Unit-III

Automatic Indexing

Classes of 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 structures associated with an item.

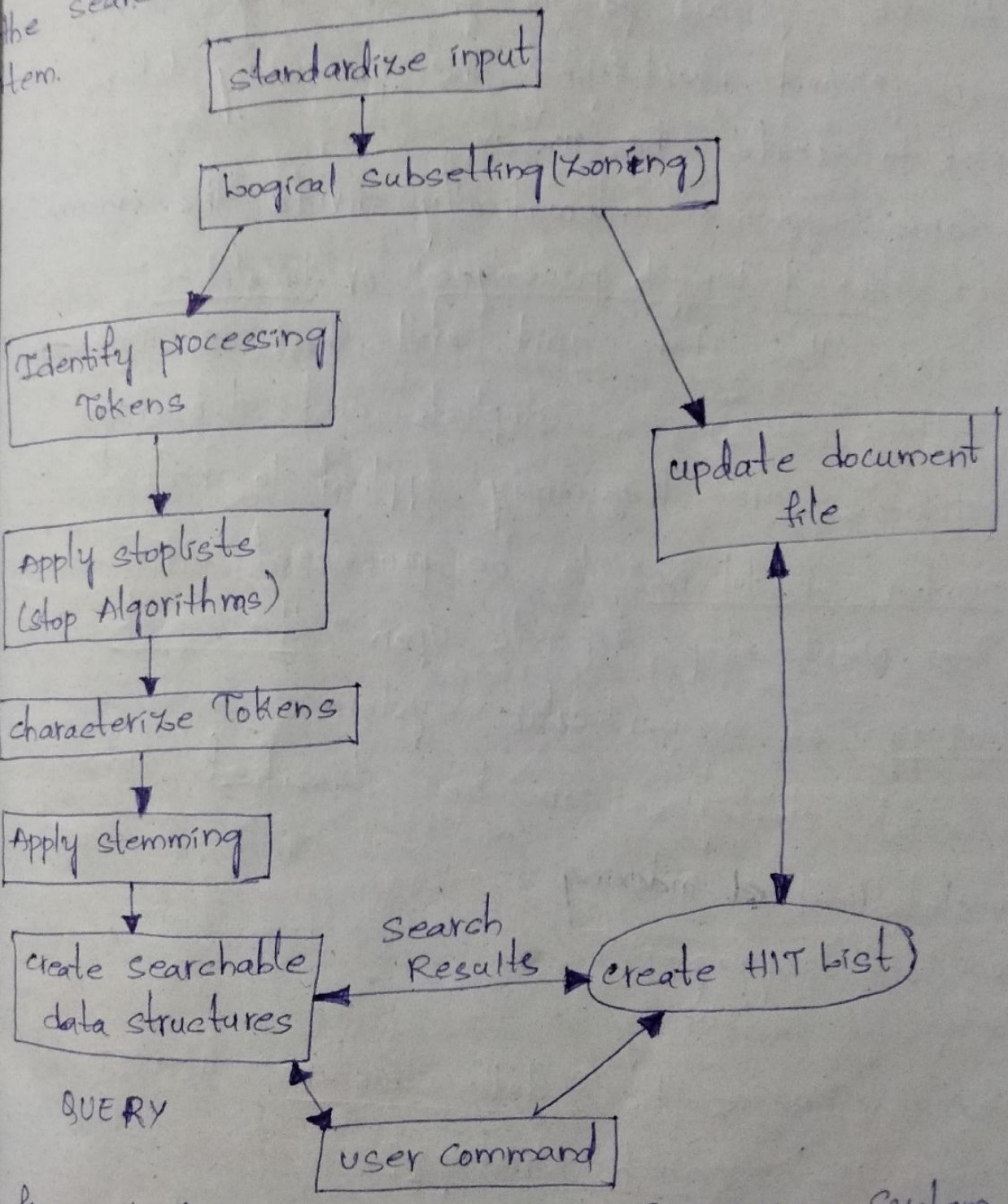


fig.: Data flow in Information Processing System.

From the figure, the left side of the figure including → Identify processing Tokens
→ Apply stop lists
→ characterize Tokens
→ Apply stemming and
→ create searchable datastructure is all the part of Indexing process.

⇒ All the systems go through an Initial stage of "Zoning" and identifying the processing tokens used to create the index.

⇒ Some systems automatically divide the document into "fixed length passages" or "localities", which became the item unit and that is indexed.

⇒ "filters", such as stoplists and stemming algorithms are frequently applied to reduce the "number of tokens" to be processed.

⇒ The Next step depends upon "the search strategy" of a particular system.

⇒ Search strategies can be classified into
* statistical indexing
* natural language
* concept indexing

→ An index is the "data structure" created to support the "search strategy"

Statistical (or) statistical indexing:

→ The statistical approach use the freguency of occurence of "events."

→ "events" are usually related to occorence of processing tokens (words/phrases) within the documents and with in the database.

The words/phrases are the domain of searchable values.

→ The statistics that are applied to the event data are probabilistic,

Bayesian,

vector space & neural

Natural language:

Natural language approaches performs the similar processing token identification

Concept indexing:

The concept indexing uses the words within an item to correlate to concepts discussed in the item.

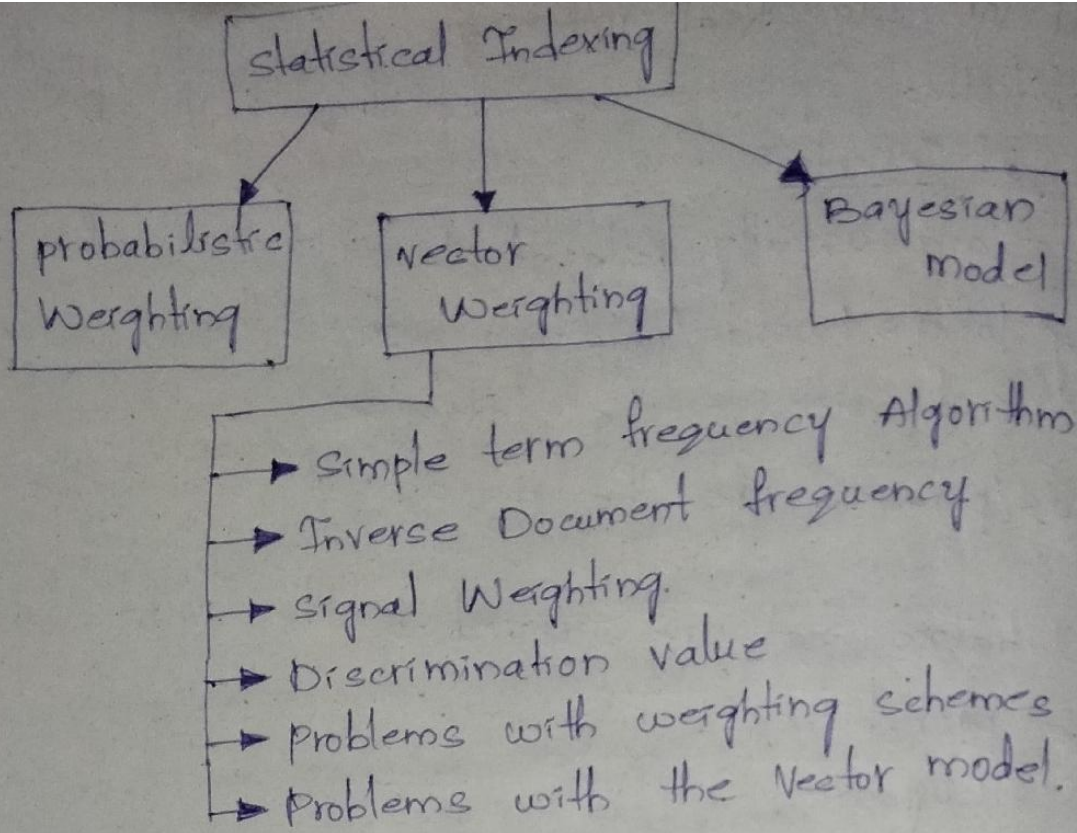
Statistical Indexing.

Statistical Indexing divided into three divisions

→ probabilistic weighting

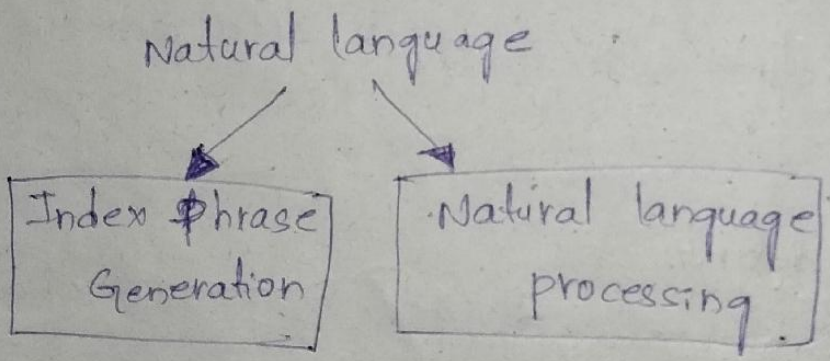
→ vector Weighting

→ Bayesian model.



Natural language processing

*Natural language approach performs the similar processing tokens identification as statistical technique but additionally it performs the varying levels of natural language processing of item.



*Index Phrase Generation: The goal of index phrase Generation is to represent the semantic concepts of an item in the information system that supports to finding the relevant information.

→ It is achieved by cohesion factor.

$COHESION_{K,h} = size\text{-}factor * (Pair\text{-}freq_{K,h} / TOTF_K * TOTF_h)$

K,h are terms.

size-factor ⇒ indicates the length of vocabulary.

$Pair\text{-}freq_{K,h}$ ⇒ indicates the total frequency of co-occurence of paired terms K and h.

Co-occurence may be defined with the help of proximity. This proximity may be word proximity or sentence proximity or adjacency.

$TOTF_K$ ⇒ Total frequency of term K.

$TOTF_h$ ⇒ Total frequency of term h.

Guidelines:

→ Any pair of adjacent non-stop word is a potential phrase

→ Any pair must exist in 25 or more than 25 items

→ phrase weighting uses a modified version of smart system single term algorithm.

→ normalization - It is achieved by dividing the length of single term sub vector.

⇒ Natural language Processing.

→ The Goal of Natural language Processing to use the semantic information in addition to the statistical information. This improves the precision of the search and reduce the

false hits of the user review.

Semantic information:

The Semantic information is extracted as a result of processing ~~token~~ language.

→ The natural language processing can also combine the concepts into higher level concepts sometimes referred as thematic representation.

→ Natural language processing can reduce the errors in determining the phrases.

→ The major advantage of the natural language approach is their ability to produce multiple term phrases to denote a single concept.

for example:

Industrious Intelligent students.

→ If we consider statistical approach, express the sentence as follows:

"Industrious Intelligent"

"Intelligent students."

→ If we consider natural language, express the sentence as follows:

"Industrious students"

"Intdligent students"

"Industrious Intelligent students."

Concept Indexing

→ The concept indexing is a "Statistical Technique."

→ The Goal of concept indexing is to determine the canonical representation of the information.

→ This concept indexing is used to representing the "relevant information" when other techniques are "miss."

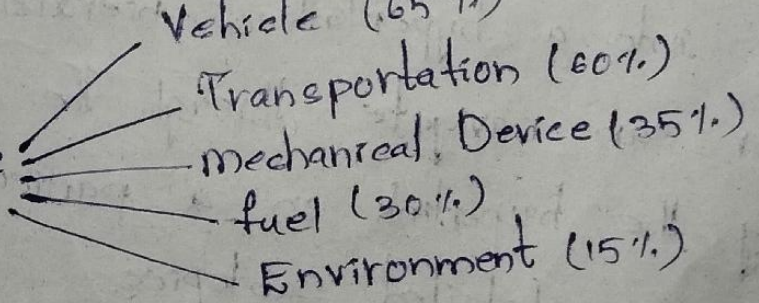
→ Concept indexing start with the number of unbalanced concept classes and the concept classes are created to define the information.

→ In this concept indexing, the single term is going to representing multiple concepts

for e.g.; Term: Automobile.

The term automobile, is related to the concepts such as "Vehicle", "transportation", "mechanical device", "fuel", "Environment".

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



Vector Representation = (65, 60, 35, 30, 15)

Hypertext linkages:

→ Hypertext linkages created "an additional dimension" to the Information Retrieval.

→ In this traditional items can be viewed as two dimensional constructs.

→ The text of the item is one-dimension representation of information.

→ Imbedded references are a logical second-dimension.

→ The Imbedding of the linkage allows the user to go immediately to the linked items for additional information.

→ The following 3 classes of mechanisms help to find the information on the internet.

They are * Manually generated indexes
* automatically generated indexes
* Web crawlers (Intelligent agents).

Manually generate Indexes:

The information sources (Home pages) are indexed manually into a hyperlinked hierarchy.

The user can navigate through the hierarchy by expanding the hyperlink on a particular topic to see the more detailed subtopics.

eg: YAHOO (http://www.yahoo.com)
Google

Automatically Generated Indexes:

The particular sites are automatically go out to other Internet sites and return the text at the sites for automatic indexing.

e.g.: Lycos (http://www.lycos.com) &
Altavista (http://www.altavista.digital.com)

Web crawlers:

The Web crawlers and Intelligent agents are tools that allow a user to define items of interest and they automatically go to the sites on the internet for searching the desired information.

Thesaurus Generation

→ The word Theasaurus comes from Greek term "Thesauros" and the meaning is a "store house" (or) "treasary of words."

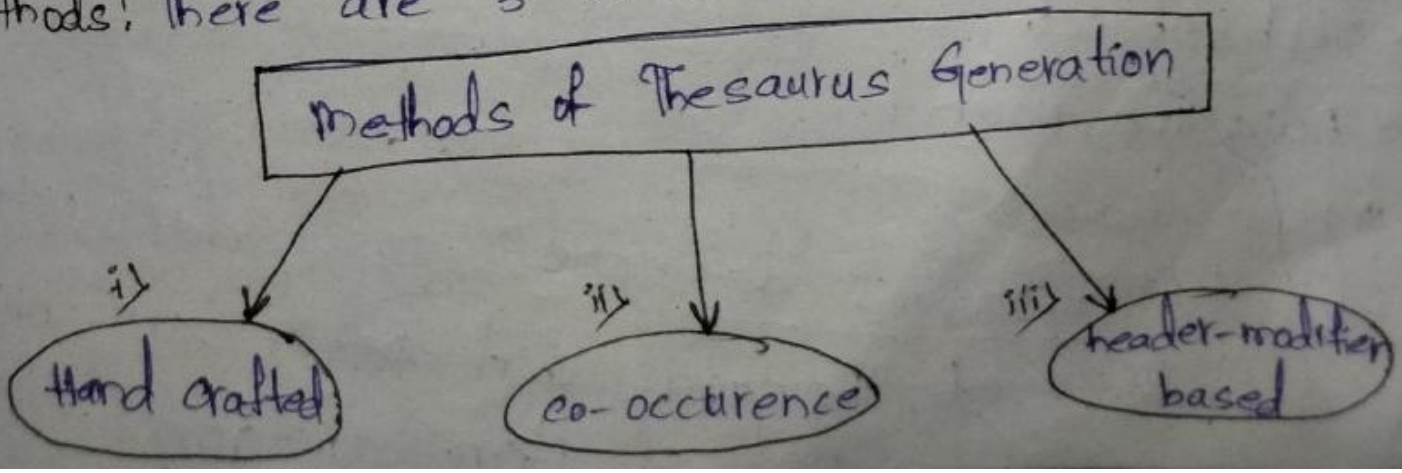
→ Thesaurus was first Introduced by "Peter Mark Roget" i.e., Thesaurus of "English words and phrases" (or)

→ A Thesaurus is a "book of words" that shows the relationship among the words it contains.

⇒ "Manual Generation" of clusters usually focus on generating a "thesaurus."

⇒ Automatically generated thesauri contains classes that reflect the use of words in the carpora.

Methods: There are 3 basic words.



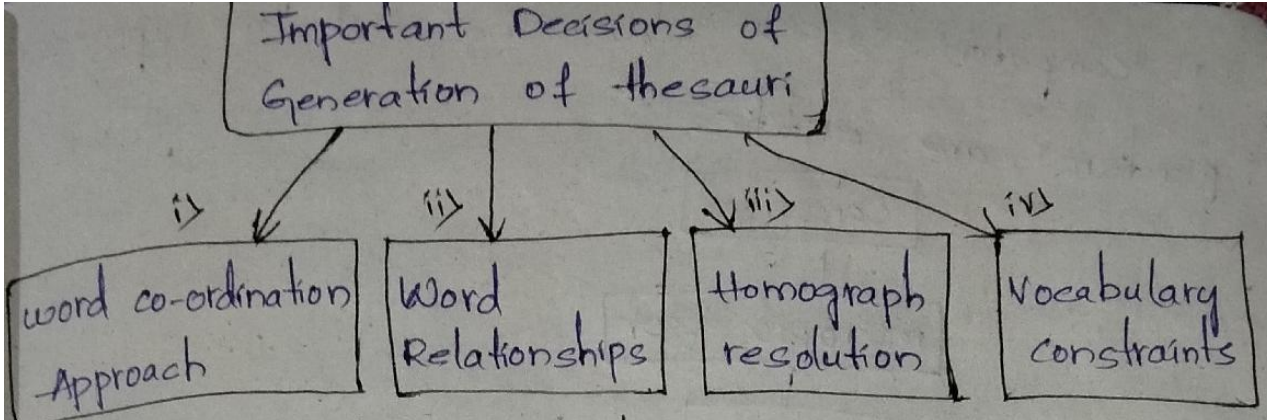
i) Hand crafted: → Manual thesauri
→ General thesauri

* The "Manually made thesauri" only helps in Query Expansion."
* The "general thesaurus (e.g.: Word Next) does not help as much because "same word" have different meanings.

ii) Co-occurence:
→ "Techniques" for co-occurence creation of thesauri are: i.e., Each "noun" has a set of "verbs", "adjectives" The nouns that it co-occurs with "a mutual information"
* mutual Information:
The "Similarity between the words" is calculated by mutual information to classify the terms.

iii) Header - Modifier Based:
The Header-modifier Based thesauri is based upon the linguistic relationships.
↓
scientific study of language or its structures
→ The linguistic parsing of document discovers the following syntatical structures:
* subject - verb
* verb - object
* Adjective - Noun &
* Noun - Noun.

Additional "Important decisions" associated with the Generation of thesauri that ~~occur~~ are not part of item clustering



i) Word co-ordination Approach:

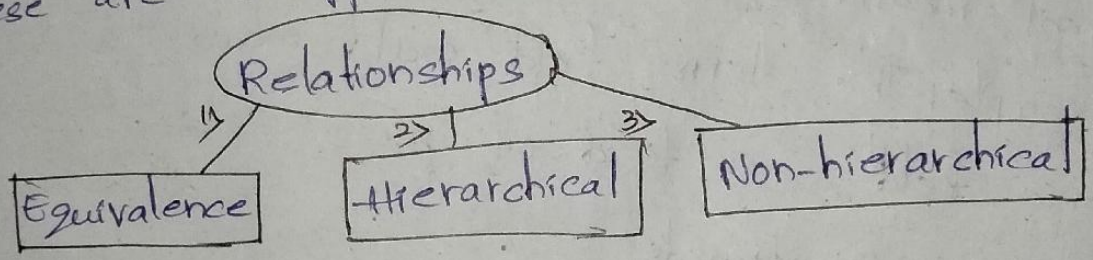
In this approach, the "phrases" and "individual terms" are to be clustered.

Hint: see pre-coordination & post-coordination.

ii) Word Relationships:

The generation thesaurus includes a "human interface" a variety of relationships" between words are possible.

These are 3 types of relationships.



1) Equivalence:

The equivalence relationships are most "common relationships" which represent the "Synonyms".

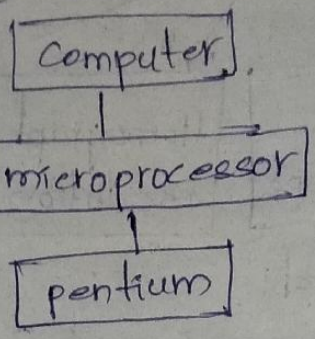
eg: 1. The term "photograph" and "print" may be defined as Synonyms.

2. The words "Genius" and "moron" may be synonyms in a class called "intellectual capability".

2) Hierarchical:

A very common Technique is Hierarchical relationships where the classname is a general term and entities are examples of general term.

e.g.: "computer" is a "classname" and "microprocessor" "pentium" are entities.



3) Non-hierarchical

The non-hierarchical cover other types of relationships such as "object-attribute" contain "employee" and "jobtitle".

iii) Homograph Resolution:

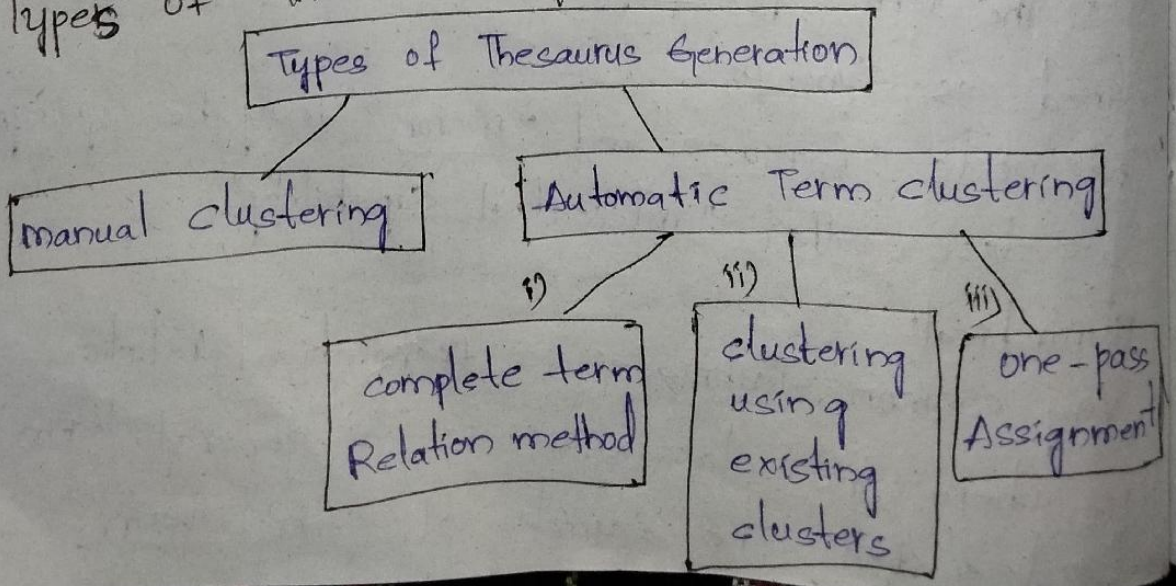
Homograph is a word that has "multiple, completely different meanings".

for example: The term "field" could mean "a electronic field", "a field of grass", etc..

iv) Vocabulary Constraints:

This includes "guidelines" on the normalization and specificity of the vocabulary.

Types of Thesaurus Generation:



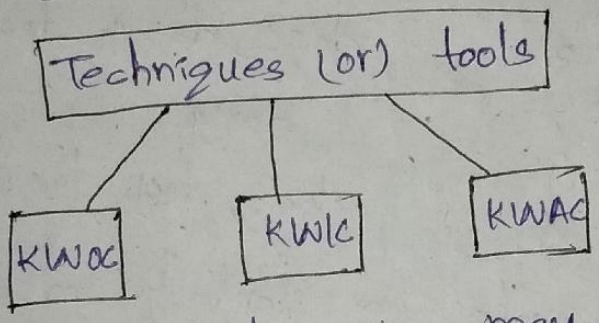
Manual clustering:

→ The Manual clustering process follows the steps in the generation of a theasaurus.

→ The first step is to determine the domain for the clustering, and identify attributes and determine strength of relationship between the attributes.

→ Both clusterings uses the same techniques.

Tools ~~are~~/ techniques used in Manual clustering:



Here, KWOC, KWIC and KWAC may help in determining useful words.

Here

*KWOC → A KeyWord Out of Context

It is another name for a "concordance".

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

*KWIC → KeyWord In Context

→ It displays a "possible term" in its phrase Context

→ It is structured to identify location of the term.

*KWAC → KeyWord And Context

→ It displays the keywords followed by their context.

The following statement/phrase representing the KWOC, KWIC and KWAC.

for example: "Computer design contains memory chips"

Note: The phrase is assumed from document.

eg: "Computer design contain memory chips"

KWOC

TERM	PREQ	ITEM Ids
chips	2	doc2, doc4
computer	3	doc1, doc4, doc10
Design	1	doc4
memory	4	doc3, doc4, doc8, doc12

→ Arrange the words in sorted order.

→ After arranging the terms check the frequency. Here frequency is indicating how many number of times the particular "Term" appear in documents.

i.e. chips "2" times appear in doc2 & doc4

KWIC

In the given sentence {"Computer design contains memory chips"}.

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

→ The words/terms, the grouping is performed like this If I want to search for a word "chips/."

The Complete sentence will be extracted like "Computer design Contains memory" and the chips will be "excluded" from the Sentence.

KWIC

chips/	Computer Design Contains Memory
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 Computer Sentence will be extracted."

i.e.

chips	Computer Design Contains Memory chips
Computer	Computer Design Contain Memory chips.
design	Computer Design Contains Memory chips.
Memory	Computer Design Contains memory chips.

fig: Example of KWOC, KWIC & KWAC

2) Automatic Term clustering.

→ Terms are automatically clustered based on their frequency.

There are many techniques for the automatic generation of term clusters to create "Statistical ~~thesauri~~" thesauri"

→ The Automatic Term clustering can be performed

by following 3 ways.

They are

1. Complete term Relation method
2. clustering Using Existing clusters
3. One pass Assignments.

1) Complete term Relation method:

→ In the complete term relation method, the Similarity between every term pair is calculated as basis for determining the clusters.

→ The ~~exact~~ easiest way to understand this approach is to consider "the vector model."

→ The Vector Model is represented by a "matrix", where, the "rows" are "individual items" and the "columns" are "Unique Words" in the items.

↓
means processing tokens.

→ The values in the matrix represent how strongly that particular word represent concepts in the item.

The following figure/example contains a database with "5 items" and "8 terms."

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
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

fig: Vector Example

→ A similarity measure is required.
→ A measure calculate the similarity between two terms

i.e.

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

where 'K' is summed across the set of all the items.

⇒ Based on the Vector matrix / Example we have to find out

* Term - Term matrix &
* Term - Relationship matrix
* Network Diagram of Term Similarities.

⇒ from figure, we can say that the "term 1" appear in item 1 is '0' and Item 2 is '3', item 3 is '3' and item 4 is '0' and item 5 is '2'.

Term to Term table:

Here we want to find out the relation between one term and other term means how term 1 is related to term1, Term2, ---- Term8.

Similary Term 2 is related to Term1 - - - - Term8
Term 3 is related to Term1 - - - - Term8
Term 4 is related to Term1 - - - Term8
⋮
Term8 is related to Term1 - - - - Term8

There are no values on the dragonal that represents the auto correlation of word to itself.

Term 1:

Term1, Term2 = 0*4 + 3*1 + 3*0 + 0*1 + 2*2
= 0+3+0+0+4 = 7

Term1, Term3 = 0*0 + 3*4 + 3*0 + 0*0 + 2*2
= 0+12+0+0+4 = 16

Term1, Term4 = 0*0 + 3*3 + 3*0 + 0*3 + 2*3
= 0+9+0+0+6 = 15

Term1, Term5 = 0*0 + 3*1 + 3*3 + 0*0 + 2*1
= 0+3+9+0+2 = 14

Term1, Term6 = 0*2 + 3*2 + 3*0 + 0*0 + 2*4
= 0+6+0+0+8 = 14

Term1, Term7 = 0*1 + 3*0 + 3*3 + 0*2 + 2*0
= 0+0+9+0+0 = 9

Term1, Term8 = 0*3 + 3*1 + 3*0 + 0*0 + 2*2
= 0+3+0+0+4 = 7

similarly, we should find for Term2, Term3, Term4, Term5, Term6, Term7 and Term8.

Then we get the table as shown below

	Term1	Term2	Term3	Term4	Term5	Term6	Term 7	Term 8
Term1	-	7	16	15	14	4	9	7
Term2	7	-	8	12	3	18	6	17
Term3	16	8	-	18	6	16	0	8
Term4	15	12	18	-	6	18	6	9
Term5	14	3	6	6	-	6	9	3
Term6	4	18	16	18	6	-	2	16
Term7	9	6	0	6	9	2	-	3
Term8	7	17	8	9	3	16	3	-

fig.: Term-Term matrix

→ Next Step is to select a "threshold" that determines if two terms are considered similar enough to each other to be in the same class.

Take Threshold value as '10' or i.e. > 10

Thus, two terms are considered similar if the similarity value between them is 10 or greater than 10.

This produce a new binary matrix called "Term Relationship matrix".

Here, >10 values indicated with '1' and <10 values are indicated with '0'.

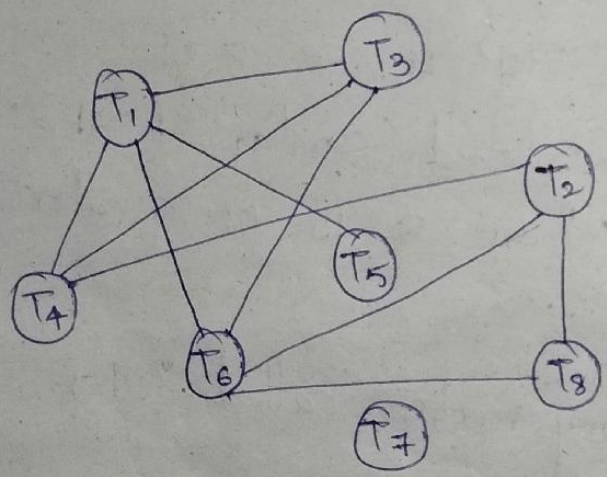
ie. ~~Term~~

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Term 1	–	0	1	1	1	1	0	0
Term 2	0	–	0	1	0	1	0	1
Term 3	1	0	–	1	0	1	0	0
Term 4	1	1	1	–	0	1	0	0
Term 5	1	0	0	0	–	0	0	0
Term 6	1	1	1	1	0	–	0	1
Term 7	0	0	0	0	0	0	–	0
Term 8	0	1	0	0	0	1	0	–

fig.: Relationship matrix.

Next step: Network Diagram.

~~T1 + T3~~ Terms with 1 are connected and Terms with 0 are not connected.



T_1 : $\{T_3, T_4, T_5, T_6\}$

T_2 : $\{T_4, T_6, T_8\}$

T_3 : $\{T_1, T_4, T_6\}$

T_4 : $\{T_1, T_2, T_3, T_6\}$

T_5 : T_4

T_6 : $\{T_1, T_2, T_3, T_4, T_8\}$

~~T_7~~ : –

T_8 : $\{T_2, T_6\}$

Two classes

class 1 ($T_1, T_2, T_3, T_4, T_5, T_6, T_8$) → There is a connectivity with other terms

class 2 (T_7) → There is no connectivity with other terms

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 illustrates below.

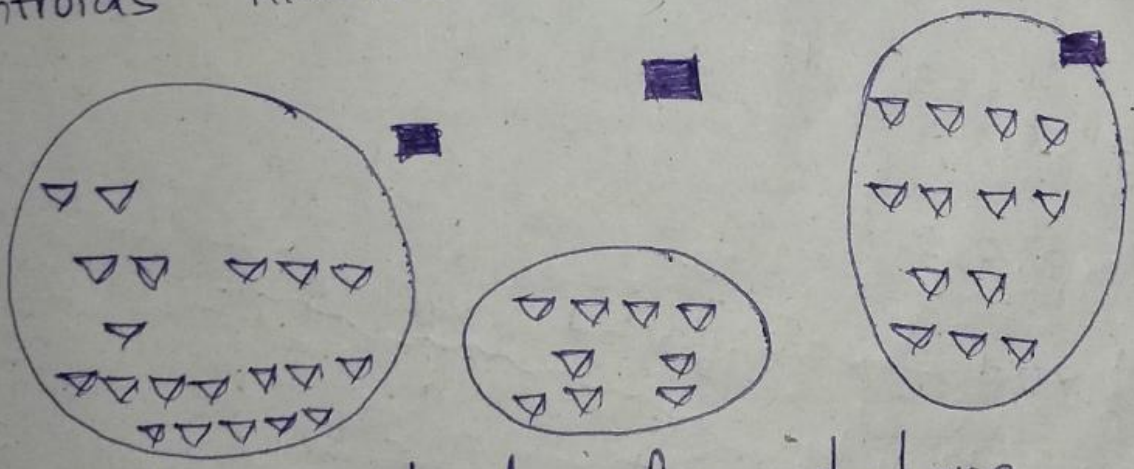


fig.: initial centroids for clusters

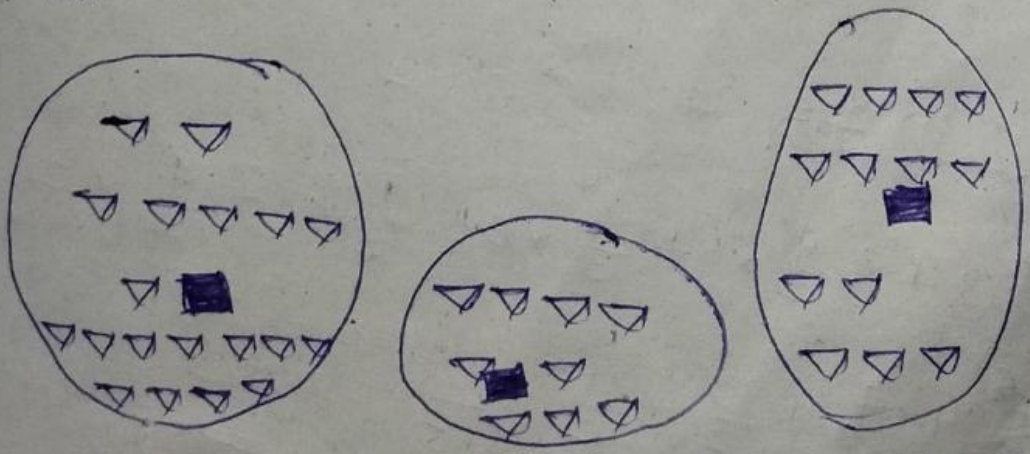


fig.: Centroids after Reassigning Terms

A centroid is viewed in physics as the center of mass of a set of objects
for a Inital Iteration
eg:

class 1 = (Term 1, Term 2)

class 2 = (Term 3, Term 4)

class 3 = (Term 5, Term 6)

This would produce the following centroids for each d...

Initial Centroids:

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

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

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

class 2 = $\left(\frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}\right)$

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

class 3 = $\left(\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}\right)$

class 1 centroids for term 1, term 2, term 3 ----

term 8 (calculate similarity)

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

<u>term 1</u> = 0, 3, 3, 0, 2

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

$= \frac{0}{2} + \frac{12}{2} + \frac{9}{2} + \frac{0}{2} + \frac{8}{2} = \frac{29}{2}$

~~class 2 =~~ <u>term 2</u> = 4, 1, 0, 1, 2

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

$= \frac{16}{2} + \frac{4}{2} + \frac{0}{2} + \frac{1}{2} + \frac{8}{2} = \frac{29}{2}$

term 3 = 0, 4, 0, 0, 2

similarity = $\frac{4}{2}\times0 + \frac{4}{2}\times4 + \frac{3}{2}\times0 + \frac{1}{2}\times0 + \frac{4}{2}\times2$

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

$= \frac{24}{2}$

term 4 = 0, 3, 0, 3, 3

similarity = $\frac{4}{2}\times0 + \frac{4}{2}\times3 + \frac{3}{2}\times0 + \frac{1}{2}\times3 + \frac{4}{2}\times3$

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

$= \frac{27}{2}$

term 5 = 0, 1, 3, 0, 1

similarity = $\frac{4}{2}\times0 + \frac{4}{2}\times1 + \frac{3}{2}\times3 + \frac{1}{2}\times0 + \frac{4}{2}\times1$

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

$= \frac{17}{2}$

term 6 = 2, 2, 0, 0, 4

similarity = $\frac{4}{2}\times2 + \frac{4}{2}\times2 + \frac{3}{2}\times0 + \frac{1}{2}\times0 + \frac{4}{2}\times4$

$= \frac{8}{2} + \frac{8}{2} + \frac{0}{2} + \frac{0}{2} + \frac{16}{2} = \frac{32}{2}$

term 7 = 1, 0, 3, 2, 0

similarity = $\frac{4}{2}\times1 + \frac{4}{2}\times0 + \frac{3}{2}\times3 + \frac{1}{2}\times2 + \frac{4}{2}\times0$

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

$= \frac{15}{2}$

term 8 = 3, 1, 0, 0, 2

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

$= \frac{12}{2} + \frac{4}{2} + \frac{0}{2} + \frac{0}{2} + \frac{8}{2}$

$= \frac{24}{2}$

* class 2 centroids for term 1, ---- term 8 (calculate similarity)

class 2 = $\left(\frac{0}{2}, \frac{7}{2}, \frac{0}{2}, \frac{3}{2}, \frac{5}{2}\right)$

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

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

$= \frac{0}{2} + \frac{21}{2} + \frac{0}{2} + \frac{0}{2} + \frac{10}{2} = \frac{31}{2}$

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

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

$= \frac{0}{2} + \frac{7}{2} + \frac{0}{2} + \frac{3}{2} + \frac{10}{2} = \frac{20}{2}$

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

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

$= \frac{0}{2} + \frac{28}{2} + \frac{0}{2} + \frac{0}{2} + \frac{10}{2} = \frac{38}{2}$

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

similarity = $\frac{0}{2}\times0 + \frac{7}{2}\times3 + \frac{0}{2}\times0 + \frac{3}{2}\times3 + \frac{5}{2}\times3$

$= \frac{0}{2} + \frac{21}{2} + \frac{0}{2} + \frac{9}{2} + \frac{15}{2} = \frac{45}{2}$

* Term 5 = 0, 1, 3, 0, 1

similarity = $\frac{0}{2}\times0 + \frac{7}{2}\times1 + \frac{0}{2}\times3 + \frac{3}{2}\times0 + \frac{5}{2}\times1$

$= \frac{0}{2} + \frac{7}{2} + \frac{0}{2} + \frac{0}{2} + \frac{5}{2} = \frac{12}{2}$

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

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

$= \frac{0}{2} + \frac{14}{2} + \frac{0}{2} + \frac{0}{2} + \frac{20}{2} = \frac{34}{2}$

* Term 7 = 1, 0, 3, 2, 0

similarity = $\frac{0}{2} \times 1 + \frac{7}{2} \times 0 + \frac{8}{2} \times 3 + \frac{3}{2} \times 2 + \frac{5}{2} \times 0$

$= \frac{0}{2} + \frac{0}{2} + \frac{0}{2} + \frac{6}{2} + 0 = \frac{6}{2}$

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

similarity = $\frac{0}{2} \times 3 + \frac{7}{2} \times 1 + \frac{0}{2} \times 0 + \frac{3}{2} \times 0 + \frac{5}{2} \times 2$

$= \frac{0}{2} + \frac{7}{2} + \frac{0}{2} + \frac{8}{2} + \frac{10}{2} = \frac{17}{2}$

* class 3 centrords for term 1 .... term 8
(calculate similarity)

class 3 = $\left(\frac{2}{2}, \frac{3}{2}, \frac{3}{2}, \frac{0}{2}, \frac{5}{2}\right)$

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

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

$= \frac{0}{2} + \frac{9}{2} + \frac{9}{2} + \frac{0}{2} + \frac{10}{2} = \frac{28}{2}$

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

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

$= \frac{8}{2} + \frac{3}{2} + \frac{0}{2} + \frac{0}{2} + \frac{10}{2} = \frac{21}{2}$

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

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

$= \frac{0}{2} + \frac{12}{2} + \frac{0}{2} + \frac{0}{2} + \frac{10}{2} = \frac{22}{2}$

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

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

$= \frac{0}{2} + \frac{9}{2} + \frac{0}{2} + \frac{0}{2} + \frac{15}{2}$

$= \frac{24}{2}$

* Term 5 = 0, 1, 3, 0, 1

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

$= \frac{0}{2} + \frac{3}{2} + \frac{9}{2} + \frac{0}{2} + \frac{5}{2} = \frac{17}{2}$

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

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

$= \frac{4}{2} + \frac{6}{2} + \frac{0}{2} + \frac{0}{2} + \frac{20}{2} = \frac{30}{2}$

* Term 7 = 1, 0, 3, 2, 0

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

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

$= \frac{11}{2}$

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

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

$= \frac{6}{2} + \frac{3}{2} + \frac{0}{2} + \frac{0}{2} + \frac{10}{2}$

$= \frac{19}{2}$

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
class 1	29/2	29/2	24/2	27/2	17/2	32/2	15/2	24/2
class 2	31/2	20/2	38/2	45/2	12/2	34/2	16/2	17/2
class 3	28/2	21/2	22/2	24/2	17/2	30/2	11/2	19/2
Assign	class 2	class 1	class 2	class 2	class 3	class 2	class 1	class

fig: Iterated class Assignment

Next Iteration

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 above table.

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

class 1 = $\left(\frac{8}{3}, \frac{2}{3}, \frac{3}{3}, \frac{3}{3}, \frac{4}{3}\right)$

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

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

class 3 = $\left(\frac{0}{1}, \frac{1}{1}, \frac{3}{1}, \frac{0}{1}, \frac{1}{1}\right)$

*class 1 centroids for term 1, term 2, ... term 8

(calculate similarity)

class 1 = $\left(\frac{8}{3}, \frac{2}{3}, \frac{3}{3}, \frac{3}{3}, \frac{4}{3}\right)$

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

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

$= \frac{0}{3} + \frac{6}{3} + \frac{9}{3} + \frac{0}{3} + \frac{8}{3} = \frac{23}{3}$

term 2 = 4, 1, 0, 1, 2

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

$= \frac{32}{3} + \frac{2}{3} + \frac{0}{3} + \frac{3}{3} + \frac{8}{3} = \frac{45}{3}$

term 3 = 0, 4, 0, 0, 2

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

similarity = $\frac{0}{3} + \frac{8}{3} + \frac{0}{3} + \frac{0}{3} + \frac{8}{3} = \frac{16}{3}$

term 4 = 0, 3, 0, 3, 3

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

$= \frac{0}{3} + \frac{6}{3} + \frac{0}{3} + \frac{9}{3} + \frac{12}{3} = \frac{27}{3}$

term 5 = 0, 1, 3, 0, 1

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

$= \frac{0}{3} + \frac{2}{3} + \frac{9}{3} + \frac{0}{3} + \frac{4}{3} = \frac{15}{3}$

term 6 = 2, 2, 0, 0, 4

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

$= \frac{16}{3} + \frac{4}{3} + \frac{0}{3} + \frac{0}{3} + \frac{16}{3} = \frac{36}{3}$

term 7 = 1, 0, 3, 2, 0

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

$= \frac{8}{3} + \frac{0}{3} + \frac{9}{3} + \frac{6}{3} + \frac{0}{3} = \frac{23}{3}$

term 8 = 3, 1, 0, 0, 2

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

$= \frac{24}{3} + \frac{2}{3} + \frac{0}{3} + \frac{0}{3} + \frac{8}{3} = \frac{34}{3}$

* class 2 centroids for term 1, term 2, ... term 8 (calculate similarity)

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

~~similarity~~ * term 1 = 0, 3, 3, 0, 2

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

$= \frac{0}{4} + \frac{36}{4} + \frac{9}{4} + \frac{0}{4} + \frac{22}{4} = \frac{67}{4}$

* term 2 = 4, 1, 0, 1, 2

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

$= \frac{8}{4} + \frac{12}{4} + \frac{0}{4} + \frac{3}{4} + \frac{22}{4} = \frac{45}{4}$

* term 3 = 0, 4, 0, 0, 2

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

$= \frac{0}{4} + \frac{48}{4} + \frac{0}{4} + \frac{0}{4} + \frac{22}{4} = \frac{70}{4}$

* term 4 = 0, 3, 0, 3, 3

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

$= \frac{0}{4} + \frac{36}{4} + \frac{0}{4} + \frac{9}{4} + \frac{33}{4} = \frac{78}{4}$

* term 5 = 0, 1, 3, 0, 1

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

$= \frac{0}{4} + \frac{12}{4} + \frac{9}{4} + \frac{0}{4} + \frac{11}{4} = \frac{32}{4}$

* term 6 = 2, 2, 0, 0, 4

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

$= \frac{4}{4} + \frac{24}{4} + \frac{0}{4} + \frac{0}{4} + \frac{44}{4} = \frac{72}{4}$

* term 7 = 1, 0, 3, 2, 0

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

$= \frac{2}{4} + \frac{0}{4} + \frac{9}{4} + \frac{6}{4} + \frac{0}{4} = \frac{17}{4}$

* term 8 = 3, 1, 0, 0, 2

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

$= \frac{6}{4} + \frac{12}{4} + \frac{0}{4} + \frac{0}{4} + \frac{22}{4}$

$= \frac{50}{4}$

class 3 centroids Term 1, Term 2, ---- Term 8

(calculate similarity)

class 3 = $\left(\frac{0}{1}, \frac{1}{1}, \frac{3}{1}, \frac{0}{1}, \frac{1}{1}\right)$

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

Similarity = $\frac{0}{1}\times0 + \frac{1}{1}\times3 + \frac{3}{1}\times3 + \frac{0}{1}\times0 + \frac{1}{1}\times2$

$= \frac{0}{1} + \frac{3}{1} + \frac{9}{1} + \frac{0}{1} + \frac{2}{1} = \frac{14}{1}$

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

Similarity = $\frac{0}{1}\times4 + \frac{1}{1}\times1 + \frac{3}{1}\times0 + \frac{0}{1}\times1 + \frac{1}{1}\times2$

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

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

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

$= \frac{0}{1} + \frac{4}{1} + \frac{0}{1} + \frac{0}{1} + \frac{2}{1} = \frac{6}{1}$

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

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

$= \frac{0}{1} + \frac{3}{1} + \frac{0}{1} + \frac{0}{1} + \frac{3}{1} = \frac{6}{1}$

*Term 5 = 0, 1, 3, 0, 1

Similarity = $\frac{0}{1}\times0 + \frac{1}{1}\times1 + \frac{3}{1}\times3 + \frac{0}{1}\times0 + \frac{1}{1}\times1$

$= \frac{0}{1} + \frac{1}{1} + \frac{9}{1} + \frac{0}{1} + \frac{1}{1} = \frac{11}{1}$

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

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

$= \frac{0}{1} + \frac{2}{1} + \frac{0}{1} + \frac{0}{1} + \frac{4}{1} = \frac{6}{1}$

*Term 7 = 1, 0, 3, 2, 0

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

$= \frac{0}{1} * \frac{0}{1} + \frac{9}{1} + \frac{0}{1} + \frac{0}{1} = \frac{9}{1}$

Terms = 3, 1, 0, 0, 2

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

$= \frac{0}{1} * \frac{1}{1} + \frac{0}{1} + \frac{0}{1} + \frac{2}{1} = \frac{3}{1}$

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
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	50/4
class 3	14/1	3/1	6/1	6/1	11/1	6/1	9/1	3/1

fig.: New centroids & cluster Assignments.

3) One pass Assignments:

This Technique has the minimum overhead in that only one pass of all of the items 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.

⇒ This process continues until all items are assigned to classes.

ie. class 1 = Term 1, Term 3, Term 4

class 2 = Term 2, Term 6, Term 8

class 3 = Term 5

class 4 = Term 7

Note: The centroid values used during the one pass process:

class 1 (Term 1, Term 3) *

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

class 3 (Term 2, Term 6)

~~clac~~ calculate Centroids:

class 1 (Term 1, Term 3) = $\frac{0+0}{2}, \frac{3+4}{2}, \frac{3+0}{2}, \frac{0+0}{2}, \frac{2+2}{2}$

$= \frac{0}{2}, \frac{7}{2}, \frac{3}{2}, \frac{0}{2}, \frac{4}{2}$

class 2 (Term 1, Term 3, Term 4) = $\frac{0+0+0}{3}, \frac{3+4+3}{3},$

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

$= \frac{0}{3}, \frac{10}{3}, \frac{3}{3}, \frac{4}{3}, \frac{7}{3}$

class 3 (Term 2, Term 6) = ~~$\frac{4+0}{2}, \frac{1+3}{2}, \frac{0+0}{2}, \frac{1+3}{2}, \frac{2+3}{2}$~~

~~$= \frac{4}{2}, \frac{4}{2}, \frac{0}{2}, \frac{4}{2}, \frac{5}{2}$~~

$= \frac{4+2}{2}, \frac{1+2}{2}, \frac{0+0}{2}, \frac{1+0}{2}, \frac{2+4}{2} = \frac{6}{2}, \frac{3}{2}, \frac{0}{2}, \frac{1}{2}, \frac{6}{2}$

Item clustering

clustering of item / Item clustering is very similar to term clustering for the generation of thesauri

→ The Techniques described for the clustering also applied to item clustering

→ In this concept similarity beteween the documents is based upon "two items".

i.e. terms in common versus terms with items in common.

∴ The similarity function is performed between rows of the item matrix.

Similarity equation:

$sim(Item_i, Item_j) = \sum (Term_{i,k})(Term_{j,k})$

i.e. The above equation indicates "set of items and their terms"

Here 'k' goes from 1 to 8 for the eight terms.

In this item clustering

* An Item-Item matrix is created.
* The Item Relationship matrix is performed by using a threshold of 10.

finding similarity between items.

Item 1

→ (Item 1, Item 2)

⇒ 0×3 + 4×1 + 0×4 + 0×3 + 0×1 + 2×2 + 1×0

⇒ 0 + 4 + 0 + 0 + 0 + 4 + 0 = 11

→ (Item 1, Item 3)

⇒ 0×3 + 4×0 + 0×0 + 0×0 + 0×3 + 2×0 + 1×3 + 3×0

⇒ 0 + 0 + 0 + 0 + 0 + 0 + 3 + 0 = 3

→ (Item 1, Item 4)

⇒ 0×0 + 4×1 + 0×0 + 0×3 + 0×0 + 2×0 + 1×2 + 3×0

⇒ 0 + 4 + 0 + 0 + 0 + 0 + 2 + 0

⇒ 4 + 2 = 6

→ (Item 1, Item 5)

⇒ 0×2 + 4×2 + 0×2 + 0×3 + 0×1 + 2×4 + 1×0 + 3×2

⇒ 0 + 8 + 0 + 0 + 0 + 8 + 0 + 6 = 22

Item 2

→ (Item 2, Item 1)

⇒ 3×0 + 1×4 + 4×0 + 3×0 + 1×0 + 2×2 + 0×1 + 1×3

⇒ 0 + 4 + 0 + 0 + 0 + 4 + 0 + 3

⇒ 11

→ (Item 2, Item 3)

⇒ 3×3 + 1×0 + 4×0 + 3×0 + 1×3 + 2×0 + 0×3 + 1×0

⇒ 9 + 0 + 0 + 0 + 3 + 0 + 0 + 0

⇒ 12

→ (Item 2, Item 4)

⇒ 3×0 + 1×1 + 4×0 + 3×3 + 1×0 + 2×0 + 0×2 + 1×0

⇒ 0 + 1 + 0 + 9 + 0 + 0 + 0 + 0 = 10

→ (Item 2, Item 5)

⇒ 3×2 + 1×2 + 4×2 + 3×3 + 1×1 + 2×4 + 0×0 + 1×2

⇒ 6 + 2 + 8 + 9 + 1 + 8 + 0 + 2

⇒ 36

Item 3

→ (Item 3, Item 1)

= 3×0 + 0×4 + 0×0 + 0×0 + 3×0 + 0×2 + 3×1 + 0×3

= 0 + 0 + 0 + 0 + 0 + 0 + 3 + 0 = 3

→ (Item 3, Item 2)

= 3×3 + 0×1 + 0×4 + 0×3 + 3×1 + 0×2 + 3×0 + 0×1

= 9 + 0 + 0 + 0 + 3 + 0 + 0 + 0 = 12

→ (Item 3, Item 4)

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

$= 0+0+0+0+0+0+6+0 = 6$

→ (Item 3, Item 5)

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

$= 6+0+0+0+3+0+0+0 = 9$

Item 4

→ (Item 4, Item 1)

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

$= 0+4+0+0+0+0+2+0 = 6$

→ (Item 4, Item 2)

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

$= 0+1+0+9+0+0+0+0 = 10$

→ (Item 4, Item 3)

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

$= 0+0+0+0+0+0+6+0 = 6$

→ (Item 4, Item 5)

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

$= 0+2+0+9+0+0+0+0 = 11$

Item 5

→ (Item 5, Item 1)

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

$= 0+8+0+0+0+8+0+6 = 22$

→ (Item 5, Item 2)

⇒ 2×3 + 2×1 + 2×4 + 3×3 + 1×1 + 4×2 + 0×0 + 9×1

= 6 + 2 + 8 + 9 + 1 + 8 + 2 = 36

→ (Item 5, Item 3)

⇒ 2×3 + 2×0 + 2×0 + 3×0 + 1×3 + 4×0 + 0×3 + 2×0

= 6 + 0 + 0 + 0 + 3 + 0 + 0 + 0 = 9

→ (Item 5, Item 4)

⇒ 2×0 + 2×1 + 2×0 + 3×3 + 1×0 + 4×0 + 0×2 + 2×0

= 0 + 2 + 0 + 9 + 0 + 0 + 0 + 0 = 11

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	–	11	3	6	22
Item 2	11	–	12	10	36
Item 3	3	12	–	6	9
Item 4	6	10	6	–	11
Item 5	22	36	9	11	–

fig: Item / Item matrix

→ Next step is Item-Relationship matrix. So, that, take threshold of 10 produces the Item Relationship matrix.

i.e. >10 equal to 1 and remaining set to 0

	Item1	Item2	Item3	Item4	Item5
Item1	-	1	0	0	1
Item2	1	-	1	1	1
Item3	0	1	-	0	0
Item4	0	1	0	-	1
Item5	1	1	0	1	-

fig.: Item Relationship Matrix.

⇒ clustering by starting with "existing clusters" can be performed in a manner similar to the term Model.

Inital Assignments:

→ lets start with

→ item 1 and item 3 in class 1

i.e. class 1 = (item 1, item 3)

→ item 2 and item 4 in class 2

i.e. class 2 = (item 2, item 4)

This would produce following centroids for each class

Initial centroids:

class 1 = item 1, item 3

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

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

class 2 = item 2, item 4

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

$= \frac{3}{2}, \frac{2}{2}, \frac{4}{2}, \frac{6}{2}, \frac{1}{2}, \frac{2}{2}, \frac{2}{2}, \frac{1}{2}$

class1 centroids for item1, item2 ---- item5

~~class 1 centroid for item1~~:

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

class 1 centroids ~~for~~ x item 1:

Item 1 (entire row) = (0, 4, 0, 0, 0, 2, 1, 3) from vector matrix

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

$= \frac{0}{2} + \frac{16}{2} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} + \frac{4}{2} + \frac{4}{2} + \frac{9}{2} = \frac{33}{2}$

class 1 centroids ~~for~~ x item 2:

Item 2 (entire row) = (3, 1, 4, 3, 1, 2, 0, 1) from vector matrix

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

$= \frac{9}{2} + \frac{4}{2} + \frac{0}{2} + \frac{0}{2} + \frac{3}{2} + \frac{4}{2} + \frac{0}{2} + \frac{3}{2} = \frac{23}{2}$

class 1 centroids ~~for~~ x item 3:

Item 3 (entire row) = (3, 0, 0, 0, 3, 0, 3, 0)

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

$= \frac{9}{2} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} + \frac{9}{2} + \frac{0}{2} + \frac{12}{2} + \frac{0}{2} = \frac{30}{2}$

class1 centroids ~~for~~ x item 4:

Item 4 (entire row) = 0, 1, 0, 3, 0, 0, 2, 0 from vector matrix

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

$= \frac{0}{2} + \frac{4}{2} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} + \frac{8}{2} + \frac{0}{2} = \frac{12}{2}$

class 1 centroids ~~for~~ x Item 5:

Item 5 (entire row) = 2, 2, 2, 3, 1, 4, 0, 2 from vector matrix

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

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

$\Rightarrow \frac{6}{2} + \frac{8}{2} + \frac{0}{2} + \frac{0}{2} + \frac{3}{2} + \frac{8}{2} + \frac{0}{2} + \frac{6}{2} = \frac{31}{2}$

class 2 centroids for Item 1, Item 2, Item 3, Item 4, Item 5

class 2 centroids = $\frac{3}{2}, \frac{2}{2}, \frac{4}{2}, \frac{6}{2}, \frac{1}{2}, \frac{2}{2}, \frac{2}{2}, \frac{1}{2}$

~~Item 1~~

class 2 centroids ~~for~~ x item 1:

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

similarity = $\frac{3}{2} \times 0 + \frac{2}{2} \times 4 + \frac{4}{2} \times 0 + \frac{6}{2} \times 0 + \frac{1}{2} \times 0 + \frac{2}{2} \times 2$

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

$= \frac{0}{2} + \frac{8}{2} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} + \frac{4}{2} + \frac{2}{2} + \frac{3}{2} = \frac{17}{2}$

class 2 centroids ~~for~~ x item 2

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

similarity = $\frac{3}{2} \times 3 + \frac{2}{2} \times 1 + \frac{4}{2} \times 4 + \frac{6}{2} \times 3 + \frac{1}{2} \times 1 +$

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

$= \frac{9}{2} + \frac{2}{2} + \frac{16}{2} + \frac{18}{2} + \frac{1}{2} + \frac{4}{2} + \frac{0}{2} + \frac{1}{2} = \frac{51}{2}$

class 2 centroids x item 3

item 3 = 3, 0, 0, 0, 3, 0, 3, 0

similarity = $\frac{3}{2}\times3 + \frac{2}{2}\times0 + \frac{4}{2}\times0 + \frac{6}{2}\times0 + \frac{1}{2}\times3 + \frac{2}{2}\times0 +$

$\frac{2}{2}\times3 + \frac{1}{2}\times0$

$= \frac{9}{2} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} + \frac{3}{2} + \frac{0}{2} + \frac{6}{2} + \frac{0}{2} = \frac{18}{2}$

class 2 centroids x item 4:

item 4 = 0, 1, 0, 3, 0, 0, 2, 0

similarity = $\frac{3}{2}\times0 + \frac{2}{2}\times1 + \frac{4}{2}\times0 + \frac{6}{2}\times3 + \frac{1}{2}\times0 + \frac{2}{2}\times0 +$

$\frac{2}{2}\times2 + \frac{1}{2}\times0$

$= \frac{0}{2} + \frac{2}{2} + \frac{0}{2} + \frac{18}{2} + \frac{0}{2} + \frac{0}{2} + \frac{4}{2} + \frac{0}{2} = \frac{24}{2}$

class 2 centroids x item 5:

item 5 = 2, 2, 2, 3, 1, 4, 0, 2

similarity = $\frac{3}{2}\times2 + \frac{2}{2}\times2 + \frac{4}{2}\times2 + \frac{6}{2}\times3 + \frac{1}{2}\times1 +$

$\frac{2}{2}\times4 + \frac{2}{2}\times0 + \frac{1}{2}\times2$

$= \frac{6}{2} + \frac{4}{2} + \frac{8}{2} + \frac{18}{2} + \frac{1}{2} + \frac{8}{2} + \frac{0}{2} + \frac{2}{2} = \frac{47}{2}$

finally, we assigning 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

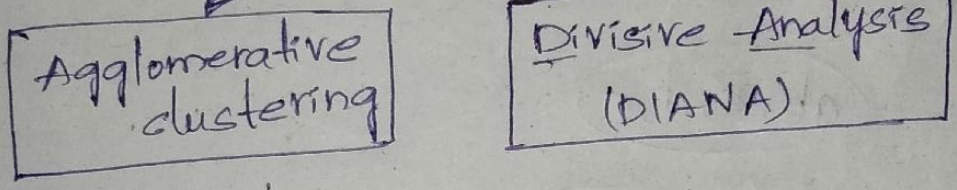
fig.: Item clustering with initial clusters.

Hierarchy of clusters

→ Hierarchical clustering in information Retrieval focuses on the area of hierarchical agglomerative clustering methods (HACM)

→ The term Agglomerative means the clustering process starts with "unclustered items" and perform pairwise similarity measures to determine the clusters.

~~Methods~~ ~~of~~ methods of Hierarchical clustering



Agglomerative clustering

→ It is a bottom-up approach

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

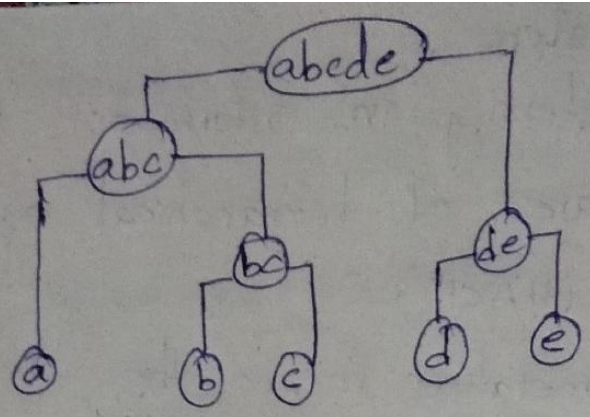
for e.g.: Take data set a, b, c, d, e

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

step 2: (bc) (de) (a) (b) (c) (d) (e)

step 3: (abc) (de) (bc) (a) (b) (c) (d) (e)

Step 4:



2. Divisive Analysis (DIANA):

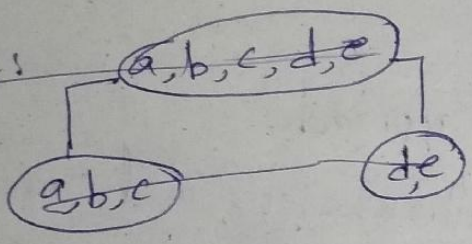
→ It is a Top-down approach

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

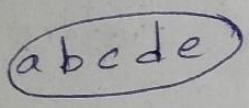
for e.g.: consider dataset {a,b,c,d,e}

step 1: ~~a,b,c,d,e~~

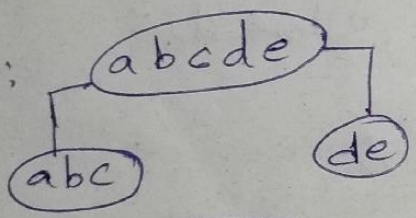
step 2: ~~a,b,c,d,e~~



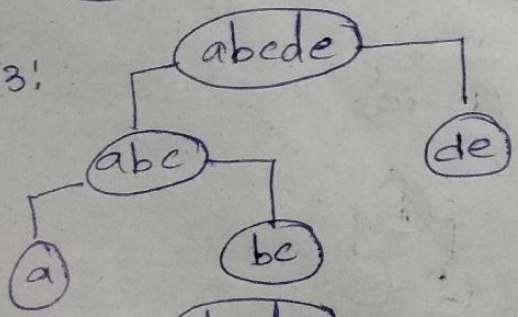
step 1:



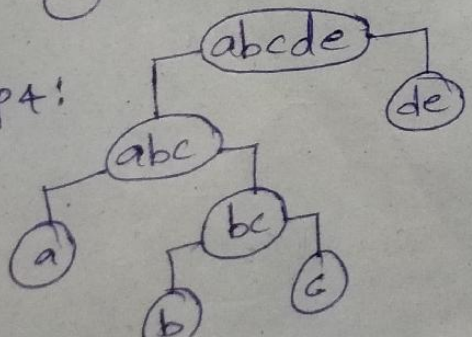
step 2:



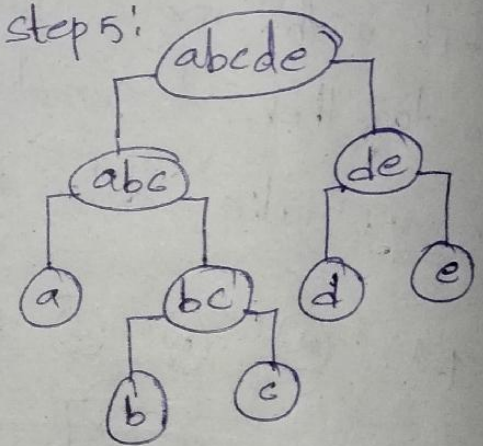
step 3:



step 4:



step 5:



The objectives of creating a hierarchy of clusters are to:

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

Here,

→ search overhead is reduced by performing "top-down searches"

→ It is difficult to create a visual display of the total item space.

Dendogram:

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

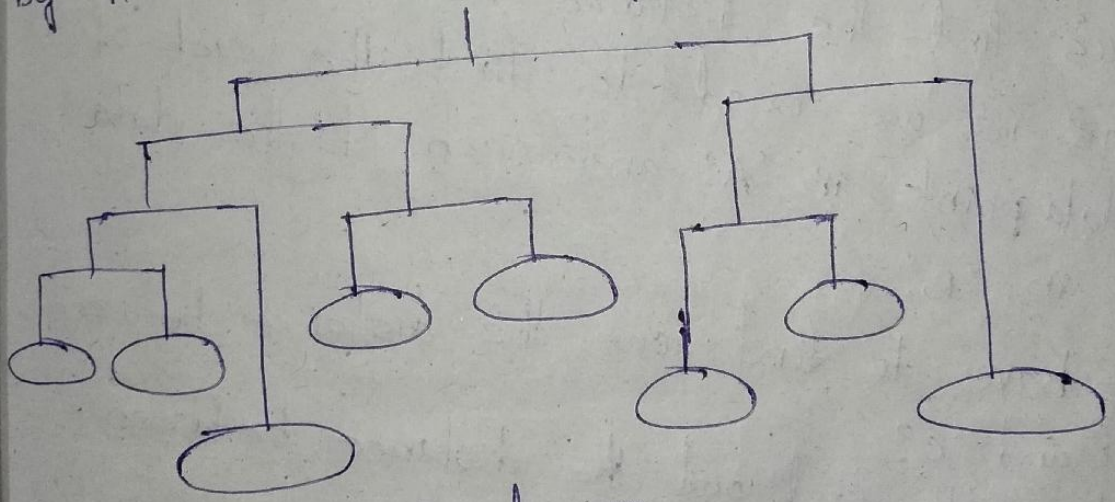


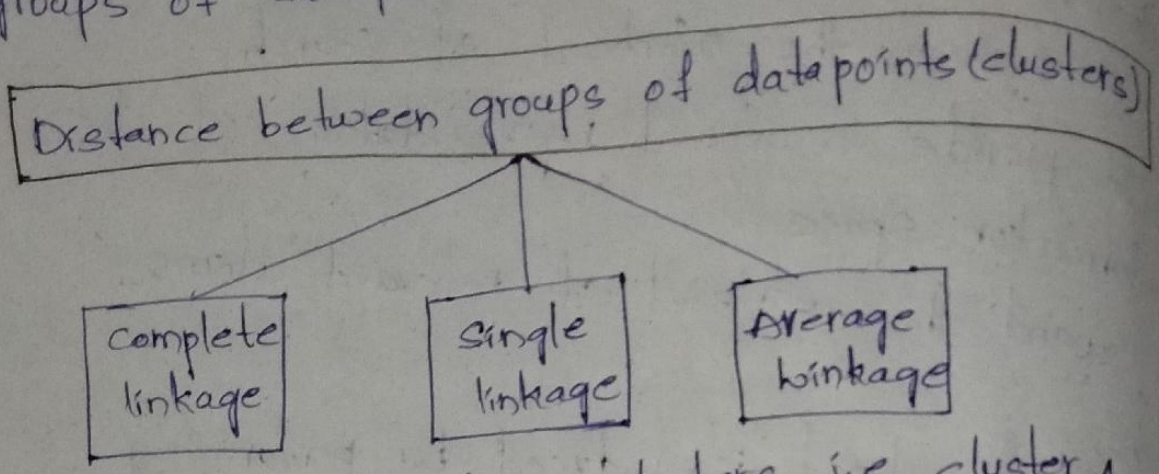
fig.: Dendogram.

Measures of Dissimilarity.

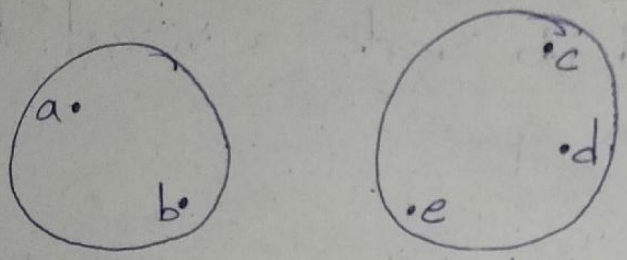
The Dissimilarity measure is used to decide 2 things i.e.

1. Which cluster should be combined.
2. Where a cluster should be split.

let see how to calculate distance between groups of datapoints (clusters).



Take an example of 2 clusters i.e. cluster A & cluster B.

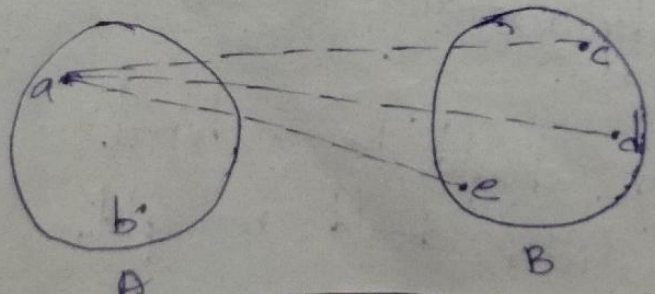


cluster-A cluster-B

for cluster 'A' data points are "a, b"
for cluster 'B' data points are "c, d, e"
Let us find the distance between 'A' and 'B' clusters i.e. we need to find the each of the data points in 'A' and each of the data point in 'B'.

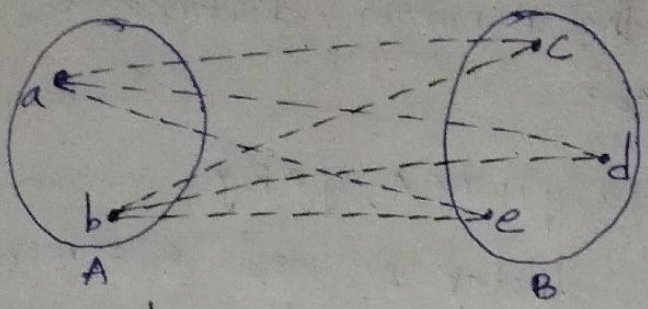
① We have to find out the distance between 'a' and 'c'.
We have to find out the distance between 'a' and 'd'.
We have to find out the distance b/w 'a' and 'e'.

i.e.



(2) We have to find out the distance between 'b' and 'c'.
We have to find out the distance b/w 'b' and 'd'.
We have to find out the distance b/w 'b' and 'e'.

i.e.

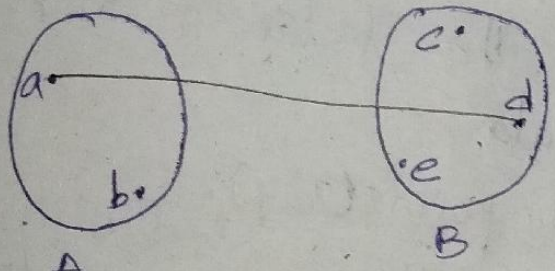


ii) Complete linkage:

The Complete linkage method indicates the maximum distance between the data points in cluster 'A' and cluster 'B'.

From the above figure, we can say that the maximum distance is achieved between 'a' and 'b' (data points)

i.e.



distance between cluster A & B = maximum of

$$(d(x,y) : x \in A, y \in B)$$

↓ distance between x, y ; ↓ x element of A, y element of B

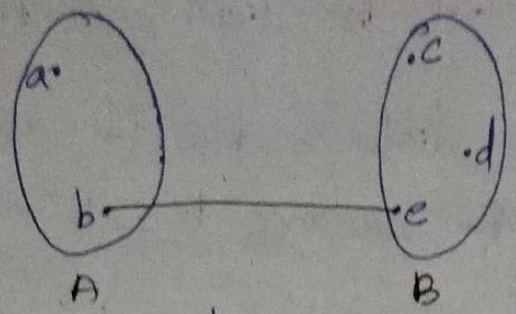
i.e.

$$d(A,B) = \max\{d(x,y) : x \in A, y \in B\}$$

iii) Single linkage:

In Single linkage method, we take the "minimum distance" from clusters.

i.e.



In this example, the minimum distance is available between 'b' and 'e'.

$d(A,B) = \min \{ d(x,y) : x \in A, y \in B \}$

distance between cluster 'A' and 'B' = minimum of distance between 'x' and 'y' i.e. $d(x,y)$ and 'x' is element of 'A' (i.e. $x \in A$) and 'y' is element of 'B' (i.e. $y \in B$)

Average linkage method:

$$d(A,B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x,y)$$

This formula indicates the $\sum$ of all the elements

→ The average of all the distances find out with the above formula.

Distance between data points (x,y):

Consider two data points $\bar{x} = (x_1, x_2, \ldots x_n)$ and $\bar{y} = (y_1, y_2, \ldots y_n)$.

These are different methods used to find out the distance between data points.

The methods are

* Euclidean distance
* Squared Euclidean distance
* Manhattan distance
* Maximum distance.

* Euclidean distance:

$$\sqrt{(x_1-y_1)^2+\cdots+(x_n-y_n)^2}$$

* squared Euclidean Distance:

$$(x_1-y_1)^2+\cdots+(x_n-y_n)^2$$

* Manhattan distance:

$$|x_1-y_1|+\cdots+|x_n-y_n|$$

* maximum distance:

$$\max\{|x_1-y_1|, |x_2-y_2|\cdots|x_n-y_n|\}$$