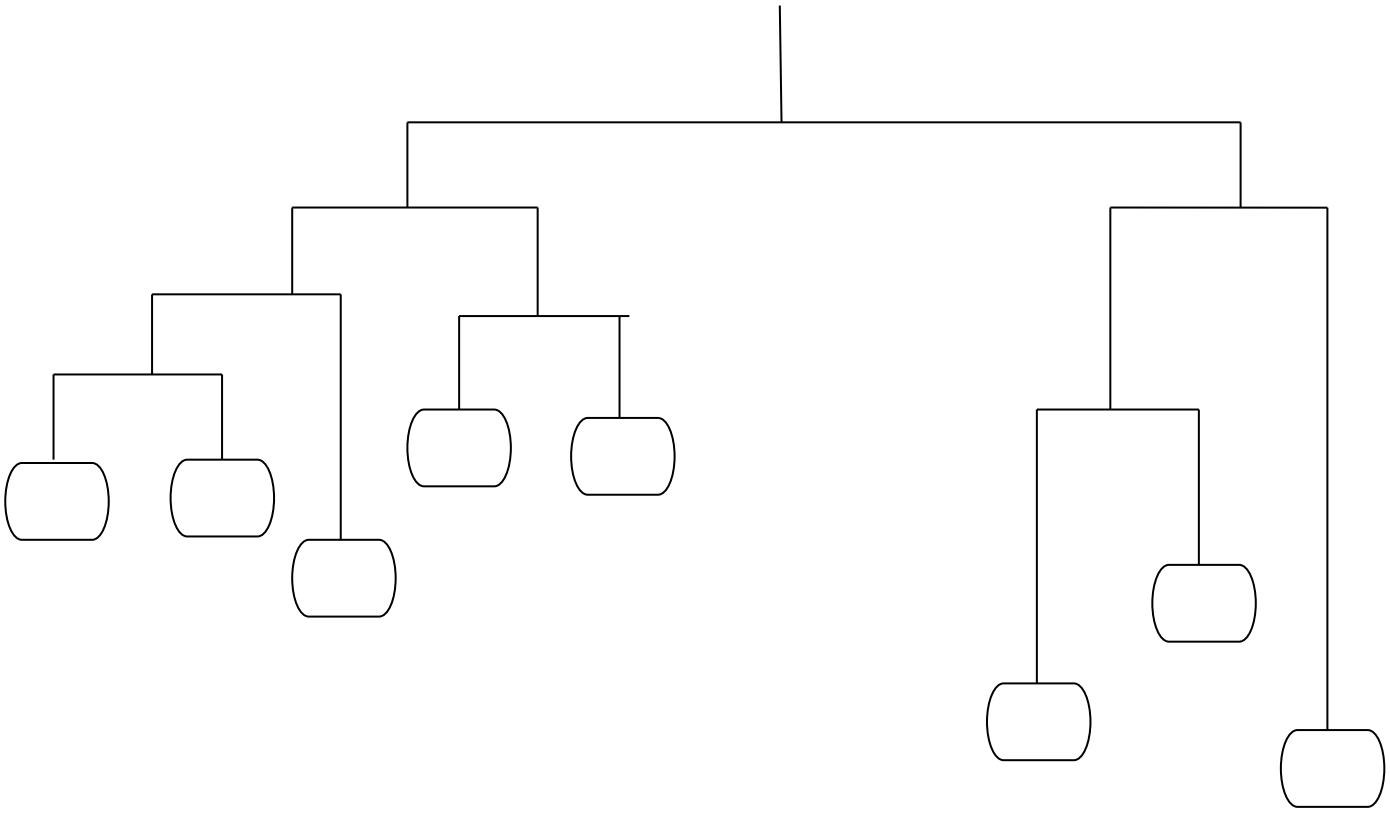
We start with Large cluster and we breaking it down into smaller cluster

Objectives of creating a Hierarchy of Clusters:-

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

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

Dendogram for Visualizing Hierarchical Clusters



This figure about of Structure of Hierarchy clustering