Unit-III

Automatic Indexing

(*) classes of Automatic Indexing

→ Automatic indexing is the process of analyzing an item to extract the information and kept that information permanently in an index.

→ This process is associated with generation of the "Searchable data structure" and Searchable data structure is associated with an item.

→ Automatic Indexing is the preprocessing stage allowing search of items in IRS.

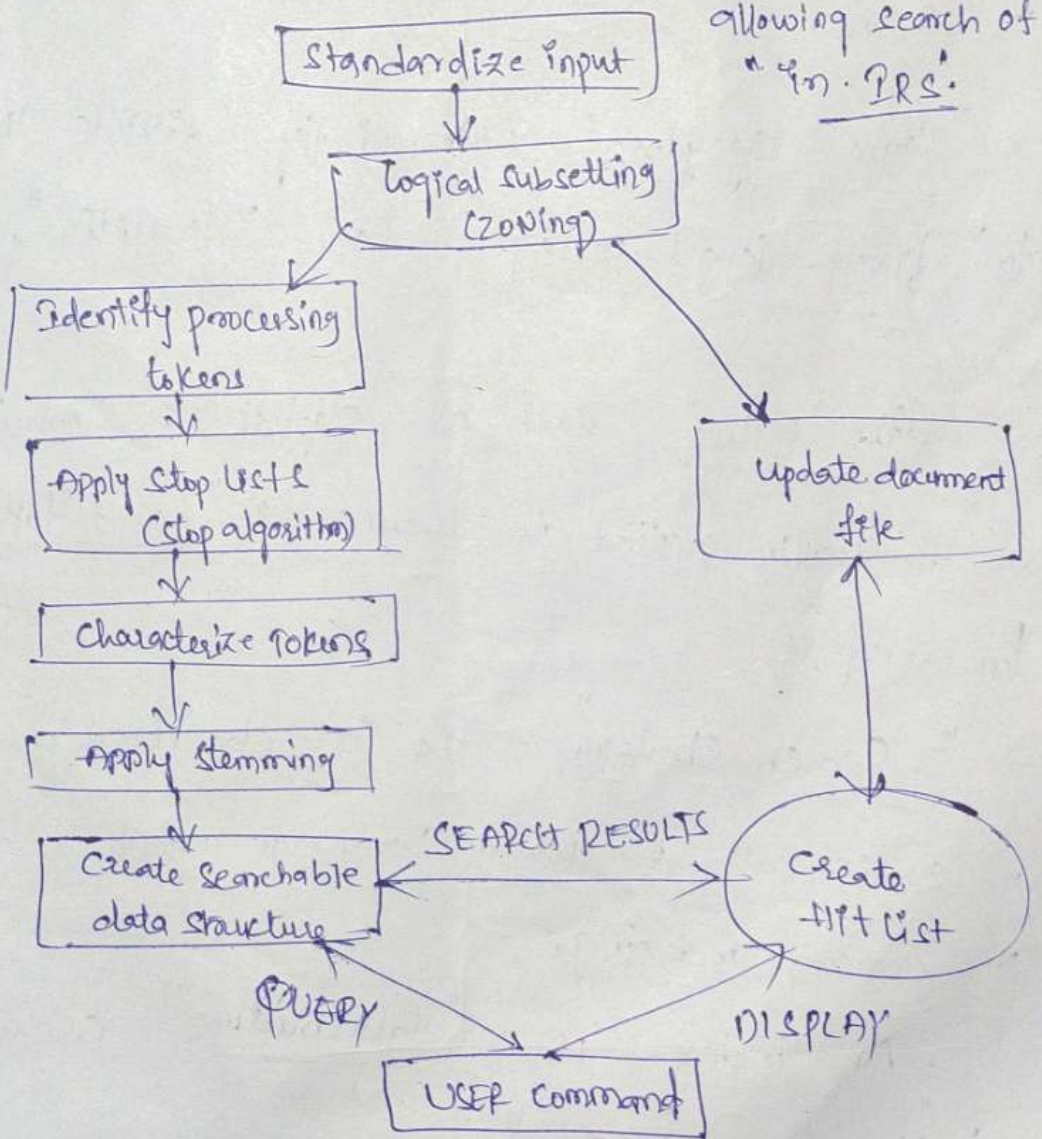


fig: Dataflow in information Processing system.

from figure, the left side including:

→ Identify Processing tokens
→ Apply Stoplists
→ Characterize tokens
→ Apply stemming &
→ create searchable data structures are all the Part of Indexing Process.

→ "All the systems" go through an "initial stage of "zoning" and identifying Processing tokens used to create the index.

→ "Some systems" Automatically divide the document into "fixed-length passages" (or) "localities".

→ "filters" such as stoplist & stemming algorithms are frequently applied to reduce the "no. of tokens" to be Processed.

→ "Search Strategy": It is classified as
* Statistical * Natural language and
* Concept Indexing.

→ An index is a "datastructure" create to support "the search strategy".

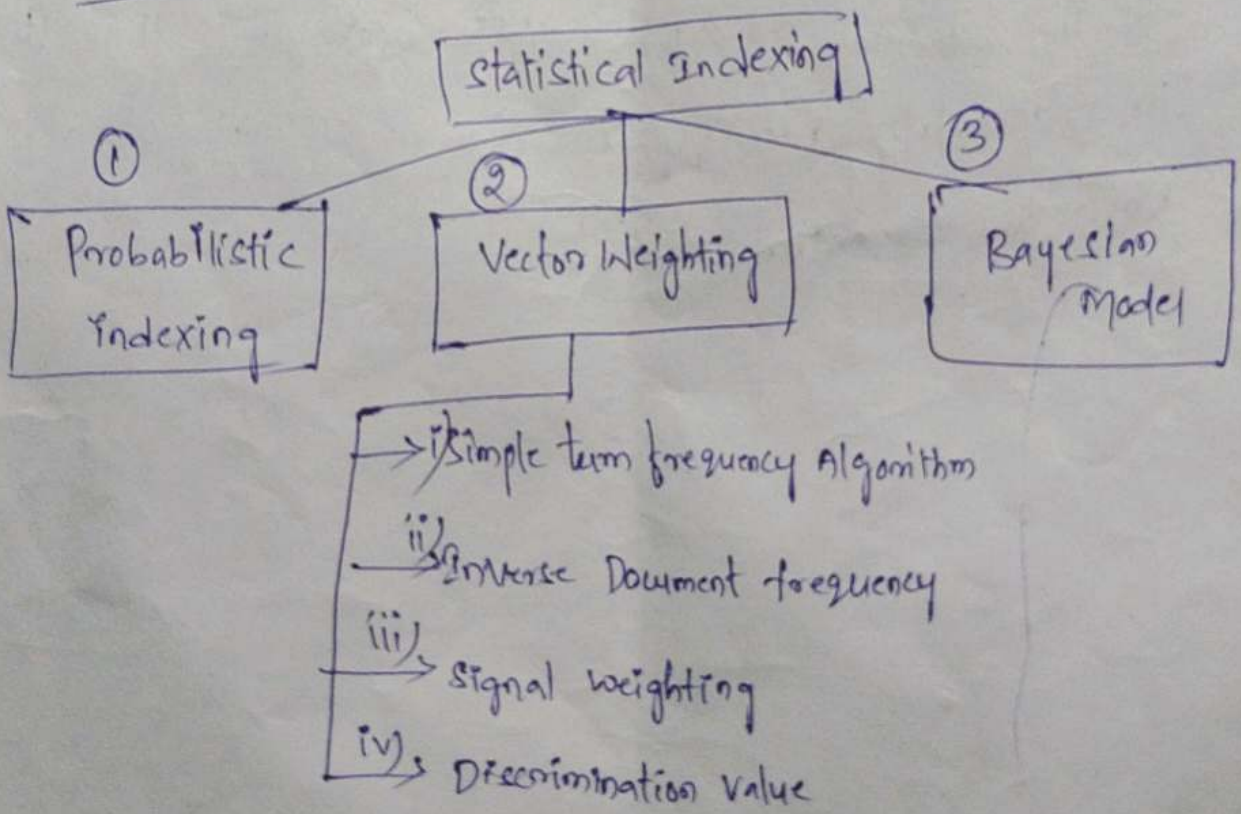
(4) Statistical Indexing :—

Statistical Indexing uses frequency of occurrence of "events" to calculate a "number" that is used to indicate the "Potential Relevance of an item".

Statistical calculations are performed on the +ltt file, ranking the document.

Eg: Term frequency algorithms.

Statistical Indexing divided into 3 types :—



① Probabilistic Indexing :-

→ Probabilistic approach is based on direct application of the theory of Probability to Information Retrieval System.

→ The use of probability theory is a "natural choice" because it is the basis of evidential reasoning (i.e drawing conclusions from evidence).

⇒ In probabilistic weighting concept the probabilities are usually based upon a "binary condition"; i.e an item is "relevant" (or) "not".

Advantage :-

The advantage of probabilistic approach is that it can accurately identify its "weak assumptions" and work to "strengthen" them.

(2) Vector Weighting :—

The semantics of every item are represented by as "Vector".

- A vector is a "One-dimensional" set of values, where the Order / Position of each value in the set is fixed and represented as a particular domain.

Vector typically represented by "Processing token".

There are two approches in vector to represent domain values:

i) Binary approche
ii) Weighted approch.

i) Binary approch :-

Under Binary approch,

The domain Contains the values of "0" (or) "1", with one repuesenting the "Processing token" in the item.

ii) Weighted approche:

Under ~~un~~weighted approch,

The domain is typically the set of all "real positive numbers"

for Example :-

The major topics discussed are indicated by the index term for each column

i.e Petroleum, mexico, Oil, Taxes, Refineries, & shipping.

Assume threshold is 1.0.

	Petroleum	mexico	Oil	Taxes	Refineries	shipping
Binary	(1	1	1	0	1	0)
Weighted	(2.8	1.6	3.5	0.3	3.1	0.1)

fig: Binary and Vector Representation of an item.

→ In the above example "tax" and "shipment" are below the threshold of importance so that they are not included in the vector.

→ After eliminating the below threshold items the remaing processing tokens/items are three they are Petroleum, mexio, Oil.

→ Vector approch allows for a mathematical and physical representation using a vector-space model.

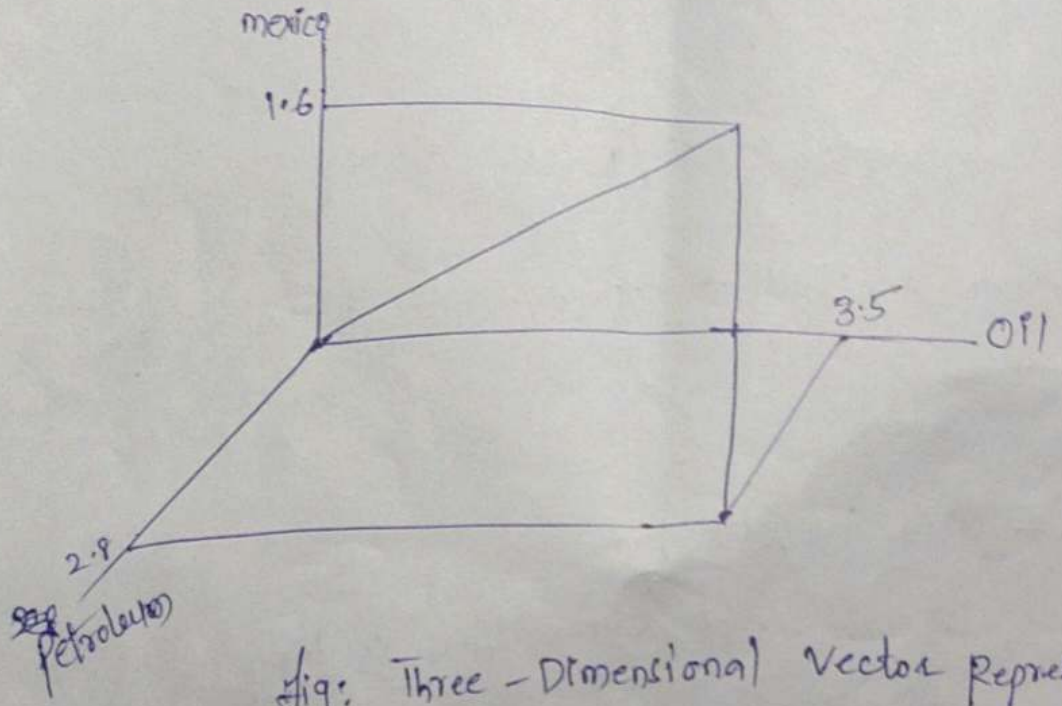


Fig: Three-Dimensional Vector Representation

i)* Simple Term frequency Algorithm:

In statistical system,

The data that are potentially available for calculating a weight are: (i) the frequency of occurence of the Processing token in an existing item (i.e term frequency (TF)), (ii) the frequency of occurence of the processing token in the existing database (i.e total frequency - TOTF) & and (iii) no. of unique items in the ~~existing~~ database that contains Processing token (i.e item frequency - If, is frequently ~~labe~~ labeled in other publications as Document frequency or).

Term frequency weighting formula:

$$\frac{(1+\log(TF))/1+\log(\text{average}(TF))}{(1-\text{slope})*\text{pivot}+\text{slope}*\text{number of unique terms.}}$$

Here

TF → Term frequency

Pivot → average no. of unique terms occuring in the collection.

ii) Inverse Document frequency :—

The frequency of occurence of the processing token in the "data base" is called Document frequency.

In the weighting algorithms that the weight assigned to an item should be inversely proportional to the frequency of occurence of the an item in the data base.

This algorithm is called "Inverse document frequency" "IDF".

formula :— Un-normalized weighted formula.

$$WEIGHT_{ij} = TF_{ij} * \left[\log_2(n) - \log_2(IF_j) + 1\right]$$

Where TF → Term frequency.

IF → item frequency

for Example

Assume the term "oil" is found in 128 items,
"mexico" is found in 16 items,
refinery is found in 1024 items.

If a new item arrives with all three terms in it,
"oil" found "4 times"
"mexico" found "8 times"
"refinery" found "10 times" and

there are 2048 items in the total database.

⇒ Using un-normalized term frequency, the item vector is (4, 8, 10) i.e TF.

Now using inverse document frequency the following calculations apply;

(how many times Dil occurs; total items in db; oil found in 128 patterns)

$$\text{Weight}_{oil} = 4 * (\log_2(2048) - \log_2(128) + 1)$$
$$= 4 * (11 - 7 + 1)$$
$$= 20$$

$$\text{Weight}_{mexio} = 8 * (\log_2(2048) - \log_2(16) + 1)$$
$$= 8 * (11 - 4 + 1)$$
$$= 64$$

$$\text{Weight}_{refinery} = 10 * (\log_2(2048) - \log(1024) + 1)$$
$$= 10 * (11 - 10 + 1)$$
$$= 20$$

∴ with resultant IDF item vector = (20, 64, 20)

fig: Example of IDF

iii) Signal weighting

The distribution of the frequency of processing tokens with in an item can affect the ability to rank items.

for example

Assume the terms "SAW" and "DRILL" are found in 5 items with following frequencies.

The theoretical basis for the algorithm:

Item Distribution	SAW	DRILL
A	10	2
B	10	2
C	10	18
D	10	10
E	10	18
	50	50

Item distribution for SAW and DRILL.

Both items are found 50 times in the 5 items.

→ The information value for an event that occurs ~~50 % of the time~~ is calculated by

$$\text{INFORMATION} = -\log_2(P)$$

where P → probability of occurence of event "p".

∴ The Information Value for an event that occurs 50 % of the time is :

$$\text{INFORMATION} = -\text{Log}(.50)$$
$$= -(-1)$$
$$= 1$$

Average Information :

A.I Values across the event is calculated by

$$\text{AVE\_INFO} = -\sum_{K=1}^{n} P_K \text{Log}(P_K)$$

$P_K \rightarrow P_K$ can be defined as $TF_{IK}/TOTF_K$.

AVE-INFO → takes maximum values for every P_K is same.

formula for Calculating Signal weight :

$$\text{Signal}_K = \text{Log}_2(\text{TOTF}) - \text{AVE-INO}$$

$\text{Signal}_K \Rightarrow \text{Signal}_{SAW}$, Signal_{DRILL}, TOTF = 50

$$\text{Signal}_{SAW} = \text{Log}_2(50) - [5 \times \{ 10/50 \text{ Log}_2(10/50) \}]$$

Log₂(50) ↓ TOTF; 10/50 ↓ P_K; $\text{Log}_2(10/50)$ ↓ $\text{Log}_2(P_K)$

$\sum_{K=1}^{n}$ i.e 1 to 5

K=1, n=5

$$Signal_{DRILL} = Log_2(50) - [2/50\ Log_2(2/50) + 2/50\ Log_2(2/50) + 18/50\ Log_2(18/50) + 10/50\ Log_2(10/50) + 18/50\ Log_2(18/50)]$$

∴ The weighting factor for term "DRILL" does not have a uniform distribution and it is larger than that term "SAW" and it gives "higher weight."

iv) Discrimination value :—

Another approach to creating a weighting algorithm is to "based it upon the "discrimination value of item".

It is used to achieve the objective of finding relevant items and it is important that the index discriminates among items.

Formula :

$$DISCRIM_i = AVESIM_i - AVESIM$$

- $AVESIM$ → Average similary b/w every item in the db.
- $DISCRIM_i$ → discrimination value for each item "i".
- $AVESIM_i$ → Same Average Calculation similary i.e Avg. similary b/w every item except that item "i" is removed from all items.

(3) Bayesian model :—

The Bayesian approach is based on "conditional probabilities".

ie Eg: Probability of "Event 1" given "Event 2" occurred.

In General Case :

The Bayesian formula is

$$P(REL/DOC_i, Query_j)$$

where P → Probability

REL → Probability of Relevance (REL)

REL :

The Probability of Relevance (REL) to search statement given a particular "Document" and "Query".

⟹ The Bayesian formulas can be used in determining the "weights associated" with Processing tokens in an item.

Bayesian Network :—

A Bayesian network can be used to determine the final set of Processing tokens and their weights.

Here "processing tokens called as "topics".

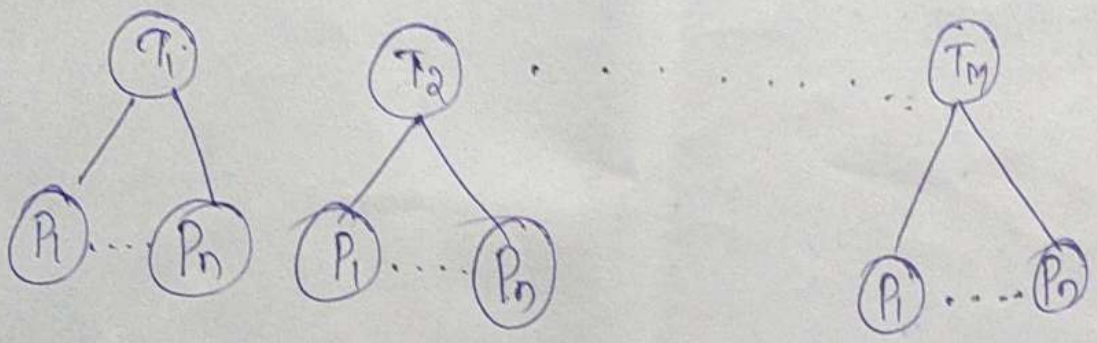


fig: Bayesian Term Weighting.

From, figure show,

A simple view of process where

"T_i" → represents the relevance of topic "i" in a particular item. and i.e ($i = 1 \dots m$)

"P_j" → represents a statistics associated with the event of processing token "j" being present in item. i.e ($j = 1 \dots n$)

(4) ~~Bayesian Networks:-~~ Extended Bayesian Networks:-

Problem Handling in Bayesian model:-

⇒ There are 2 approaches to handling the problems.

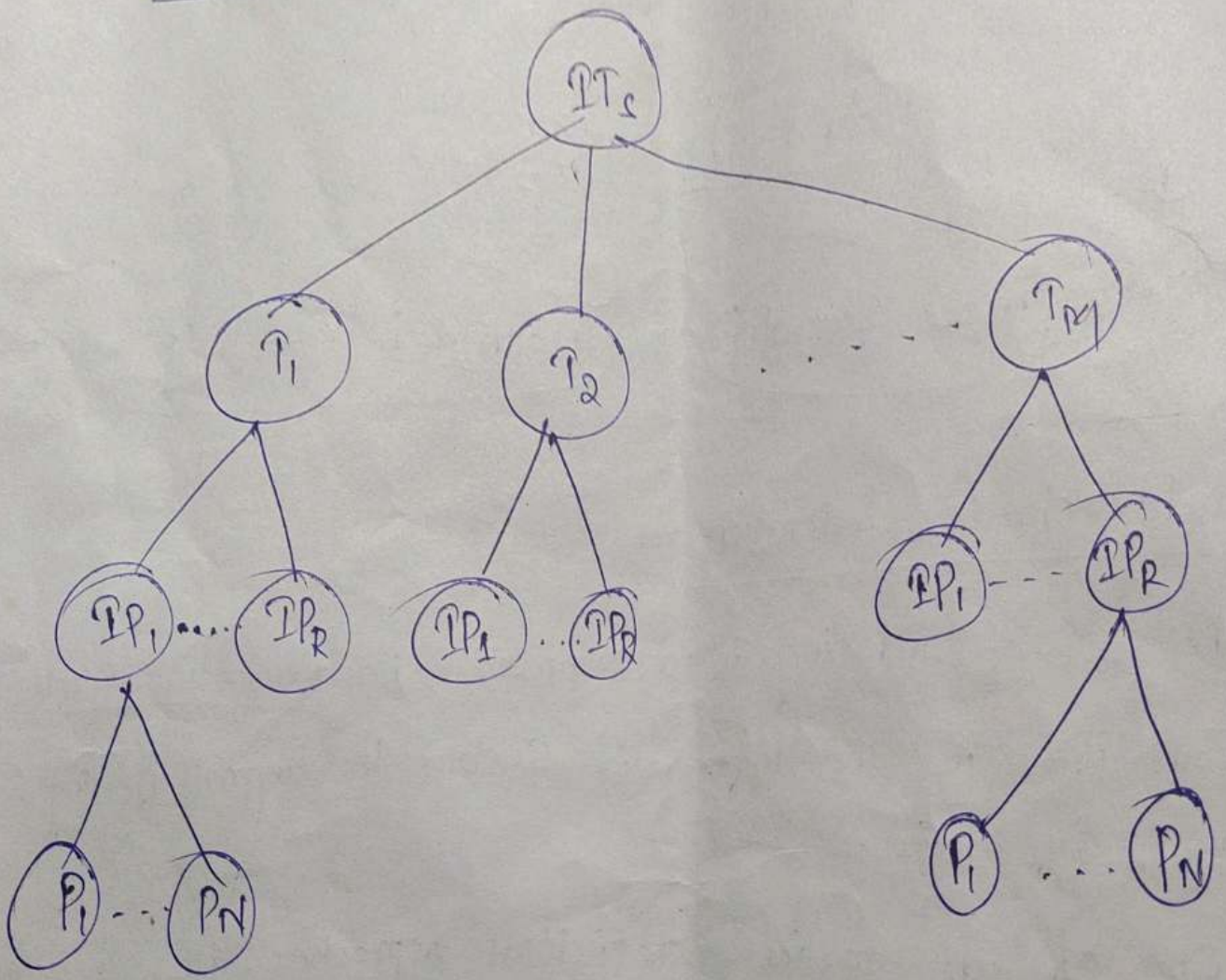
First Approach:-

first is to assume that there are "dependencies". these errors are created by assuming manual independence.

Second Approach:-

Second Approach extend the network to additional layers to handle "inter dependencies".

Extended Bayesian Network :-



Where

IT_S → Independent topics

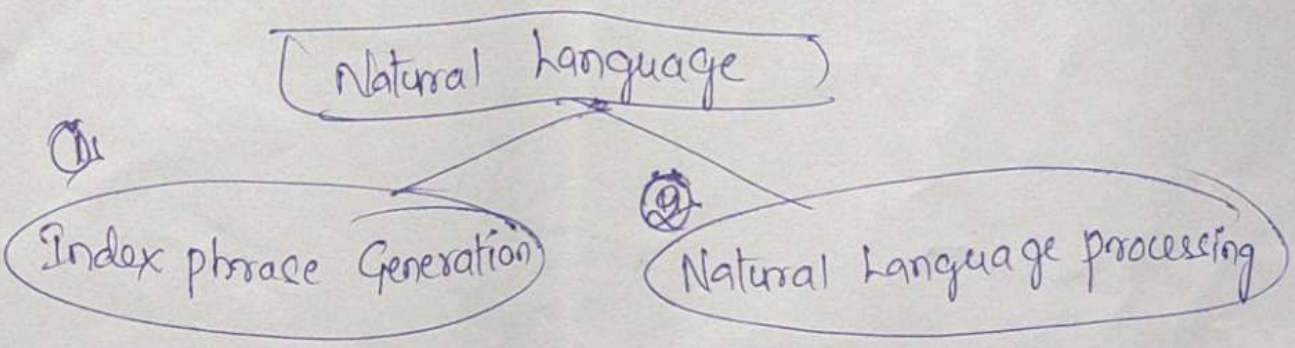
IP → Independent processing Tokens.

T → Relevance topic

P → Processing token

(*) Natural Language :-

Natural Language Concept divided into 2 types



(1) Index phrase Generation :-

The Goal of Indexing is to represent the "Semantic Concepts" of an Item in the Information System to support finding relevant information.

One of the earliest statistical approches to determining term phrases proposed by "Salton" was. Use a "COHESION" factor between terms:

$$COHESION_{k,h} = SIZE\text{-}FACTOR * \left(\frac{PAIR\text{-}FREQ_{k,h}}{TOTF_k * TOTF_h} \right)$$

Where, "SIZE-FACTOR" is a normalization factor & it is based on the size of the "Vocabulary"

"$PAIR\text{-}FREQ_{k,h}$" → Total frequency of occurence of pair terms $Term_k$, $Term_h$ in the Item collection.

TOTF → Total frequency (i.e Total frequency of k * " " h).

If a "phrase" such as "Industrious Intelligent students" was used often in both approaches. i.e

① Statistical Approach ② Natural language approach.

Statistical Approach :-

① ②

Word phrase : "Industrious Intelligent Students"

Statistical Approach create the phrases such as "Industrious Intelligent" and "Intelligent Student".

Natural Language Approach :-

Word phrase : "Industrious Intelligent Students"

Natural Language approach would create phrases such as "Industrious student", "Intelligent Student" and "Industrious Intelligent student".

Guide lines to generate Index phrases:

i) Any pair of adjacent non-stop words, is a potential phrase.

ii) Any pair must exist in 25 or more items.

iii) Phrase weighting uses a modified version of the SMART system single term algorithm.

iv) Normalization is achieved by dividing the length of single-term subvectors.

(2) Natural Language Processing:-

Natural language processing can combine the concepts into higher-level concepts sometimes referred as thematic representations.

The first step in a natural language determination of phrases is a "lexical analysis of the input".

The Lexical Analysis determine the

* Verb tense,
* plurality
* Part of speech.

→ Natural language processing not only producing / produces "more accurate" term phrases but also provide higher level semantic information that identifying the relationships b/w concepts.

→ Part of speech tags:

class	examples
(1) determiners	a, the
(2) Singular nouns	Paper, notation, structure, language
(3) Plural nouns	operations, data processes
(4) Preposition	in, by, of, for

class.	Example
(5) adjective	high, concurrent.
(6) Present tense verb	Presents, associates
(7) Present participal	multiprogramming.

Fig: part of Speech tage.

(*) concept Indexing :-

→ The concept Indexing is a "statistical technique".

Goal:

The Goal of concept Indexing is "determine the Canonical representation" of "concepts".

Uses:-

The concept Indexing is "used" to representing the "relevant information" when other techniques are "miss".

Start Point

The concept Indexing "start with" the "no. of unbalanced" concept classes and these are created to define the information.

In Concept indexing "single term" is representing the "multiple concepts".

Example :-

TERM : Automobile

TERM	Weights for associated concepts:
Vehicle	.65
Transportation	.60
Environment	.35
Fuel	.33
Mechanical Device	.15

Vector representation Automobile: (.65, ..60, ..., .35 ..., .33, ..., 15)

fig:- Concept Vector for Automobile.

From fig:-

→ The term "Automobile" is related to the concepts such as "Vehicle", "Transportation", "Environment", "fuel" and "Mechanical Device".

→ Here, Automobile is strongly related to the "vehicle" and "lesser" related to the "transportation" and "much lesser" related to other terms.

→ Vector representation of Automobile: (.65 ..., .60 ..., .35 ... 30 ... 15).

Streaming Op.

(*) Hypertext Linkages :-

A new class of information representation is the "Hypertext data structure" evolving on "Internet".

Hypertext (linkages created on) data structure must be generated manually through User Interface tools.

Hypertext linkages :

Hypertext linkages are creating an additional information retrieval dimension (or) an "additional dimension" to the information retrieval system.

In this concept, traditional items can be viewed as 2-Dimensional construct.

One - Dimension :

The text of the item is One - Dimension representation of information.

Two - Dimension :

In Two - Dimension, Embedded references are logical Second 2-Dimensions.

⇒ The Embedding of the linkage allows the user to go immediately to the linked terms for additional Information.

Document and Team Clustering

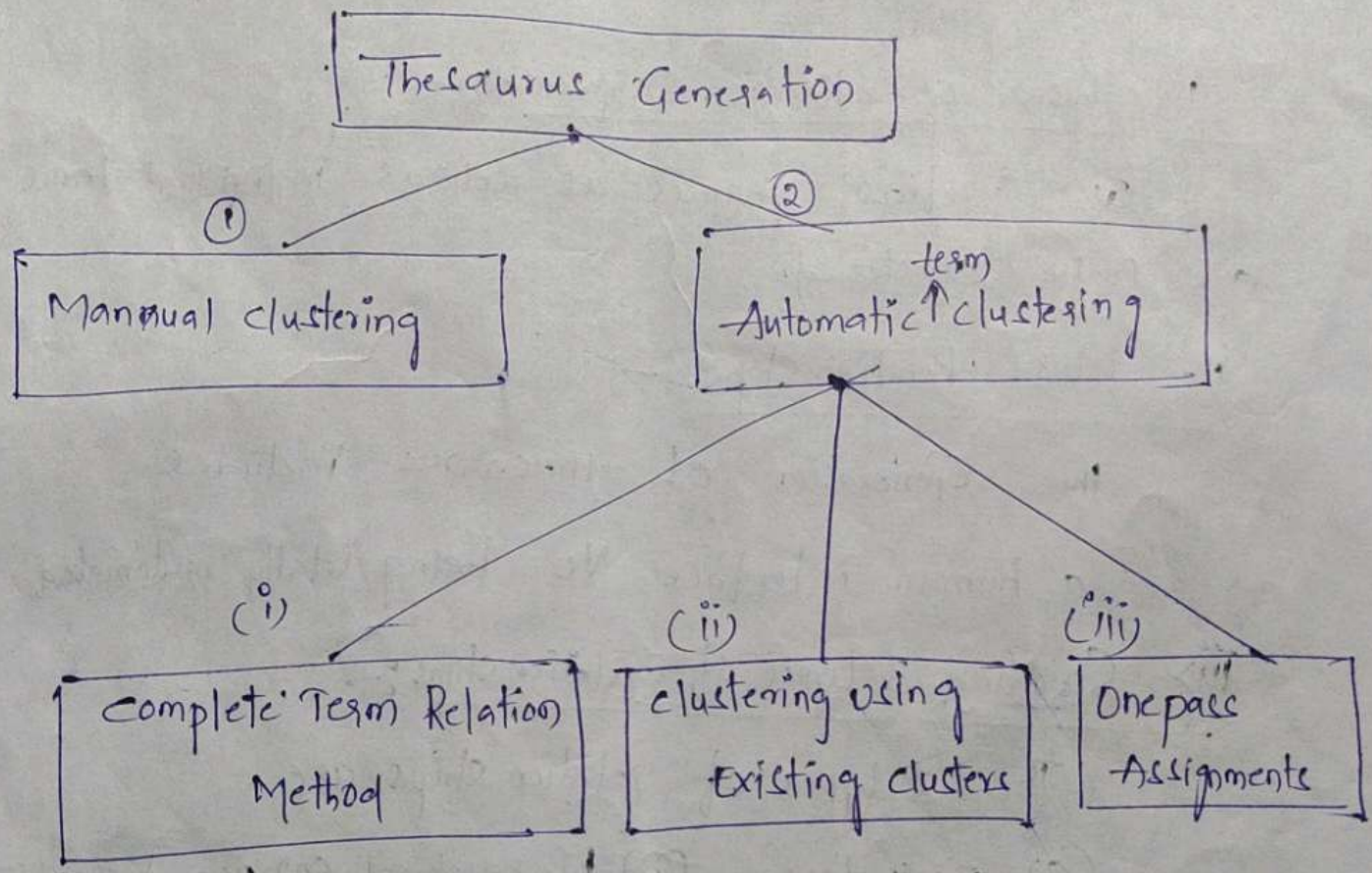
(*) Introduction to clustering :-

→ Clustering (Document) is the process of grouping a set of documents into clusters of similar documents.

→ Documents within a cluster should be similar.

→ In clustering documents are clustered into groups either automatically or manually.

(X) Thesaurus Generation :-



Thesaurus Generation :

The word "Thesaurus" coming from the Latin word meaning "treasure" → store house.

Treasure is similar to dictionary in that it "store words"

Instead of definition it provides "synonyms and Antonyms for the words.

A Thesaurus is a "Book of words" that shows relationships among the words.

* The important decisions associated with the generation of thesauri are not part of item clustering.

i) Word coordination approach:-

It specifies phrases as well as individual terms are to be clustered.

ii) Word Relationships :-

The Generation of thesaurus includes a human Interface Vs being totally automated.

iii) Specified 3 types of relationships:-

The 3 types of relationships are.

(a) Equivalence (b) hierarchical (c) non-hierarchical.

Equivalence relationships :-

Equivalence relationships are "most common" and represent "Synonyms".

Hierarchical relationships:-

Hierarchical relationships where the "class name" is a "general term" and entries are specific examples of general term.

Non-Hierarchical relationships:-

Non-Hierarchical relationships covers other types of relationships such as "Object" - "attribute".

iv) Homograph resolution :-

Homograph is a word that has "Multiple, Completely different meanings".

For Example:

Term : field

The term field means a Electrical field & a field of grass etc.

v) Vocabulary constraints:-

It includes guidelines on the Normalization and Specificity of the Vocabulary.

* Methods for Generation of thesaurus:-

There are 3 basic methods for generation of a thesaurus.

(1) Hand crafted method :-

↙ Manual thesauri

↘ General thesauri

* The Manual thesauri only helps in query expansion.
* The General thesauses does not help because "same word" have different meaning.

(2) CO-Occurence :-

It provide Techniques for creation of thesauri.

(3) Header - Modifier based :-

It is based on the Linguistic relationships.

↳ Scientific study of languages & it's structures.

Types of Thesaurus Generation :-

There are 2 types of thesaurus generations.

1. Manual clustering
2. Automatic Term clustering.

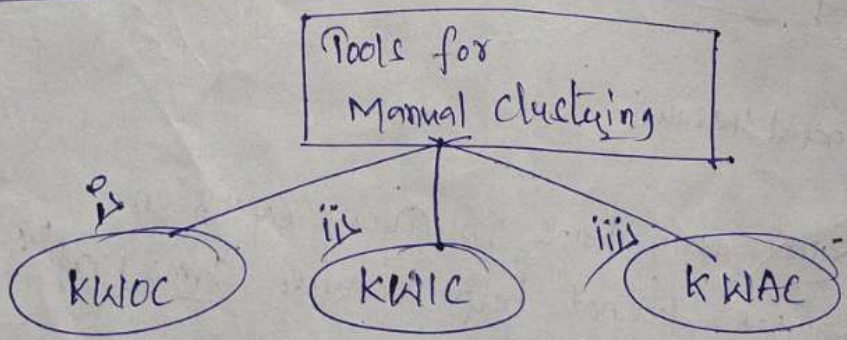
(1) Manual clustering :-

Manual clustering Process follows the steps in the generation of thesaurus.

steps:

→ Determine the domain for clustering.
→ Determine the strength of the relationships b/w attributes.
→ Total set of objects and the strength of the relationships b/w the Objects have been determined.
→ Apply some algorithms to determine the class.

Tools (or) Techniques Used for Manual clustering :-



i) KWIOC :-

KWIOC → keyWord Out of Context

KWIOC is the another name for "concordance".

Concordance :-

→ A concordance is an alphabetical listing of words from set of items.

ii) KWIC :-

KWIC → Key Word In Context

It is structured to identify location of the term.

iii) KWAC :-

KWAC → KeyWord And Context

It is display the keywords followed by their context.

The Following phrase representing the KWOC, KWIC, KWAC:

Note :- The phrase is assumed from the document.

for Example :- "Computer design contains memory chips"

KWOC :-

TERM	FREQ	ITEM ID's
chips	2	doc2, doc4
Computer	3	doc1, doc4, doc10
Design	1	doc4
memory	4	doc3, doc4, doc8, doc12

⇒ first Arrange the Words in "sorted order".

⇒ After assanging the terms check frequency. Here frequency is indicating how many no. of times the particular "term" appear in documents.

i.e chips → "2" times appear in documents (i.e doc2, doc4)

Similarly all.

KWIC :-

In the given sentence / phrase

{ "Computer design contains memory chips" }

→ The "chips" is last word in the phrase so that, it is indicated with "/" "slash".

→ Here grouping is performed as follows:

* If User want to search for a word "chips/"

The complete sentence will be extracted like this

"Computer design contains memory chips/" and the chips will be "extracted" from the sentence.

KWIC extracted word Computer Design contains memory chips/.

* chips/ → Computer Design contains Memory

extracted Computer Design contains memory chips/.

* Computer → Design contains memory chips/.

* Design → Contains memory chips/ Computer.

* Memory chips/ Computer Design Contains.

KWAC :-

The term which you want to search chips (or) computer (or) Design (or) Memory, the complete sentence will be extracted.

i.e

chips	Computer Design contains Memory chips
computer	Computer Design contains Memory chips.
design	Computer Design contains Memory chips.
Memory	Computer Design contains Memory chips.

Fig: Example for KWOC, KWIC & KWAC.

(2) Automatic Term clustering :-

The Terms are automatically clustered based on their frequency.

* The Automatic Term clustering can be performed by following 3 ways.

1. complete term Relation Method
2. clustering using existing clusters
3. One pass Assignments

(1) Complete term Relation Method :-

In this Method,

The similarity between every term pair is calculated as a basis for determining the clusters.

The easiest way to understand this model / approach is "Vector Model".

Vector Model :-

The Vector Model is represented by a "Matrix" where rows indicate the "individual items" and columns are the "Unique words" (processing tokens) in the items.

The values in the Matrix represent the how strongly word represents concepts in the item.

The ~~similarity~~ similarity b/w the two terms are calculated by :

$$SIM(Term_i, Term_j) = \Sigma (Term_{K,i})(Term_{K,j})$$

Where, SIM ⟶ Similarity b/w 2 terms

K ⟶ K is summed across the set of all items.

Threshold value as > 10

For Example :-

	Team1	Team2	Team3	Team4	Team5	Team6	Team7	Team8
Item1	0	4	0	0	0	2	1	3
Item 2	3	1	4	3	1	2	0	1
Item 3	3	0	0	0	3	0	3	0
Item4	0	1	0	3	0	0	2	0
Item 5	2	2	2	3	1	4	0	2

fig: Vector Example.

Sol :→ Based on the Vector Matrix we have to find out

(i)* Team - Team Matrix

(ii)* Team Relationship Matrix

(iii) Network Diagram of Team Similarities

→ from the figure, the "Team 1" appes in "Item 1" is "0"

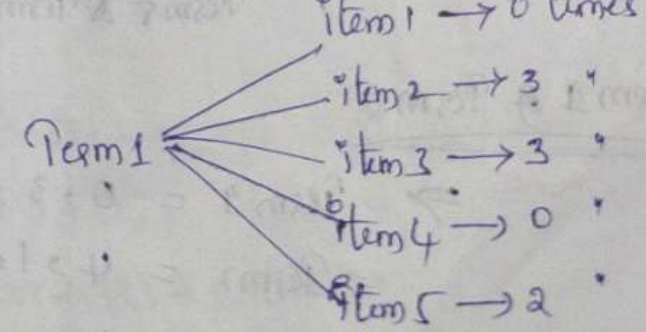
Item 2 is "3",

Item 3 is "3"

Item 4 is "0"

Item 5 is "2"

i.e,



Similarly, Team 2, Team 3 --- Team 8.

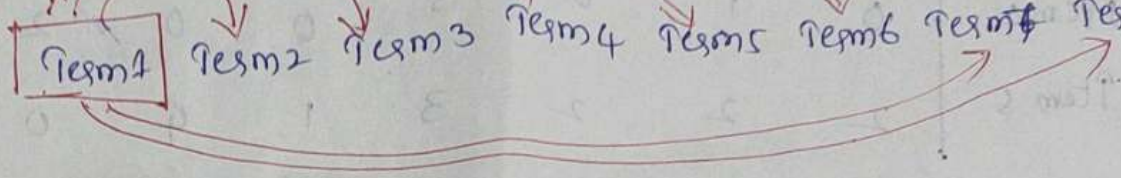
↑ find

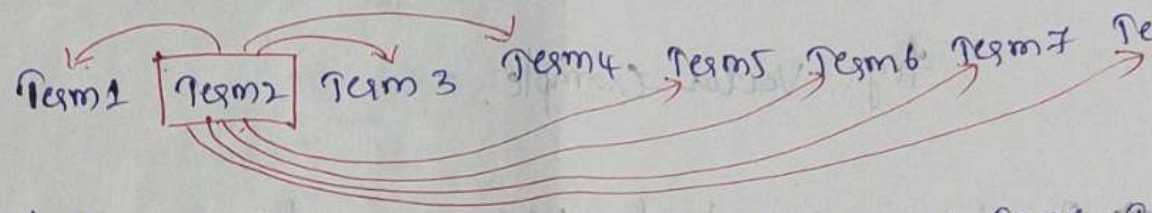
1) Term to Term table / matrix (M-matrix) :-

Find the relationship b/w one term and other terms means how Term1 is related to Term1, Term2, Term3 - Term8.

Similarly, Term2 is related to Term1, Term2, Term3 --- Term8.

i.e





Similarly all terms : i.e Term3, Term4, Term5, Term6, Term7, Term8.

for Term 1 :-

Relationship b/w
- Term1 & Term1
- Term2 & Term2
- Term3 & Term3
- Term4 & Term4
- Term5 & Term5
- Term6 & Term6
- Term7 & Term7
- Term8 & Term8

} Diagonal Elements i.e These are "no values" on diagonal matrix

Term 1 & Term 2

⇒ Term1 = 0, 3, 3, 0, 2

Term2 = 4, 1, 0, 1, 2

formula

$$SIM(Term_i, Term_j) = \sum (Term_{k,i})(Term_{k,j})$$

Term1 , Term2 = 0×4 + 3×1 + 3×0 + 0×1 + 2×2

= 0 + 3 + 0 + 0 + 4

∴ Term1 , Term2 = 7

Term1 , Term3 :

Term1 = 0 , 3 , 3 , 0 , 2

Term3 = 0 , 4 , 0 , 0 , 2

(Term1 , Term3) = 0×0 + 3×4 + 3×0 + 0×0 + 2×2

= 12 + 0 + 4

∴ Term1 , Term3 = 16

Term1 , Term4 :

Term1 = 0 , 3 , 3 , 0 , 2

Term4 = 0 , 3 , 0 , 3 , 3

(Term1 , Term4) = 0×0 + 3×3 + 3×0 + 0×3 + 2×3

= 9 + 6

∴ Term1 , Term4 = 15

Term1 , Term5 :

Term1 = 0 , 3 , 3 , 0 , 2

Term5 = 0 , 1 , 3 , 0 , 1

(Term1 , Term5) = 0×0 + 3×1 + 3×3 + 0×0 + 2×1

= 3 + 9 + 2

∴ Term1 , Term5 = 14

Term 1, Term 6 :

Term 1 = 0, 3, 3, 0, 2

Term 6 = 2, 2, 0, 0, 4

(Term 1, Term 6) = 0×2 + 3×2 + 3×0 + 0×0 + 2×4

= 6 + 8

∴ Term 1, Term 6 = 14

Term 1, Term 7 :

Term 1 : 0, 3, 3, 0, 2

Term 7 : 1, 0, 3, 2, 0

(Term 1, Term 7) = 0×1 + 3×0 + 3×3 + 0×2 + 2×0

= 9.

∴ Term 1, Term 7 = 9

Term 1, Term 8 :

Term 1 : 0, 3, 3, 0, 2

Term 8 : 3, 1, 0, 0, 2

(Term 1, Term 8) = 0×3 + 3×1 + 3×0 + 0×0 + 2×2

= 0 + 3 + 4

∴ Term 1, Term 8 = 7

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Term1	–	7	16	15	14	14	9	7
Term2	7	–						
Term3	16		–					
Term4	15			–				
Term5	14				–			
Term6	14					–		
Term7	9						–	
Term8	7							–

Similarly Term2

Term 2, Term1

Term2 = 4, 1, 0, 1, 2

Term 1 = 0, 3, 3, 0, 2

Term2, Term1 = 4×0 + 1×3 + 0×3 + 1×0 + 2×2

= 3 + 4

∴ Term2, Term1 = 7

Term2, Term3 : Term2 = 4, 1, 0, 1, 2

Term3 = 0, 4, 0, 0, 2

Term2, Term3 = 4×0 + 1×4 + 0×0 + 1×0 + 2×2

= 4 + 4

∴ Term2, Term3 = 08

∴ Term2, Term4 :

Term2 = 4, 1, 0, 1, 2

Term4 = 0, 3, 0, 3, 3

(Term2, Term4) = 4×0 + 1×3 + 0×0 + 1×3 + 2×3

= 3 + 3 + 6

∴ Term2, Term4 = 12

Term2, Term5 :

Term2 = 4, 1, 0, 1, 2

Term5 = 0, 1, 3, 0, 1

(Term2, Term5) = 4×0 + 1×1 + 0×3 + 1×0 + 2×1

= 1 + 2

∴ Term2, Term5 = 3

Term2, Term6 :

Term2 = 4, 1, 0, 1, 2

Term6 = 2, 2, 0, 0, 4

(Term2, Term6) = 4×2 + 1×2 + 0×0 + 1×0 + 2×4

= 8 + 2 + 8

∴ Term2, Term6 = 18

Term2, Term7 :

Term2 = 4, 1, 0, 1, 2

Term7 = 1, 0, 3, 2, 0

(Term2, Term7) = 4×1 + 1×0 + 0×3 + 1×2 + 2×0

= 4 + 2

∴ Term2, Term7 = 6

Term 2, Term 8 :

Term 2 = 4, 1, 0, 1, 2

Term 8 = 3, 1, 0, 0, 2

(Term 2 , Term 8) = 4×3 + 1×1 + 0×0 + 1×0 + 2×2

= 12 + 1 + 4

∴ Term 2, Term 8 = 17

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Term 1	—	4	16	15	14	14	9	7
Term 2	4	—	8	12	3	18	16	17
Term 3	16	8	—					
Term 4	15	12		—				
Term 5	14	3			—			
Term 6	14	18				—		
Term 7	9	6					—	
Term 8	7	17						—

Similarly find for all the combinations.

i.e Term 3,	Term 4	Term 5	Term 6	Term 7	Term 8
Term 3, Term 1 = 16	Term 4, Term 1	Term 5, Term 1	Term 6, Term 1	Term 7, Term 1	Term 8 - Term 1
Term 3, Term 2 = 8	Term 4, Term 2	Term 5, Term 2	Term 6, Term 2	Term 7, Term 2	Term 8 - Term 2
Term 3, Term 4 = 18	Term 4, Term 3	Term 5, Term 3	Term 6, Term 3	Term 7, Term 3	Term 8 - Term 3
Term 3, Term 5 = 6	Term 4, Term 5	Term 5, Term 4	Term 6, Term 4	Term 7, Term 4	Term 8 - Term 4
Term 3, Term 6 = 16	Term 4, Term 6	Term 5, Term 6	Term 6, Term 5	Term 7, Term 5	Term 8 - Term 5
Term 3, Term 7 = 0	Term 4, Term 7	Term 5, Term 7	Term 6, Term 7	Term 7, Term 6	Term 8 - Term 6
Term 3, Term 8 = 8	Term 4, Term 8	Term 5, Term 8	Term 6, Term 8	Term 7, Term 8	Term 8 - Term 7

Resm2, Resm8 :

Resm2 = 4, 1, 0, 1, 2

Resm8 = 3, 1, 0, 0, 2

(Resm2, Resm8) = 4×3 + 1×1 + 0×0 + 1×0 + 2×2

= 12 + 1 + 4

∴ Resm2, Resm8 = 17

	Resm1	Resm2	Resm3	Resm4	Resm5	Resm6	Resm7	Resm8
Resm1	—	4	16	15	14	14	9	7
Resm2	4	—	8	12	3	18	16	17
Resm3	16	8	—					
Resm4	15	12		—				
Resm5	14	3			—			
Resm6	14	18				—		
Resm7	9	6					—	
Resm8	7	17						—

Similarly find for all the combinations.

i.e Resm3,	Resm4	Resm5	Resm6	Resm7	Resm8
Resm3, Resm1 = 16	Resm2, Resm1	Resm5, Resm1	Resm6, Resm1	Resm7, Resm1	Resm8 - Resm1
Resm3, Resm2 = 8	Resm4, Resm2	Resm5, Resm2	Resm6, Resm2	Resm7, Resm2	Resm8 - Resm2
Resm3, Resm4 = 18	Resm4, Resm3	Resm5, Resm3	Resm6, Resm3	Resm7, Resm3	Resm8 - Resm3
Resm3, Resm5 = 6	Resm4, Resm5	Resm5, Resm4	Resm6, Resm4	Resm7, Resm4	Resm8 - Resm4
Resm3, Resm6 = 16	Resm4, Resm6	Resm5, Resm6	Resm6, Resm5	Resm7, Resm5	Resm8 - Resm5
Resm3, Resm7 = 0	Resm4, Resm7	Resm5, Resm7	Resm6, Resm7	Resm7, Resm6	Resm8 - Resm6
Resm3, Resm8 = 8	Resm4, Resm8	Resm5, Resm8	Resm6, Resm8	Resm7, Resm8	Resm8 - Resm7

* **Term3**: $T_3 \rightarrow 0,4,0,0,2$
$T_1 \rightarrow 0,3,3,0,2$

$(Term3, Term1) = 0\times0+4\times3+0\times3+0\times0+2\times2$
$= 12+4$
$\therefore$ **Term3, Term1 = 16**

$(Term3, Term2)$ $T_3 \rightarrow 0,4,0,0,2$
$T_2 \rightarrow 4,1,0,1,2$

$(Term3, Term2) = 0\times4+4\times1+0\times0+0\times1+2\times2$
$= 4+4$
$\therefore$ **Term3, Term2 = 8**

Term3, Term4 :

$Term3 = 0,4,0,0,2$
$Term4 = 0,3,0,3,3$

$(T_3, T_4) = 0\times4\times3+0\times0+2\times3$
$= 12+6$
$T_3, T_4 = 18$

Term3, Term5 :

$Term3 \rightarrow 0,4,0,0,2$
$Term5 \rightarrow 0,1,3,0,1$

$(T_3, T_5) = 0+4\times1+0+0+2\times1$
$= 4+2$
$(T_3, T_5) = 6$

Term3, Term6 :

$Term3 = 0,4,0,0,2$
$Term6 = 2,2,0,0,4$

$(T_3, T_6) = 0\times2+4\times2+2\times4$
$= 8+8$
$T_3, T_6 = 16$

Term3, Term7 :

$Term3 = 0,4,0,0,2$
$Term7 = 1,0,3,2,0$

$(T_3, T_7) = 0\times1+4\times0+0\times3+0\times0+0\times2$
$= 0$

Term3, Term8

$Term3 = 0,4,0,0,2$
$Term8 = 3,1,0,0,2$

$(T_3, T_8) = 0\times3+4\times1+0\times0+0+2\times2$
$= 4+4$
$= 8$

i.e

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1	-	7	16	15	14	14	9	7
T_2	07	-	8	12	03	18	6	17
T_3	15	8	-	18	6	16	0	8
T_4	15	12	18	-				
T_5	14	03	6		-			
T_6	14	18	16			-		
T_7	9	6	0				-	
T_8	7	17	8					-

Prob) 4

$(T_4, T_1) = T_1 = 0,3,3,0,2$

$T_4 = 0,3,0,3,3$

$0\times0 + 3\times3 + 3\times0 + 0\times3 + 2\times3$

$= 9+6$

$T_4, T_2 = 15$

$(T_4, T_2) = T_4 = 0,3,0,3,3$

$T_2 = 4,1,0,1,2$

$0\times4 + 3\times1 + 0 + 3\times1 + 3\times2$

$= 3+3+6$

$T_4, T_2 = 12$

$(T_4, T_3) = T_4 = 0,3,0,3,3$

$T_3 = 0,4,0,0,2$

$0\times0 + 3\times4 + 0\times0 + 3\times0 + 3\times2$

$12+6$

$T_4, T_8 = 18$

$(T_4, T_5) = T_4 = 0,3,0,3,3$

$T_5 = 0,1,3,0,1$

$\Rightarrow 0\times0 + 3\times1 + 0\times3 + 3\times0 + 3\times1$

$= 3+3$

$T_4, T_5 = 6$

$(T_4, T_6) = T_4 = 0,3,0,3,3$

$T_6 = 2,2,0,0,4$

$= 0\times2 + 3\times2 + 0 + 0 + 3\times4$

$= 6+12$

$T_4, T_6 = 18$

$(T_4, T_7) = T_4 = 0,3,0,3,3$

$T_7 = 1,0,3,2,0$

$\Rightarrow 0+0+0+3\times2+0$

$T_4, T_7 = 6$

$(T_4, T_8) = T_4 = 0,3,0,3,3$

$T_8 = 3,1,0,0,2$

$0\times3 + 3\times1 + 0 + 0 + 3\times2$

$3+6$

$T_4, T_8 = 9$

i.e

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1	–	7	16	15	14	14	9	7
T_2	7	–	8	12	3	18	6	17
T_3	16	8	–	18	6	16	0	8
T_4	15	12	18	–	6	18	6	9
T_5	14	3	6	6	–			
T_6	14	18	16	18		–		
T_7	9	6	0	6			–	
T_8	7	17	8	9				–

Prob 4

$(T_4, T_1) = T_1 = 0,3,3,0,2$

$T_4 = 0,3,0,3,3$

$0\times0 + 3\times3 + 3\times0 + 0\times3 + 2\times3$

$= 9+6$

$T_4T_2 = 15$

$(T_4, T_2) = T_4 = 0,3,0,3,3$

$T_2 = 4,1,0,1,2$

$0\times4 + 3\times1 + 0 + 3\times1 + 3\times2$

$= 3+3+6$

$T_4T_2 = 12$

$(T_4, T_3) = T_4 = 0,3,0,3,3$

$T_3 = 0,4,0,0,2$

$0\times0 + 3\times4 + 0\times0 + 3\times0 + 3\times2$

$12+6$

$T_4T_3 = 18$

$(T_4, T_5) = T_4 = 0,3,0,3,3$

$T_5 = 0,1,3,0,1$

$\Rightarrow 0\times0 + 3\times1 + 0\times3 + 3\times0 + 3\times1$

$= 3+3$

$T_4T_5 = 6$

$(T_4, T_6) = T_4 = 0,3,0,3,3$

$T_6 = 2,2,0,0,4$

$= 0\times2 + 3\times2 + 0 + 0 + 3\times4$

$= 6+12$

$T_4T_6 = 18$

$(T_4, T_7) = T_4 = 0,3,0,3,3$

$T_7 = 1,0,3,2,0$

$\Rightarrow 0+0+0+3\times2+0$

$T_4T_7 = 6$

$(T_4, T_8) = T_4 = 0,3,0,3,3$

$T_8 = 3,1,0,0,2$

$0\times3 + 3\times1 + 0 + 0 + 3\times2$

$3+6$

$T_4T_8 = 9$

i.e

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1	–	7	16	15	14	14	9	7
T_2	7	–	8	12	3	18	6	17
T_3	16	8	–	18	6	16	0	8
T_4	15	12	18	–	6	18	6	9
T_5	14	3	6	6	–			
T_6	14	18	16	18		–		
T_7	9	6	0	6			–	
T_8	7	17	8	9				–

Terms

$(T_{terms5}, T_{terms1}) \Rightarrow T_{terms5} = 0,1,3,0,1$

$T_{terms1} = 0,3,3,0,2$

$0\times0+1\times3+3\times3+0+1\times2$

$= 3+9+2$

$T_5, T_1 = 14$

$(T_{terms5}, T_{terms2}) = T_{terms5} = 0,1,3,0,1$

$T_{terms2} = 4,1,0,1,2$

$= 0\times4+1\times1+3\times0+0\times1+1\times2$

$= 1+2$

$T_5, T_2 = 3$

$(T_{terms5}, T_{terms3}) = T_{terms5} = 0,1,3,0,1$

$T_{terms3} = 0,4,0,0,2$

$0\times0+1\times4+3\times0+0+1\times2$

$4+2$

$T_5, T_3 = 6$

$(T_{terms5}, T_{terms4}) = T_{terms5} = 0,1,3,0,1$

$T_{terms4} = 0,3,0,3,3$

$0\times0+1\times3+3\times0+0\times3+1\times3$

$3+3$

$T_5, T_4 = 6$

$(T_{terms5}, T_{terms6}) \Rightarrow T_{terms5} = 0,1,3,0,1$

$T_{terms6} = 2,2,0,0,4$

$0\times2+1\times2+3\times0+0+1\times4$

$2+4$

$T_5, T_6 = 6$

$(T_{terms5}, T_{terms7}) = T_{terms5} = 0,1,3,0,1$

$T_{terms7} = 1,0,3,2,0$

$0\times1+1\times0+3\times3+0\times2+1\times0$

$9+0$

$T_5, T_7 = 9$

$(T_{terms5}, T_{terms8}) = T_{terms5} = 0,1,3,0,1$

$T_{terms8} = 3,1,0,0,2$

$0\times3+1\times1+3\times0+0+1\times2$

$= 1+2$

$T_5, T_8 = 3$

i.e

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1	–	7	16	15	14	14	9	7
T_2	7	–	8	12	3	18	6	17
T_3	16	8	–	18	6	16	0	8
T_4	15	12	18	–	6	18	6	9
T_5	14	3	6	6	–	6	9	3
T_6	14	18	16	18	6	–		
T_7	9	6	0	6	9		–	
T_8	7	17	8	9	3			–

Term6

Term6 = 2, 2, 0, 0, 4

(Term6, Term1) ⇒ Term1 = 0, 3, 3, 0, 2

$2\times0+2\times3+0\times3+0+4\times2$

$= 6+8$

(T6, T1) = 14

Term6 = 2, 2, 0, 0, 4

(Term6, Term2) ⇒ Term2 = 4, 1, 0, 1, 2

$2\times4+2\times1+0+0+4\times2$

$8+2+8$

(T6, T2) = 18

Term6 = 2, 2, 0, 0, 4

(Term6, Term3) ⇒ Term3 = 0, 4, 0, 0, 2

$2\times0+2\times4+0+0+4\times2$

$8+8.$

(T6, T3) = 16

Term6 = 2, 2, 0, 0, 4

(Term6, Term4) ⇒ Term4 = 0, 3, 0, 3, 3

$2\times0+2\times3+0+0+4\times3$

$6+12$

(T6, T4) = 18

Term6 = 2, 2, 0, 0, 4

(Term6, Term5) ⇒ Term5 = 0, 1, 3, 0, 1

⇒ $2\times0+2\times1+0\times3+0\times0+4\times1$

⇒ $2+4$

(T6, T5) ⇒ 6

Term6 = 2, 2, 0, 0, 4

(Term6, Term7) ⇒ Term7 = 1, 0, 3, 2, 0

⇒ $2\times1+2\times0+0\times3+0\times2+4\times0$

(T6, T7) ⇒ 2

Term6 = 2, 2, 0, 0, 4

(Term6, Term8) ⇒ Term8 = 3, 1, 0, 0, 2

⇒ $2\times3+2\times1+0+0+4\times2$

⇒ $6+2+8$

(T6, T8) ⇒ 16

i.e	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1	–	7	16	15	14	14	9	7
T_2	7	–	8	12	3	18	6	17
T_3	16	8	–	18	6	16	0	8
T_4	15	12	18	–	6	18	6	9
T_5	14	3	6	6	–	6	9	3
T_6	14	18	16	18	6	–	2	16
T_7	9	6	0	6	9	2	–	
T_8	7	17	8	9	3	16		–

fig: Term-Term matrix (m)

Team7 ⇒ Team7 = 1,0,3,2,0
Team1 = 0,3,3,0,2

(Team7, Team1)

1×0+0×3+3×3+2×0+0×2

(T7, T1) = 9

Team7 = 1,0,3,2,0

(Team7, Team2) ⇒ Team2 = 4,1,0,1,2

1×4+0×1+3×0+2×1+0×2

4+2

(T7, T2) = 6

Team7 = 1,0,3,2,0

(Team7, Team3) ⇒ Team3 = 0,4,0,0,2

1×0+0×4+3×0+2×0+0×2

(T7, T3) = 0

Team7 = 1,0,3,2,0

(Team7, Team4) ⇒ Team4 = 0,3,0,3,3

1×0+0×3+3×0+2×3+0×3

(T7, T4) = 6

Team7 = 1,0,3,2,0

(Team7, Team5) ⇒ Team5 = 0,1,3,0,1

1×0+0×1+3×3+2×0+0×1

(T7, T5) = 9

Team7 = 1,0,3,2,0

(Team7, Team6) ⇒ Team6 = 2,2,0,0,4

1×2+0×2+3×0+2×0+0×4

(T7, T5) = 2

(Team7, Team8)

Team7 = 1,0,3,2,0

Team8 = 3,1,0,0,2

⇒ 1×3+0×1+3×0+2×0+0×2

⇒ 3+0

(T7, T8) = 3

Fig:

	T1	T2	T3	T4	T5	T6	T7	T8
T1	–	7	16	15	14	14	9	7
T2	7	–	8	12	3	18	6	17
T3	16	8	–	18	6	16	0	8
T4	15	12	18	–	6	18	6	9
T5	14	3	6	6	–	6	9	3
T6	14	18	16	18	6	–	2	16
T7	9	6	0	6	9	2	–	3
T8	7	17	8	9	3	16	3	–

Fig: Team–Team Matrix (m)

~~ii) Term Relationship Matrix~~

⇒ After two terms are considered similar, if the similarity value between them is 10 (or) > than 10.

∴ Threshold Value = 10 (or) > 10

ii) Term Relationship Matrix:

After comparing the ~~two~~ term values with threshold value. It produces a new "binary matrix" called the "Term Relationship Matrix".

Here, >10 values indicated with "1" and <10 values indicated with "0"

(Term2, Term1) = 4 threshold is >10 i.e (4>10)* condition failed so it is replaced by "0".

(Term3, Term1) = 16 16>10 ⇒ condition success, it is replaced by "1".

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
T_1	-	0	1	1	1	1	0	0
T_2	0	-	0	1	0	1	0	1
T_3	1	0	-	1	0	1	0	0
T_4	1	1	1	-	0	1	0	0
T_5	1	0	0	0	-	0	0	0
T_6	1	1	1	1	0	-	0	1
T_7	0	0	0	0	0	0	-	0
T_8	0	1	0	0	0	1	0	-

Fig: Term-Relationship Matrix.

Comparision

i.e:

Term1	Term2	Term3	Term4	T5	T6	T7	T8
7>10 = 0	*.7>10=0	16>10=1	15>10=1	14>10=1	14>10=1	9>10=0	7>10=0
16>10 = 1	* 8>10 = 0	8>10=0	12>10=1	3>10=0	18>10=1	6>10=0	17>10=1
15>10 = 1	* 12>10=1	18>10=1	18>10=1	6>10=0	16>10=1	0>10=0	8>10=0
14>10 = 1	* 3>10 = 0	6>10=0	6>10=0	6>10=0	18>10=1	6>10=0	9>10=0
14>10 = 1	* 18>10 = 1	16>10=1	18>10=1	6>10=0	6>10=0	9>10=0	3>10=0
9>10 = 0	* 6>10 = 0	0>10=0	6>10=0	9>10=0	2>10=0	2>10=0	16>10=1
7>10 = 0	* 17>10=1	8>10=0	9>10=0	3>10=0	16>10=1	3>10=0	3>10=0

iii) Network Diagram of Term Similarity :-

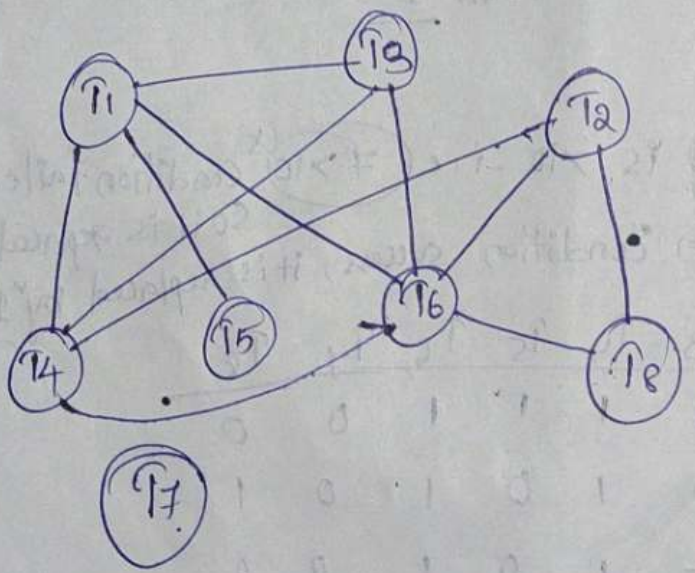


fig: Network diagram.

T1 : T3, T4, T5, T6

T2 : T4, T6, T8

T3 : T1, T4, T6

T4 : T1, T2, T3, T6

T5 : T1

T6 : T1, T2, T3, T4, T8

T8 : T2, T6

T7 : —

from Network diagram we get 2 classes :

class1 ($T_1, T_2, T_3, T_4, T_5, T_6, T_8$) :

There is a connectivity with other terms.

class2 (T_7) → There is no connectivity with other terms

(2) clustering using Existing clusters :-

An alternative methodology for creating clusters is to start with a set of "existing clusters".

A Graphical representation of terms and centroids illustrate below.

Centroid :-

Centroid is center of mass of a set of objects

Graphical representation :-

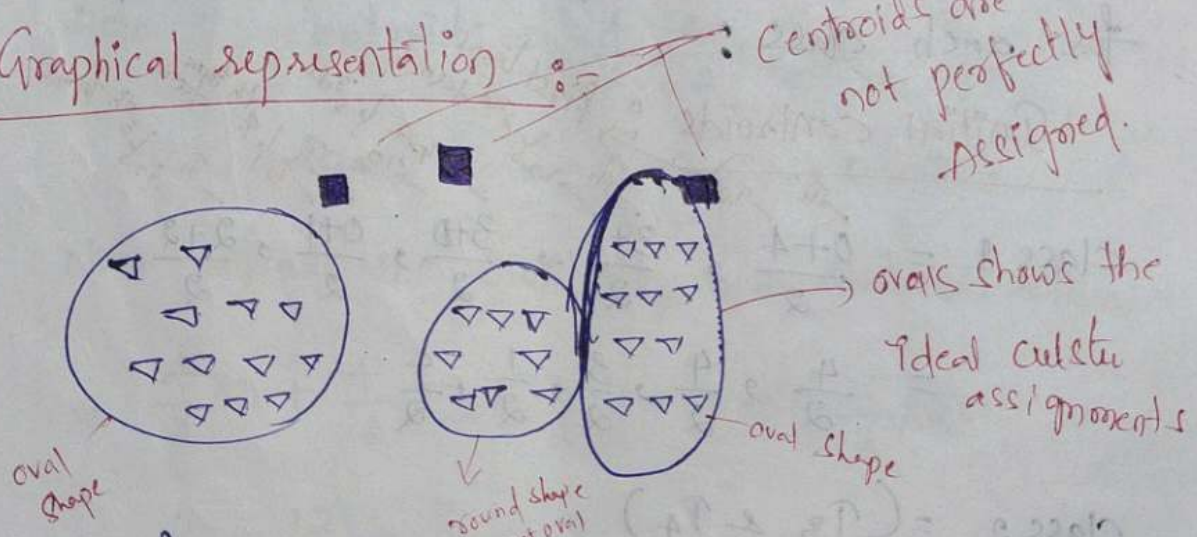
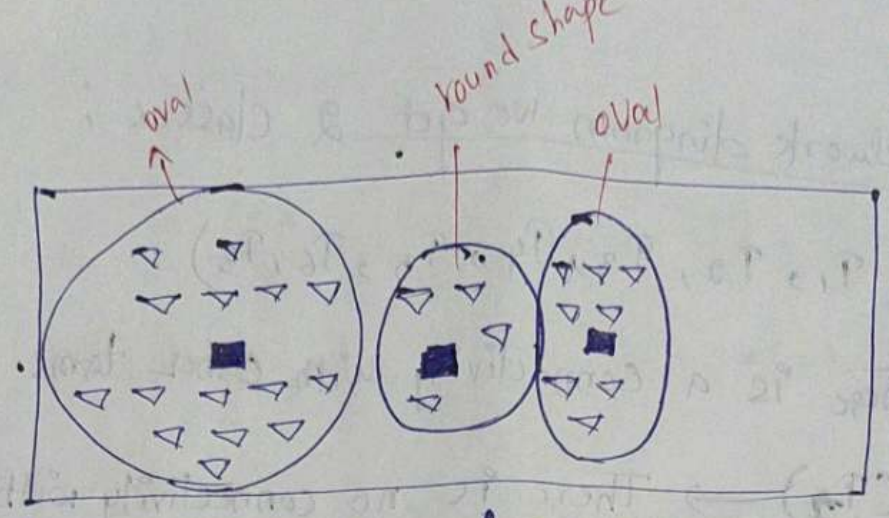


fig: Initial Centroids for clusters.



■ → represent the centriod for each classes.

fig: centroids after Reassigning Terms

For Example :—

Initial assigning for classes :—

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
I_1	0	4	0	0	0	2	1	3
I_2	3	1	4	3	1	2	0	1
I_3	3	0	6	0	3	0	3	0
I_4	0	1	0	3	0	0	2	0
I_5	2	2	2	3	1	4	0	2

fig: Vector Example

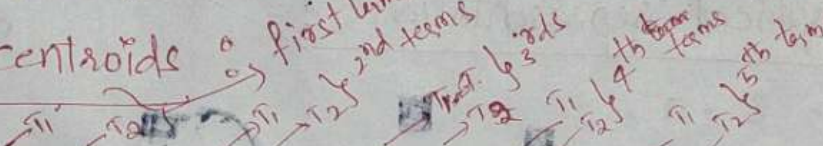
class 1 = (Term1 and Term2)

class 2 = (Term3 and Term4)

class 3 = (Term5 and Term6)

The above classes produce the following centroids for each class :

Initial centroids :



$$\text{class 1} = \frac{0+4}{2}, \frac{3+1}{2}, \frac{3+0}{2}, \frac{0+1}{2}, \frac{2+2}{2}$$

$$= \frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$$

class 2 = (T_3 & T_4)

$$= \frac{0+0}{2}, \frac{4+3}{2}, \frac{0+0}{2}, \frac{0+3}{2}, \frac{2+3}{2}$$

$$= \frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$$

class 3 = (Term 5 & Term 6)

$= \frac{0+2}{2}, \frac{1+2}{2}, \frac{3+0}{2}, \frac{0+0}{2}, \frac{1+4}{2}$

$= \frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$

i.e In class 1 the first value is calculated by averaging the weights of Term 1 & Term 2 in Item 1

Leave the values in fraction form.

Calculate the "Similary" for class 1, class 2, class 3:

(1) class-1 centroids $(\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2})$ for term 1, term 2, ---- term 8 :—

Term 1 : centroids x term 1

Here : centroids = $\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$

term 1 = 0, 3, 3, 0, 2 .

Now find the Similarity b/w centroids & term 1 .

$= \frac{4}{2}\times 0 + \frac{4}{2}\times 3 + \frac{3}{2}\times 3 + \frac{1}{2}\times 0 + \frac{4}{2}\times 2$

$= \frac{0}{2} + \frac{12}{2} + \frac{9}{2} + \frac{0}{2} + \frac{8}{2}$

Term 1 $= \frac{29}{2}$

Term2 :

Centroids : $\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$

Term 2 : 4, 1, 0, 1, 2

Similarity b/w them is :

$$\frac{4}{2}\times 4 + \frac{4}{2}\times 1 + \frac{3}{2}\times 0 + \frac{1}{2}\times 1 + \frac{4}{2}\times 2$$

$$\frac{16}{2} + \frac{4}{2} + \frac{0}{2} + \frac{1}{2} + \frac{8}{2}$$

$$\text{Term2} = \frac{29}{2}$$

Term 3

centroids : $\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$

Term 3 : 0, 4, 0, 0, 2

Similarity : $\frac{4}{2}\times 0 + \frac{4}{2}\times 4 + \frac{3}{2}\times 0 + \frac{1}{2}\times 0 + \frac{4}{2}\times 2$

$$= \frac{0}{2} + \frac{16}{2} + \frac{0}{2} + \frac{0}{2} + \frac{8}{2}$$

$$\text{Term3} = \frac{24}{2}$$

Term4 :

centroids : $\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$

Term 4 : 0, 3, 0, 3, 3

Similarity : $\frac{4}{2}\times 0 + \frac{4}{2}\times 3 + \frac{3}{2}\times 0 + \frac{1}{2}\times 3 + \frac{4}{2}\times 3$

$$= \frac{12}{2} + \frac{3}{2} + \frac{12}{2}$$

$$\text{Term4} = \frac{27}{2}$$

Term 5 :

centroids : $\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$

Term 5 : 0, 1, 3, 0, 1

Similarity : $\frac{4}{2}\times0 + \frac{4}{2}\times1 + \frac{3}{2}\times3 + \frac{1}{2}\times0 + \frac{4}{2}\times1$

$= \frac{4}{2} + \frac{9}{2} + \frac{4}{2}$

Term 5 = $\frac{17}{2}$

Term 6 :

Centroids : $\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$

Term 6 : 2, 2, 0, 0, 4

Similarity : $\frac{4}{2}\times2 + \frac{4}{2}\times2 + \frac{3}{2}\times0 + \frac{1}{2}\times0 + \frac{4}{2}\times4$

$= \frac{8}{2} + \frac{8}{2} + \frac{16}{2}$

Term 6 = $\frac{32}{2}$

Term 7

Centroids : $\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$

Term 7 : 1, 0, 3, 2, 0

Similarity : $\frac{4}{2}\times1 + \frac{4}{2}\times0 + \frac{3}{2}\times3 + \frac{1}{2}\times2 + \frac{4}{2}\times0$

$= \frac{4}{2} + \frac{9}{2} + \frac{2}{2}$

Term 7 = $\frac{15}{2}$

Term 8

centroids : $\frac{4}{2}, \frac{4}{2}, \frac{3}{2}, \frac{1}{2}, \frac{4}{2}$

Term 8 : 3, 1, 0, 0, 2

Similarity : $\frac{4}{2}\times3 + \frac{4}{2}\times1 + \frac{3}{2}\times0 + \frac{1}{2}\times0 + \frac{4}{2}\times2$

$= \frac{12}{2} + \frac{4}{2} + \frac{8}{2} \Rightarrow \frac{24}{2}$ i.e Term 8 = 24/2

class-2 centroids $(\frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2})$ for Term1, Term2

--- Term3) Calculate Similarity :-

centroids : $\frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term1

centroids : $\frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term1 : 0, 3, 3, 0, 2 ; from Vector Matrix.

Similarity :

$\Rightarrow \frac{0}{2}\times0 + \frac{7}{2}\times3 + \frac{0}{2}\times3 + \frac{3}{2}\times0 + \frac{5}{2}\times2$

$\Rightarrow \frac{21}{2} + \frac{10}{2}$

Term1 $\Rightarrow \frac{31}{2}$

Term2

centroids : $\frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term2 : 4, 1, 0, 1, 2

Similarity :

$= \frac{0}{2}\times4 + \frac{7}{2}\times1 + \frac{0}{2}\times0 + \frac{3}{2}\times1 + \frac{5}{2}\times2$

$= \frac{7}{2} + \frac{3}{2} + \frac{10}{2}$

Term2 $= \frac{20}{2}$

Term3 : centroids : $\frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term3 : 0, 4, 0, 0, 2

Similarity : $\frac{0}{2}\times0 + \frac{7}{2}\times4 + \frac{0}{2}\times0 + \frac{3}{2}\times0 + \frac{5}{2}\times2$

$= \frac{28}{2} + \frac{10}{2}$

Term3 $= \frac{38}{2}$

Term4 :

Centroids : $\frac{0}{2}, \frac{1}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term4 : 0, 3, 0, 3, 3

Similarity :

$= \frac{0}{2}\times0 + \frac{1}{2}\times3 + \frac{0}{2}\times0 + \frac{3}{2}\times3 + \frac{5}{2}\times3$

$= \frac{21}{2} + \frac{9}{2} + \frac{15}{2}$

Term4 = $\frac{45}{2}$

Term 5 :

Centroids: $\frac{0}{2}, \frac{1}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term 5 : 0, 1, 3, 0, 1

Similarity :

$\Rightarrow \frac{0}{2}\times0 + \frac{1}{2}\times1 + \frac{0}{2}\times3 + \frac{3}{2}\times0 + \frac{5}{2}\times1$

$= \frac{1}{2} + \frac{5}{2}$

Term5 = $\frac{12}{2}$

Term6 :

Centroids : $\frac{0}{2}, \frac{1}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term6 : 2, 2, 0, 0, 4

Similarity :

$= \frac{0}{2}\times2 + \frac{1}{2}\times2 + \frac{0}{2}\times0 + \frac{3}{2}\times0 + \frac{5}{2}\times4$

$= \frac{14}{2} + \frac{20}{2}$

Term6 = $\frac{34}{2}$

Term7 :

Centroids : $\frac{0}{2}, \frac{1}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term7 : 1, 0, 3, 2, 0

Similarity :

$= \frac{0}{2}\times1 + \frac{1}{2}\times0 + \frac{0}{2}\times3 + \frac{3}{2}\times2 + \frac{5}{2}\times0$

Term7 = $\frac{6}{2}$

Term8 :

Centroids: $\frac{0}{2}, \frac{1}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}$

Term8 : 3, 1, 0, 0, 2

Similarity :

$= \frac{0}{2}\times3 + \frac{1}{2}\times1 + \frac{0}{2}\times0 + \frac{3}{2}\times0 + \frac{5}{2}\times2$

$= \frac{1}{2} + \frac{10}{2}$

Term8 = $\frac{17}{2}$

Class-3 : centroids $(\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2})$ for Term1, Term2 -- Term8

Calculate: Similarity :-

centroids : $(\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2})$

Term1

centroids : $\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$

Term1 : 0, 3, 3, 0, 2

Similarity :

$\frac{2}{2} \times 0 + \frac{3}{2} \times 3 + \frac{3}{2} \times 3 + \frac{0}{2} \times 0 + \frac{5}{2} \times 2$

$= \frac{9}{2} + \frac{9}{2} + \frac{10}{2}$

Term1 = $\frac{28}{2}$

Term2

centroids : $\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$

Term2 : 4, 1, 0, 1, 2

Similarity :

$= \frac{2}{2} \times 4 + \frac{3}{2} \times 1 + \frac{3}{2} \times 0 + \frac{0}{2} \times 1 + \frac{5}{2} \times 2$

$= \frac{8}{2} + \frac{3}{2} + \frac{10}{2}$

Term2 = $\frac{21}{2}$

Term3

centroids : $\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$

Term3 : 0, 4, 0, 0, 2

Similarity :

$\Rightarrow \frac{2}{2} \times 0 + \frac{3}{2} \times 4 + \frac{3}{2} \times 0 + \frac{0}{2} \times 0 + \frac{5}{2} \times 2$

$\Rightarrow \frac{12}{2} + \frac{10}{2}$

Term3 = $\frac{22}{2}$

Term4

centroids : $\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$

Term4 : 0, 3, 0, 3, 3

Similarity :

$= \frac{2}{2} \times 0 + \frac{3}{2} \times 3 + \frac{3}{2} \times 0 + \frac{0}{2} \times 3 + \frac{5}{2} \times 3$

$= \frac{9}{2} + \frac{15}{2}$

Term4 = $\frac{24}{2}$

Term5

centroids : $\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$

Term5 : 0, 1, 3, 0, 1

Similarity :

$= \frac{2}{2}\times 0 + \frac{3}{2}\times 1 + \frac{3}{2}\times 3 + \frac{0}{2}\times 0 + \frac{5}{2}\times 1$

$= \frac{3}{2} + \frac{9}{2} + \frac{5}{2}$

Term5 = $\frac{17}{2}$

Term6

centroids: $\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$

Term6 : 2, 2, 0, 0, 4

Similarity:

$= \frac{2}{2}\times 2 + \frac{3}{2}\times 2 + \frac{3}{2}\times 0 + \frac{0}{2}\times 0 + \frac{5}{2}\times 4$

$= \frac{4}{2} + \frac{6}{2} + \frac{20}{2}$

Term6 = $\frac{30}{2}$

Term 7

centroids: $\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}$

Term7 : 1, 0, 3, 2, 0

Similarity:

$= \frac{2}{2}\times 1 + \frac{3}{2}\times 0 + \frac{3}{2}\times 3 + \frac{0}{2}\times 2 + \frac{5}{2}\times 0$

$= \frac{2}{2} + \frac{9}{2}$

Term7 = $\frac{11}{2}$

Term8

centroids : $(\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2})$

Term 8 : 3, 1, 0, 0, 2

Similarity:

$= \frac{2}{2}\times3 + \frac{3}{2}\times1 + \frac{3}{2}\times0 + \frac{0}{2}\times0 + \frac{5}{2}\times2$

$= \frac{6}{2} + \frac{3}{2} + \frac{10}{2}$

Term8 = $\frac{19}{2}$

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
class1	$\frac{29}{2}$	$\frac{29}{2}$	$\frac{24}{2}$	$\frac{27}{2}$	$\frac{17}{2}$	$\frac{32}{2}$	$\frac{15}{2}$ ✓	$\frac{24}{2}$ ✓
class2	31/2	20/2	38/2	47/2 ✓	12/2	34/2 ✓	6/2	17/2
class3	28/2	21/2	29/2	24/2	17/2	30/2	11/2	19/2
Assign	class 2	class 1	class 2	class 2	class 3	class 2	class1	class1

fig: Iterated class Assignments.

i.e To Assign class first compare class1, class2, class3 (weights) of Term1

Terms of Term1 ⇒ $\frac{29}{2}$, 31/2, ... column wise

	Term1
class1	29/2
class2	31/2 ← Highest value
class3	28/2

assign class2 for term 1

Similarly assign classes for all terms.

Next Iteration :—

Group the class 1, class 2, class 3 terms as follows.

Class 1 = (Term 2, Term 7, Term 8)

Class 2 = (Term 1, Term 3, Term 4, Term 6)

class 3 = (Term 5)

These New centroid classes are taken from fig: Iterated class Assignments.

$$\text{Class1} = \frac{\text{Term2} + \text{Term7} + \text{Term8}}{3}$$

$T_2 = 4, 1, 0, 1, 2$

$T_7 = 1, 0, 3, 2, 0$

$T_8 = 3, 1, 0, 0, 2$

$$= \left(\frac{4+1+3}{3}, \frac{1+0+1}{3}, \frac{0+3+0}{3}, \frac{1+2+0}{3}, \frac{2+0+2}{3}\right)$$

$$\text{class1} = \left(\frac{8}{3}, \frac{2}{3}, \frac{3}{3}, \frac{3}{3}, \frac{4}{3}\right)$$

$$\text{class2} = \frac{\text{Term1} + \text{Term3} + \text{Term4} + \text{Term6}}{4}$$

[$\because T_1 = 0, 3, 3, 0, 2$

$T_3 = 0, 4, 0, 0, 2$

$T_4 = 0, 3, 0, 3, 3$

$T_6 = 2, 2, 0, 0, 4$]

$$= \left(\frac{0+0+0+2}{4}, \frac{3+4+3+2}{4}, \frac{3+0+0+0}{4}, \frac{0+0+3+0}{4}, \frac{2+2+3+4}{4}\right)$$

$$\text{class2} = \left(\frac{2}{4}, \frac{12}{4}, \frac{3}{4}, \frac{3}{4}, \frac{11}{4}\right)$$

Class3 = Team5

$$= \frac{0, 1, 3, 0, 1}{1}$$

$$\text{Class3} = \frac{0}{1}, \frac{1}{1}, \frac{3}{1}, \frac{0}{1}, \frac{1}{1}$$

Class1 centroids for Team1, Team2 ... Team8

(Calculate Similarity)

Class1 centroids = $\left(\frac{8}{3}, \frac{2}{3}, \frac{3}{3}, \frac{3}{3}, \frac{4}{3}\right)$

Team1 = 0, 3, 3, 0, 2

Similarity:

$$= \frac{8}{3} \times 0 + \frac{2}{3} \times 3 + \frac{3}{3} \times 3 + \frac{3}{3} \times 0 + \frac{4}{3} \times 2$$

$$= \frac{6}{3} + \frac{9}{3} + \frac{8}{3}$$

$$\text{Team1} = \frac{23}{3}$$

Team2

Centroids: $\left(\frac{8}{3}, \frac{2}{3}, \frac{3}{3}, \frac{3}{3}, \frac{4}{3}\right)$

Team2: 4, 1, 0, 1, 2

Similarity:

$$= \frac{8}{3} \times 4 + \frac{2}{3} \times 1 + \frac{3}{3} \times 0 + \frac{3}{3} \times 1 + \frac{4}{3} \times 2$$

$$= \frac{32}{3} + \frac{2}{3} + \frac{3}{3} + \frac{8}{3}$$

$$\text{Team2} = \frac{45}{3}$$

Term3

Centroids: 8/3, 2/3, 3/3, 3/3, 4/3

Term3 : 0, 4, 0, 0, 2

Similarity:

$= \frac{8}{3} \times 0 + \frac{2}{3} \times 4 + \frac{3}{3} \times 0 + \frac{3}{3} \times 0 + \frac{4}{3} \times 2$

$= \frac{8}{3} + \frac{8}{3}$

Term3 = $\frac{16}{3}$

Term4

Centroids: 8/3, 2/3, 3/3, 3/3, 4/3

Term4 : 0, 3, 0, 3, 3

Similarity:

$= \frac{8}{3} \times 0 + \frac{2}{3} \times 3 + \frac{3}{3} \times 0 + \frac{3}{3} \times 3 + \frac{4}{3} \times 3$

$= \frac{6}{3} + \frac{9}{3} + \frac{12}{3}$

Term4 = $\frac{27}{3}$

Term5

Centroids: 8/3, 2/3, 3/3, 3/3, 4/3

Term5 : 0, 1, 3, 0, 1

Similarity:

$= \frac{8}{3} \times 0 + \frac{2}{3} \times 1 + \frac{3}{3} \times 3 + \frac{3}{3} \times 0 + \frac{4}{3} \times 1$

$= \frac{2}{3} + \frac{9}{3} + \frac{4}{3}$

Term5 = $\frac{15}{3}$

Term6

Centroid: 8/3, 2/3, 3/3, 3/3, 4/3

Term6 : 2, 2, 0, 0, 4

Similarity :

$$= \frac{8}{3}\times 2 + \frac{2}{3}\times 2 + \frac{3}{3}\times 0 + \frac{3}{3}\times 0 + \frac{4}{3}\times 4$$

$$= \frac{16}{3} + \frac{4}{3} + \frac{0}{3} + \frac{0}{3} + \frac{16}{3}$$

Term6 = $\frac{36}{3}$

Term7

Centroid: 8/3, 2/3, 3/3, 3/3, 4/3

Term7 : 1, 0, 3, 2, 0

Similarity

$$= \frac{8}{3}\times 1 + \frac{2}{3}\times 0 + \frac{3}{3}\times 3 + \frac{3}{3}\times 2 + \frac{4}{3}\times 0$$

$$= \frac{8}{3} + 0 + \frac{9}{3} + \frac{6}{3} + 0$$

Term7 = $\frac{23}{3}$

Term8

Centroid: 8/3, 2/3, 3/3, 3/3, 4/3

Term8 : 3, 1, 0, 0, 2

Similarity

$$= \frac{8}{3}\times 3 + \frac{2}{3}\times 1 + \frac{3}{3}\times 0 + \frac{3}{3}\times 0 + \frac{4}{3}\times 2$$

$$= \frac{24}{3} + \frac{2}{3} + \frac{8}{3}$$

Term8 = $\frac{34}{3}$

* Class2 centroids for term1, term2 --- term8

Calculate Similarity :-

Class2 centroid $= \left(\frac{2}{4}, \frac{12}{4}, \frac{3}{4}, \frac{3}{4}, \frac{11}{4}\right)$.

Term1

centroid : 2/4, 12/4, 3/4, 3/4, 11/4

Term1 : 0, 3, 3, 0, 2

Similarity

$$= \frac{2}{4}\times 0 + \frac{12}{4}\times 3 + \frac{3}{4}\times 3 + \frac{3}{4}\times 0 + \frac{11}{4}\times 2$$

$$= \frac{36}{4} + \frac{9}{4} + \frac{22}{4}$$

$$\text{Term1} = \frac{67}{4}$$

Term2

centroid : 2/4, 12/4, 3/4, 3/4, 11/4

Term2 : 4, 1, 0, 1, 2

Similarity :

$$= \frac{2}{4}\times 4 + \frac{12}{4}\times 1 + \frac{3}{4}\times 0 + \frac{3}{4}\times 1 + \frac{11}{4}\times 2$$

$$= \frac{8}{4} + \frac{12}{4} + \frac{3}{4} + \frac{22}{4}$$

$$\text{Term2} = \frac{45}{4}$$

Term3

centroid : 2/4, 12/4, 3/4, 3/4, 11/4

Term3 : 0, 4, 0, 0, 2

Similarity

$$= \frac{2}{4}\times 0 + \frac{12}{4}\times 4 + \frac{3}{4}*0 + \frac{3}{4}*0 + \frac{11}{4}\times 2$$

$$= \frac{48}{4} + \frac{22}{4} \Rightarrow \text{Term3} = 70/4$$

Term4

Centroid: 2/4, 12/4, 3/4, 3/4, 11/4

Term 4 : 0, 3, 0, 3, 3

Similarity :

$$= \frac{2}{4}\times 0 + \frac{12}{4}\times 3 + \frac{3}{4}\times 0 + \frac{3}{4}\times 3 + \frac{11}{4}\times 3$$

$$= \frac{36}{4} + \frac{9}{4} + \frac{33}{4}$$

Term4 = $\frac{78}{4}$

Term5

Centroid: 2/4, 12/4, 3/4, 3/4, 11/4

Term 5 : 0, 1, 3, 0, 1

Similarity

$$= \frac{2}{4}\times 0 + \frac{12}{4}\times 1 + \frac{3}{4}\times 3 + \frac{3}{4}\times 0 + \frac{11}{4}\times 1$$

$$= \frac{12}{4} + \frac{9}{4} + \frac{11}{4}$$

Term5 = $\frac{32}{4}$

Term6

Centroid: 2/4, 12/4, 3/4, 3/4, 11/4

Term 6 : 2, 2, 0, 0, 4

Similarity

$$= \frac{2}{4}\times 2 + \frac{12}{4}\times 2 + \frac{3}{4}\times 0 + \frac{3}{4}\times 0 + \frac{4}{4}\times 4$$

$$= \frac{4}{4} + \frac{24}{4} + \frac{44}{4}$$

Term6 = $\frac{72}{4}$

Term 7

Centroid : 2/4, 12/4, 3/4, 3/4, 11/4

Term 7 : 1, 0, 3, 2, 0

Similarity

$= \frac{2}{4} \times 1 + \frac{12}{4} \times 0 + \frac{3}{4} \times 3 + \frac{3}{4} \times 2 + \frac{11}{4} \times 0$

$= \frac{2}{4} + \frac{9}{4} + \frac{6}{4}$

Term 7 = $\frac{17}{4}$

Term 8

Centroid : 2/4, 12/4, 3/4, 3/4, 11/4

Term 8 : 3, 1, 0, 0, 2

Similarity

$= \frac{2}{4} \times 3 + \frac{12}{4} \times 1 + \frac{3}{4} \times 0 + \frac{3}{4} \times 0 + \frac{11}{4} \times 2$

$= \frac{6}{4} + \frac{12}{4} + \frac{22}{4}$

Term 8 = $\frac{40}{4}$

* Class 3 Centroids for Term 1, Term 2, --- Term 8.
(Calculate Similarity)

class 3 centroids : $\frac{0}{T}, \frac{1}{T}, \frac{3}{T}, \frac{0}{T}, \frac{1}{T}$

Term 1

Centroid : 0/1, 1/1, 3/1, 0/1, 1/1

Term 1 : 0, 3, 3, 0, 2

Similarity:

$\frac{0}{1} \times 0 + \frac{1}{1} \times 3 + \frac{3}{1} \times 3 + \frac{0}{1} \times 0 + \frac{1}{1} \times 2$

$\Rightarrow \frac{0}{1}\times 0 + \frac{1}{1}\times 3 + \frac{3}{1}\times 3 + \frac{0}{1}\times 0 + \frac{1}{1}\times 2$

$\Rightarrow 3+9+2$

Term1 = 14/1

Term2

centroid: $(\frac{0}{1}, \frac{1}{1}, \frac{3}{1}, \frac{0}{1}, \frac{1}{1})$

Term2 : 4, 1, 0, 1, 2

Similarity

$= \frac{0}{1}\times 4 + \frac{1}{1}\times 1 + \frac{3}{1}\times 0 + \frac{0}{1}\times 1 + \frac{1}{1}\times 2$

$= \frac{1}{1} + \frac{2}{1}$

Term2 = 3/1

Term3

centroid: 0/1, 1/1, 3/1, 0/1, 1/1

Term3 : 0, 4, 0, 0, 2

Similarity

$= 0/1 \times 0 + 1/1 \times 4 + 3/1 \times 0 + 0/1 \times 0 + 1/1 \times 2$

$= 4/1 + 2/1$

Term3 = 6/1

Term4

Centroid: 0/1, 1/1, 3/1, 0/1, 1/1

Term4 : 0, 3, 0, 3, 3

Similarity

$= 0/1 \times 0 + 1/1 \times 3 + 3/1 \times 0 + 0/1 \times 3 + 1/1 \times 3$

$= 3/1 + 3/1$

Term4 = 6/1

Term5

centroid: 0/1, 1/1, 3/1, 0/1, 1/1

Term5 : 0, 1, 3, 0, 1

Similarity

$= 0/1 \times 0 + 1/1 \times 1 + 3/1 \times 3 + 0/1 \times 0 + 1/1 \times 1$

$= 1/1 + 9/1 + 1/1$

Term5 = 11/1

Term6

Centroid: 0/1, 1/1, 3/1, 0/1, 1/1

Term6 : 2, 2, 0, 0, 4

Similarity

$= 0/1 \times 2 + 1/1 \times 2 + 3/1 \times 0 + 0/1 \times 0 + 1/1 \times 4$

$= 2/1 + 4/1$

Term6 = 6/1

Term7

Centroid: 0/1, 1/1, 3/1, 0/1, 1/1

Term7 : 1, 0, 3, 2, 0

Similarity

$= 0/1 \times 1 + 1/1 \times 0 + 3/1 \times 3 + 0/1 \times 2 + 1/1 \times 0$

$= 9/1 + 0$

Term7 = 9/1

Terms

Centroid: 0/1, 1/1, 3/1, 0/1, 1/1

Terms : 3, 1, 0, 0, 2

Similarity:

$= 0/1 \times 3 + 1/1 \times 1 + 3/1 \times 0 + 0/1 \times 0 + 1/1 \times 2$

$= 0 + 1/1 + 0/1 + 2/1$

Terms = 3/1

"Highest weight" similarly assign for all term classes

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
class 1	23/3	45/3	16/3	27/3	15/3	36/3	23/3	34/3
class 2	67/4	45/4	70/4	78/4	32/4	72/4	17/4	40/4
class 3	14/1	3/1	6/1	6/1	11/1	6/1	9/1	3/1
Assign	class 2	class 1	class 2	class 2	class 3	class 2	class 3	class 1

[* assign higest weight classes to Terms i.e in Term 1 highest weight is 67/4 => i.e class 2]

fig: New centroids & cluster Assignments

In this iteration of the Process, only the change is "Term 7" moves from "class 1 to class 3".

i.e Compare class 7 in "Iterated class Assignments" with class 7 in "New centroid & cluster Assignments".

* It was not strongly related to other terms in class 1.

class 7: 15/2, 6/2, 11/2 → Iterated class Assignment

class 7: 23/3, 17/4, 9/1 → New centroid & cluster Assignments

(3) One pass Assignments :-

This Technique has minimum "Overhead" that is in "onepass" of all the terms is used to assign terms to classes.

In this process,

→ The "first term" is assigned to the "first class".

→ Each additional term is compared to the centroids of the existing classes.

→ A Threshold is choosen.

→ Threshold is compared with highest ↑ similary.

from the previous term-term matrix, threshold value is 10.

~~for~~ following classes are generated.

Class 1 = Term 1, Term 3, Term 4 } from Network security.

Class 2 = Term 2, Term 6, Term 8 }

Class 3 = Term 5 → only one connection } diagram.

Class 4 = Term 7 → alone

Centroid Values during one-pass process :-

Class 1 = (Term 1, Term 3)

Class 2 (Term 1, Term 3, Term 4)

Class 3 (Term 2, Term 6)

* Class 1 (Term 1, Term 3) :-

Term 1 = 0, 3, 3, 0, 2

Term 3 = 0, 4, 0, 0, 2

$$\left(\frac{0+0}{2}, \frac{3+4}{2}, \frac{3+0}{2}, \frac{0+0}{2}, \frac{2+2}{2}\right)$$

Class 1 : $0, \frac{7}{2}, \frac{3}{2}, 0, \frac{4}{2}$
centroids

* Class 2 (Term 1, Term 3, Term 4) :-

Term 1 = 0, 3, 3, 0, 2

Term 3 = 0, 4, 0, 0, 2

Term 4 = 0, 3, 0, 3, 3

$$= \left(\frac{0+0+0}{3}, \frac{3+4+3}{3}, \frac{3+0+0}{3}, \frac{0+0+3}{3}, \frac{2+2+3}{3}\right)$$

Class 2 : $\left(0, \frac{10}{3}, \frac{3}{3}, \frac{3}{3}, \frac{7}{3}\right)$
Centroids

* Class 3 (Term 2, Term 6) :-

Term 2 = 4, 1, 0, 1, 2

Term 6 = 2, 2, 0, 0, 4

$$= \left(\frac{4+2}{2}, \frac{1+2}{2}, \frac{0+0}{2}, \frac{1+0}{2}, \frac{2+4}{2}\right)$$

Class 3 = $\frac{6}{2}, \frac{3}{2}, 0, \frac{1}{2}, \frac{6}{2}$
Centroids

(*) Item clustering:-

Item clustering is very similar to term-clustering for the generation of thesauri.

This Technique described for the clustering of terms (or) item clustering.

Concept:-

In this concept the "Similarity" between the document is based on "two items" that have "terms in Common" Versus "terms with items in common".

Similarity fuction:

The Similarity function is performed between rows of item index.

Similarity Equation:

$$SIM(Item_i, Item_j) = \Sigma (Term_{i,k})(Term_{j,k})$$

The above equation indicates the "Set of items and their terms"

Here, "K" goes from 1 to 8 for the eight terms.

See example

The following actions (Calculations) are performed in "item Clustering".

They are:

i) An "item - item Cos" "item / item Matrix" Created

ii) The "Item Relationship Matrix" is created by using a threshold value of "10"

Step-1

Find the Similarity between "items":-

For Example:

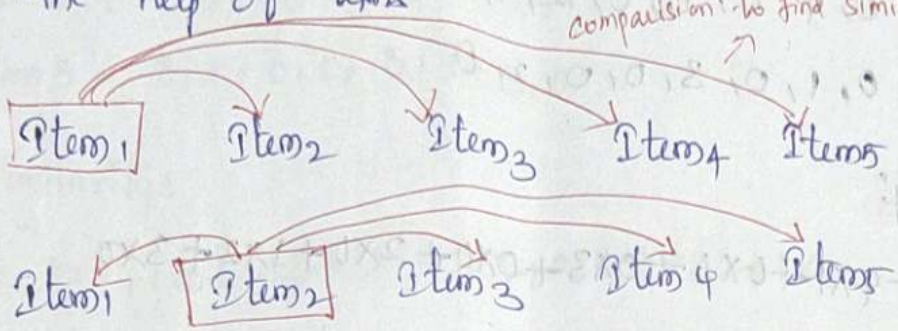
	Item1	I2	I3	I4	I5	I6	I7	I8
Item1	0	4	0	0	0	2	1	3
Item2	3	1	4	3	1	2	0	1
Item3	3	0	0	0	3	0	3	0
Item4	0	1	0	3	0	0	2	0
Item5	2	2	2	3	1	4	0	2

"row wise" Similarly all.

fig: Vector Example.

Now find the similarity between Items (i.e Item1 & Item---) with the help of items.

i.e



Similarly remaining items Compare with other items.

1) Item1 . Item2 (Similarity)

Item1 : 0, 4, 0, 0, 0, 2, 1, 3
Item2 : 3, 1, 4, 3, 1, 2, 0, 1
} similarity b/w rows

Similarity:

$= 0\times3 + 4\times1 + 0\times4 + 0\times3 + 0\times1 + 2\times2 + 1\times0 + 3\times1$

$= 4 + 4 + 3$

$= 11$

i.e **Item1 . Item2 = 11**

(Item1 . Item3):

Item1 : 0, 4, 0, 0, 0, 2, 1, 3

Item3 : 3, 0, 0, 0, 3, 0, 3, 0

Similarity:

$= 0\times3 + 4\times0 + 0\times0 + 0\times0 + 0\times3 + 2\times0 + 1\times3 + 3\times0$

$= 3$

∴ **Item1 . Item3 = 3**

(Item 1 . Item 4)

Item 1 : 0, 4, 0, 0, 0, 2, 1, 3

Item 4 : 0, 1, 0, 3, 0, 0, 2, 0

Similarity:

$= 0\times0 + 4\times1 + 0\times0 + 0\times3 + 0\times0 + 2\times0 + 1\times2 + 3\times0$

$= 4 + 2$

$= 6$

∴ Item 1 . Item 4 = 6

(Item 1 . Item 5)

Item 1 : 0, 4, 0, 0, 0, 2, 1, 3

Item 5 : 2, 2, 2, 3, 1, 4, 0, 2

Similarity:

$= 0\times2 + 4\times2 + 0\times2 + 0\times3 + 0\times1 + 2\times4 + 1\times0 + 3\times2$

$= 8 + 8 + 6$

$= 22$

∴ Item 1 . Item 5 = 22

	Item 1	Item 2	I 3	I 4	I 5
Item 1	–	11	3	6	22
Item 2	11	–			
Item 3	3		–		
Item 4	6			–	
Item 5	22				–

fig: Item/Item matrix.

2) Item 2 .

(Item 2 . Item 1)

Item 2 : 3, 1, 4, 3, 1, 2, 0, 1

Item 1 : 0, 4, 0, 0, 0, 2, 1, 3

Similarity : $3\times0 + 1\times4 + 4\times0 + 3\times0 + 1\times0 + 2\times2 + 0\times1 + 1\times3$

$= 4 + 4 + 3$

∴ Item 2 . Item 1 = 11

Item 2 . Item 3

Item 2 : 3, 1, 4, 3, 1, 2, 0, 1

Item 3 : 3, 0, 0, 0, 3, 0, 3, 0

Similarity:

$= 3\times3+1\times0+4\times0+3\times0+1\times3+2\times0+0\times3+1\times0$

$= 9+3+0$

$= 12$

$\therefore$ **Item 2 . Item 3 = 12**

Item 2 . Item 4

Item 2 : 3, 1, 4, 3, 1, 2, 0, 1

Item 4 : 0, 1, 0, 3, 0, 0, 2, 0

Similarity:

$= 3\times0+1\times1+4\times0+3\times3+1\times0+2\times0+0\times2+1\times0$

$= 1+9$

$= 10$

$\therefore$ **Item 2 . Item 4 = 10**

Item 2 . Item 5

Item 2 : 3, 1, 4, 3, 1, 2, 0, 1

Item 5 : 2, 2, 2, 3, 1, 4, 0, 2

Similarity:

$= 3\times2+1\times2+4\times2+3\times3+1\times1+2\times4+1\times2$

$= 6+2+8+9+1+8+2$

$= 36$

$\therefore$ **Item 2 . Item 5 = 36**

ii e

	Item1	Item2	Item3	Item4	Item5
Item1	–	11	3	6	22
Item2	11	–	12	10	36
Item3	3	12	–		
Item4	6	10		–	
Item5	22	36			

Item 3

(Item 3. Item 1)

Item 3 : 3, 0, 0, 0, 3, 0, 3, 0

Item 1 : 0, 4, 0, 0, 0, 2, 1, 3

Similarity:

$= 3\times0 + 0\times4 + 0 + 0 + 3\times0 + 0\times2 + 3\times1 + 0\times3$

$= 3$

$\therefore$ Item3. Item1 = 3

(Item 3 · Item2)

Item 3: 3, 0, 0, 0, 3, 0, 3, 0

Item 2: 3, 1, 4, 3, 1, 2, 0, 1

Similarity:

$= 3\times3 + 0\times1 + 0\times4 + 0\times3 + 3\times1 + 0\times2 + 3\times0 + 0\times1$

$= 9+3$

Item3·I4 = 12

Item3 . Item4

Item3 : 3, 0, 0, 0, 3, 0, 3, 0

Item4 : 0, 1, 0, 3, 0, 0, 2, 0

Similarity:

$= 3\times0 + 0\times1 + 0\times0 + 0\times3 + 0\times3 + 0\times0 + 3\times2 + 0\times0$

$= 6$

Item3 . Item4 = 6

(Item3 . Item5)

Item3 = 3, 0, 0, 0, 3, 0, 3, 0

Item5 = 2, 2, 2, 3, 1, 4, 0, 2

Similarity:

$= 3\times2 + 0\times2 + 0\times2 + 0\times3 + 3\times1 + 0\times4 + 3\times0 + 0\times2$

$= 6 + 3 + 0$

$= 9$

Item3 . Item5 = 9

i.e.

	I_1	I_2	I_3	I_4	I_5
Item1	-	11	3	6	22
Item2	11	-	12	10	36
Item3	3	12	-	6	9
Item4	6	10	6	-	
Item5	22	36	9		-

Fig: Item / Item Matrix

Item4 :

(Item 4 . Item1) :

Item4 : 0,1,0,3,0,0,2,0

Item1 : 0,4,0,0,0,2,1,3

Similarity:

$= 0\times0+1\times4+0\times0+3\times0+0\times0+0\times2+2\times1+0\times3$

$= 4+2+0$

$= 6$

$\therefore$ **Item4 . Item1 = 6**

(Item4 . Item2) :

Item4 : 0,1,0,3,0,0,2,0

Item2 : 3,1,4,3,1,2,0,1

Similarity:

$= 0\times3+1\times1+0\times4+3\times3+0\times1+0\times2+2\times0+0\times1$

$= 1+9+0$

$= 10$

$\therefore$ **Item 4 . Item2 = 10**

(Item4 . Item5) :

Item4 : 0,1,0,3,0,0,2,0

Item5 : 2,2,2,3,1,4,0,2

Similarity:

$= 0\times2+1\times2+0\times2+3\times3+0\times1+0\times4+2\times0+0\times2$

$= 2+9$

$= 11$

$\therefore$ **Item 4 . Item5 = 11**

(Item4 · Item3) :

Item 4 : 0, 1, 0, 3, 0, 0, 2, 0

Item 3 : 3, 0, 0, 0, 3, 0, 3, 0

Similarity:

$= 0\times3 + 1\times0 + 0\times0 + 3\times0 + 0\times3 + 0\times0 + 2\times3 + 0\times0$

$= 6$

$\therefore$ **Item4 · Item3 = 6**

i.e	Item1	Item2	Item3	Item4	Item5
Item1	—	11	3	6	22
Item2	11	—	12	10	36
Item3	3	12	—	6	9
Item4	6	10	6	—	11
Item5	22	36	9	11	—

fig: item / item Matrix

→ Next step is "Item - Relationship Matrix".

ii) Item - Relationship Matrix :-

The Threshold of 10 produces the Item Relationship - Matrix.

Threshold Value is 10 (or) >10

If similarity is > threshold value i.e >10 replace it with "1" if it is <10 replace value with "zero"

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	–	1	0	0	1
Item 2	1	–	1	1	1
Item 3	0	1	–	0	0
Item 4	0	1	0	–	1
Item 5	1	1	0	1	–

Fig : Item Relationship Matrix

ie see how 0 & 1's assign:

11>10=1
12>10=1
10>10=1
36>10=1

3>10=0
6>10=0
22>10=1

3>10=0
12>10=1
6>10=0
9>10=0

6>10=0
10≥10=1
6>10=0
11>10=1

22>10=1
36>10=1
9>10=0
11>10=1

clustering starting with "existing clusters" can be performed similar to item model. (ie centroids)

Initial Assignment classes for centroids :

Item 1 and Item 3 in class 1

Item 2 and Item 4 in class 2

i.e Class 1 = (Item 1, Item 3)

Class 2 = (Item 2, Item 4)

Initial centroids :-

Class 1 = (Item 1, Item 3)

Item 1 = 0, 4, 0, 0, 0, 2, 1, 3

Item 3 = 3, 0, 0, 0, 3, 0, 3, 0

Centroids : $\left(\frac{0+3}{2}, \frac{4+0}{2}, \frac{0+0}{2}, \frac{0+0}{2}, \frac{0+3}{2}, \frac{2+0}{2}, \frac{1+3}{2}, \frac{3+0}{2}\right)$

Class 1 centroids = $\frac{3}{2}, \frac{4}{2}, \frac{0}{2}, \frac{0}{2}, \frac{3}{2}, \frac{2}{2}, \frac{4}{2}, \frac{3}{2}$

Class 2 = (Item 2, Item 4)

Item 2 : 3, 1, 4, 3, 1, 2, 0, 1

Item 4 : 0, 1, 0, 3, 0, 0, 2, 0

Centroids : $\left(\frac{3+0}{2}, \frac{1+1}{2}, \frac{4+0}{2}, \frac{3+3}{2}, \frac{1+0}{2}, \frac{2+0}{2}, \frac{0+2}{2}, \frac{1+0}{2}\right)$

Class 2 centroids : $\frac{3}{2}, \frac{2}{2}, \frac{4}{2}, \frac{6}{2}, \frac{1}{2}, \frac{2}{2}, \frac{2}{2}, \frac{1}{2}$

(Calculate Similarity)

* Class1 centroids for item1, item2 items

Code: class1 : centroids x item1 :-

class1 ⇒ centroids : $\frac{3}{2}, \frac{4}{2}, \frac{0}{2}, \frac{0}{2}, \frac{3}{2}, \frac{2}{2}, \frac{4}{2}, \frac{3}{2}$

item1 : 0, 4, 0, 0, 0, 2, 1, 3

similarity :

$\Rightarrow \frac{3}{2}\times 0 + \frac{4}{2}\times 4 + \frac{0}{2}\times 0 + \frac{0}{2}\times 0 + \frac{3}{2}\times 0 + \frac{2}{2}\times 2 + \frac{4}{2}\times 1 + \frac{3}{2}\times 3$

$= \frac{0}{2} + \frac{16}{2} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} + \frac{4}{2} + \frac{4}{2} + \frac{9}{2}$

$= \frac{16+4+4+9}{2}$

item1 $= \frac{33}{2}$

class 1 : centroids x item2 :

centroids : 3/2, 4/2, 0/2, 0/2, 3/2, 2/2, 4/2, 3/2

item2 : 3, 1, 4, 3, 1, 2, 0, 1

Similarity :

$= \frac{3}{2}\times 3 + \frac{4}{2}\times 1 + \frac{0}{2}\times 4 + \frac{0}{2}\times 3 + \frac{3}{2}\times 1 + \frac{2}{2}\times 2 + \frac{4}{2}\times 0 + \frac{3}{2}\times 1$

$= 9/2 + 4/2 + 0 + 0 + 3/2 + 4/2 + 3/2$

item2 $= 23/2$

Class1 : centroids x item3 :

centroids : 3/2, 4/2, 0/2, 0/2, 3/2, 2/2, 4/2, 3/2

item3 : 3, 0, 0, 0, 3, 0, 3, 0

Similarity :

$= 3/2\times 3 + 4/2\times 0 + 0/2\times 0 + 0/2\times 0 + 3/2\times 3 + 2/2\times 0 + 4/2\times 3 + 3/2\times 0$

$= 9/2 + 9/2 + 12/2$

Item 3 = $30/2$

Class 1 centroids x item4 :

Centroids : $3/2, 4/2, 0/2, 0/2, 3/2, 2/2, 4/2, 3/2$

item4 : 0, 1, 0, 3, 0, 0, 2, 0

Similarity :

$= 3/2 \times 0 + 4/2 \times 1 + 0/2 \times 0 + 0/2 \times 3 + 3/2 \times 0 + 2/2 \times 0 + 4/2 \times 2 + 3/2 \times 0$

$= 4/2 + 8/2$

item4 = $12/2$

Class 1 Centroids x item5 :

Centroids : $3/2, 4/2, 0/2, 0/2, 3/2, 2/2, 4/2, 3/2$

item5 : 2, 2, 2, 3, 1, 4, 0, 2

Similarity :

$= 3/2 \times 2 + 4/2 \times 2 + 0/2 \times 2 + 0/2 \times 3 + 3/2 \times 1 + 2/2 \times 4 + 4/2 \times 0 + 3/2 \times 2$

$= 6/2 + 8/2 + 3/2 + 8/2 + 6/2$

item5 = $8/2$

* class 2 centroids for item1, item2 - - - item5.

Calculate Similarity: Class 2 centroid: 3/2, 2/2, 4/2, 6/2, 1/2, 2/2, 2/2, 1/2

Class 2: centroids x item1

centroids: 3/2, 2/2, 4/2, 6/2, 1/2, 2/2, 2/2, 1/2

item1 : 0, 4, 0, 0, 0, 2, 1, 3

Similarity:

$= 3/2 \times 0 + 2/2 \times 4 + 4/2 \times 0 + 6/2 \times 0 + 1/2 \times 0 + 2/2 \times 2 + 2/2 \times 1 + 1/2 \times 3$

$= 8/2 + 4/2 + 2/2 + 3/2$

Class 2 centroid for item 1 $= 17/2$

Class2: centroids x item2

Centroids: 3/2, 2/2, 4/2, 6/2, 1/2, 2/2, 2/2, 1/2

item2 : 3, 1, 4, 3, 1, 2, 0, 1

Similarity:

$= 3/2 \times 3 + 2/2 \times 1 + 4/2 \times 4 + 6/2 \times 3 + 1/2 \times 1 + 2/2 \times 2 + 2/2 \times 0 + 1/2 \times 1$

$= 9/2 + 2/2 + 16/2 + 18/2 + 1/2 + 4/2 + 1/2$

item2 = 51/2

Class 2: centroids x item3

Centroids: 3/2, 2/2, 4/2, 6/2, 1/2, 2/2, 2/2, 1/2

item3 : 3, 0, 0, 0, 3, 0, 3, 0

226

Similarity:

$= 3/2 \times 3 + 2/2 \times 0 + 4/2 \times 0 + 6/2 \times 0 + 1/2 \times 3 + 2/2 \times 0 + 2/2 \times 3 + 1/2 \times 0$

$= 9/2 + 3/2 + 6/2 + 0$

Item3 = 18/2

Class2 centriods x Item4 :

Centriods: 3/2, 2/2, 4/2, 6/2, 1/2, 2/2, 2/2, 1/2

Item4 : 0, 1, 0, 3, 0, 0, 2, 0

Similarity:

$= 3/2 \times 0 + 2/2 \times 1 + 4/2 \times 0 + 6/2 \times 3 + 1/2 \times 0 + 2/2 \times 0 + 2/2 \times 2 + 1/2 \times 0$

$= 2/2 + 18/2 + 4/2$

Item4 = 24/2

Class2 Centsiods x Item5 :

Centsiods : 3/2, 2/2, 4/2, 6/2, 1/2, 2/2, 2/2, 1/2

Item5 : 2, 2, 2, 3, 1, 4, 0, 2

Similarity:

$= 3/2 \times 2 + 2/2 \times 2 + 4/2 \times 2 + 6/2 \times 3 + 1/2 \times 1 + 2/2 \times 4 + 2/2 \times 0 + 1/2 \times 2$

$= 6/2 + 4/2 + 8/2 + 18/2 + 1/2 + 8/2 + 2/2$

Item5 = 47/2

Finally, assign the classes as follows:

	class 1	class 2	Assign
Item 1	33/2	17/2	class 1
Item 2	23/2	51/2	class 2
Item 3	30/2	18/2	class 1
Item 4	12/2	24/2	class 2
Item 5	31/2	47/2	class 2

fig: Item clustering with partial clusters.

Note: Assign class for Highest weight.

Ex:

	class 1	class 2	Assign
Item 1	33/2	17/2	class 1

Similarly all. //

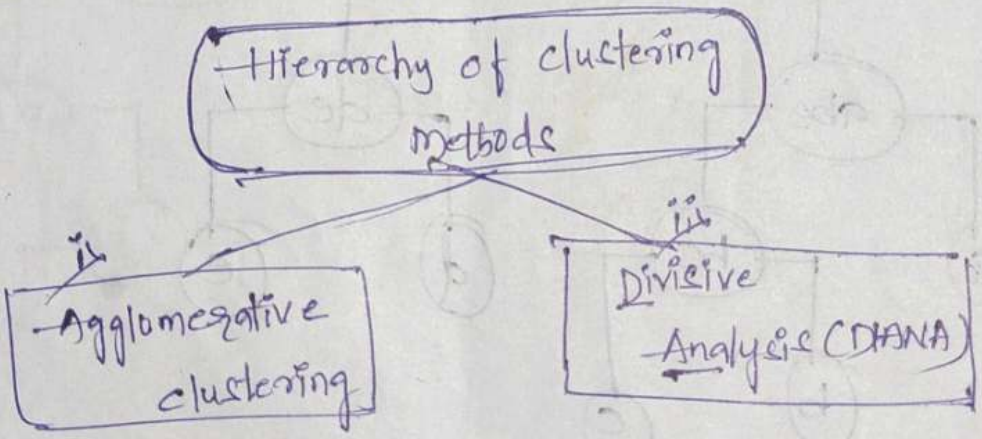
(X) Hierarchy of clusters:-

Hierarchical clustering in Information Retrieval focuses on the area of hierarchical agglomerative clustering methods (HAGM).

Agglomerative:

The term Agglomerative means "the clustering process" starts with unclustered items and it performs pairwise similarity measures to determine the cluster.

Methods of Hierarchy of clustering:-



i) Agglomerative Clustering:-

→ It is a "Bottom-up" approach.

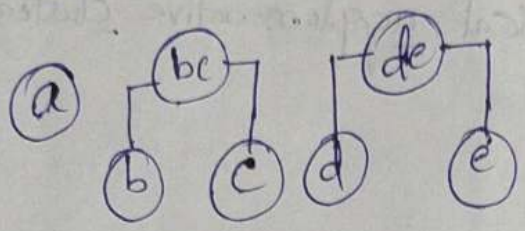
→ It "start" with many "small clusters" and "merge" them together to create "Bigger clusters".

For example :-

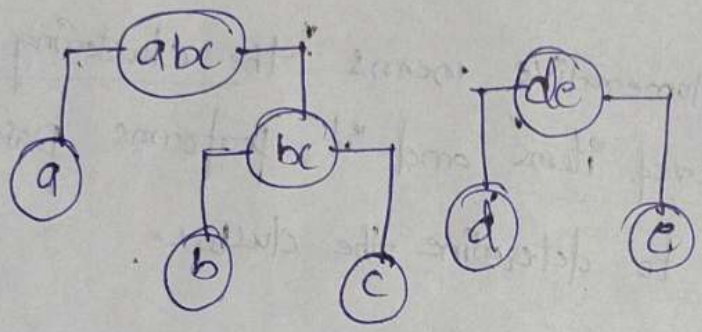
Data set a, b, c, d, e

Step 1 : (a) (b) (c) (d) (e)

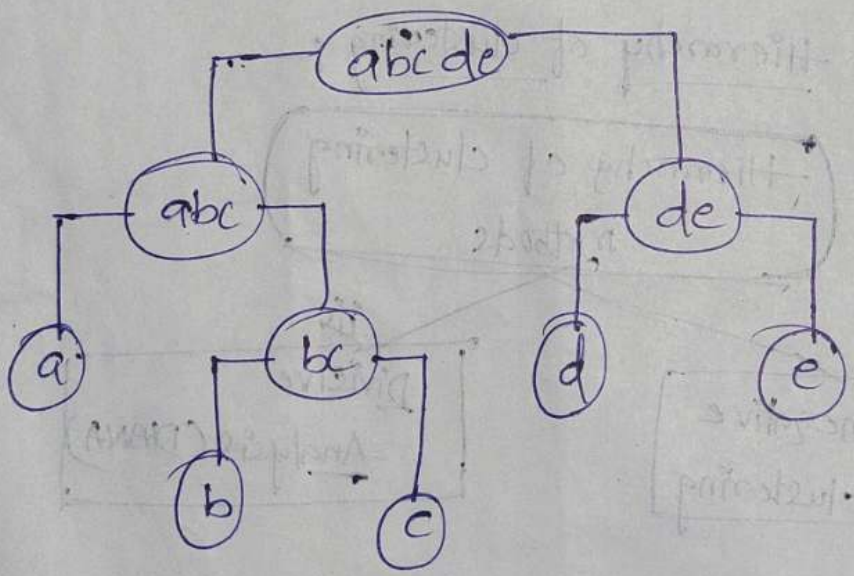
Step 2 :



Step 3 :



Step 4 :



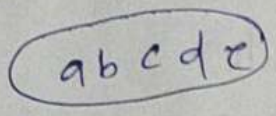
ii) Divisive Analysis (DIANA) :

→ It is a "Top-down" approach.

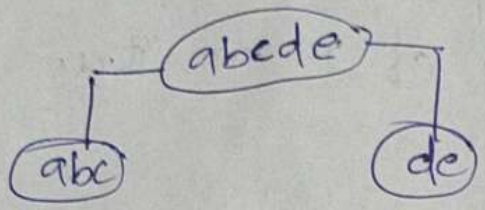
→ It start with a "Single cluster" then break it up into "smaller clusters".

For Example :- Consider Data set {a, b, c, d, e}

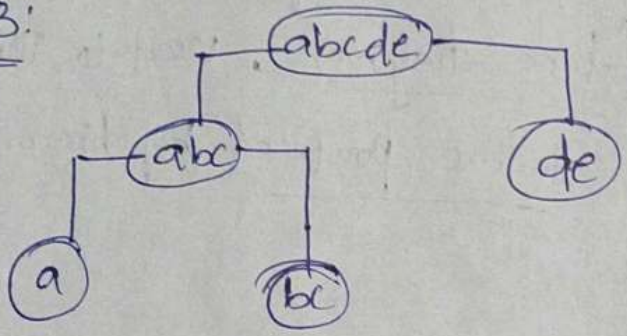
Step (1) :



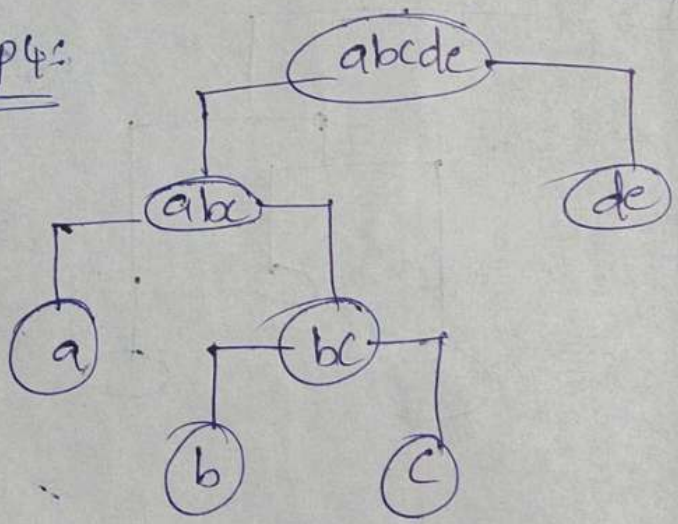
Step 2 :



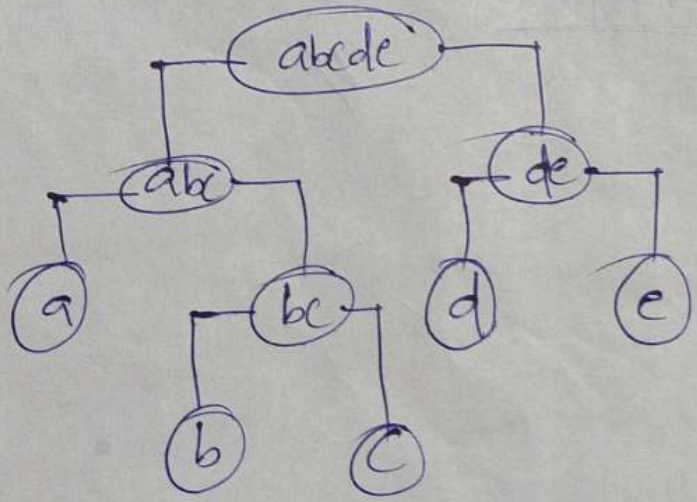
Step 3:



Step 4:



Step 5:



The Objectives of Creating a hierarchy clusters are:

* It Reduce the overhead of search
* It provide a visual representation of the information space.
* It Expand the retrieval of relevant items.

Dendogram :-

A Dendogram is a "tree diagram", used to illustrate the arrangement of the clusters produced by hierarchical clustering.

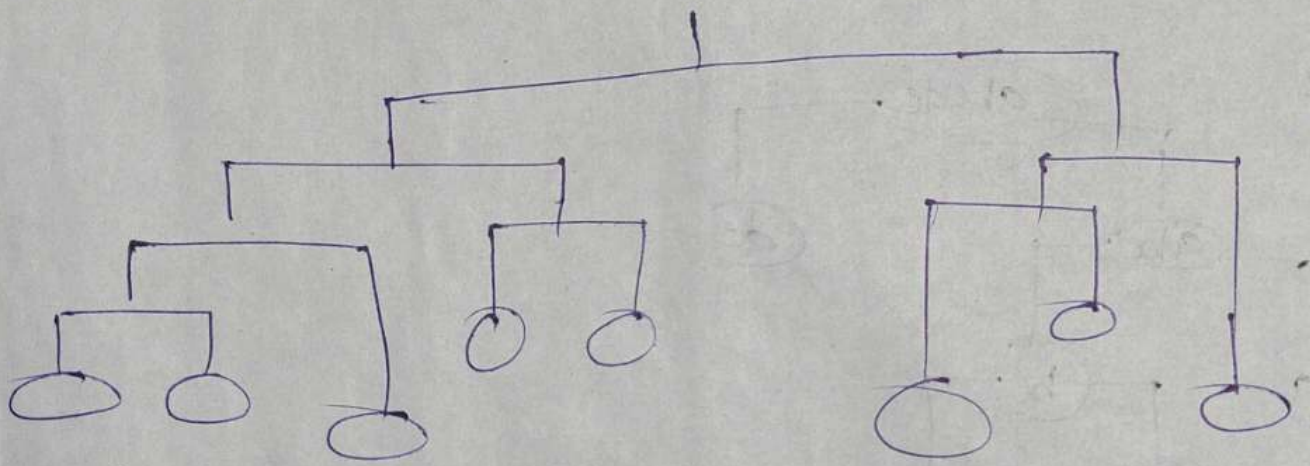


fig: Dendogram