UNIT-III

Automatic Indexing: Classes of automatic indexing, Statistical indexing, Natural language, Concept indexing, Hypertext linkages **Document and Term Clustering:** Introduction, Thesaurus generation, Item clustering, Hierarchy of clusters.

AUTOMATIC INDEXING

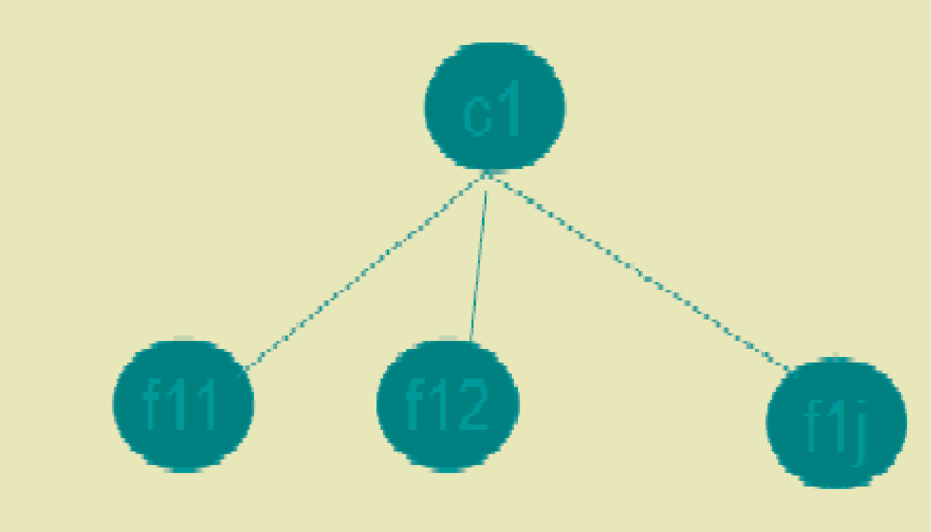
- Case: Total document indexing.
- Automatic indexing requires few seconds based on the processor and complexity of algorithms to generate indexes.'
- Adv. is consistency in index term selection process.
- Index resulting form automated indexing fall into two classes , weighted and un weighted .
- Un weighted indexing system : the existence of an index term in a

  document and some times its word location are kept as part of searchable data structure .
- Weighted indexing system: a attempt is made to place a value on the index term associated with concept in the document . Based on the frequency of occurrence of the term in the item .
- Values are normalized between 0 and 1.
- The results are presented to the user in order of rank value from highest number to lowest number .
- Indexing By term
- Terms (vocabulary) of the original item are used as basis of index process .
- There are two major techniques for creation of index statistical and natural language.
- Statistical can be based upon vector models and probabilistic models with a special case being Bayesian model(accounting for uncertainty inherent in the model selection process) .

- Called statistical because their calculation of weights use information such as frequency of occurrence of words .
- Natural language also use some statistical information , but perform more complex parsing to define the final set of index concept.
- Other weighted systems discussed as vectorised information system .
- The system emphasizes weights as a foundation for information detection and stores these weights in a vector form.
- Each vector represents a document. And each position in a vector represent a unique word(processing token) in a data base..
- The value assigned to each position is the weight of that term in the document.
- 0 indicates that the word was not in the document .
- Search is accomplished by calculating the distance between the query vector and document vector.
- Bayesian approach: based on evidence reasoning(drawing conclusion from evidence)
- Could be applied as part of index term weighing. But usually applied as part of retrieval process by calculating the relation ship

  between an item and specific query.
- Graphic representation each node represents a random variable arch between the nodes represent a probabilistic dependencies between the node and its parents .

  Two level Bayesian network
- “ c”“ represents concept in a query
- “f” representing concepts in an item



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

INFORMATION EXTRACTION

There are two processes associated with information extraction:

1.determination of facts to go into structured fields in a database and

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

The process of extracting facts to go into indexes is called Automatic File Build.

In establishing metrics to compare information extraction, precision and recall are applied with slight modifications.

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

6.1 Introduction to Clustering

The goal of the clustering was to assist in the location of information. Clustering of words originated with the generation of thesauri. Thesaurus, coming from the Latin word meaning “treasure,” is similar to a dictionary in that it stores words. Instead of definitions, it provides the synonyms and antonyms for the words. Its primary purpose is to assist authors in selection of vocabulary. The goal of clustering is to provide a grouping of similar objects (e.g., terms or items) into a “class” under a more general title. Clustering also allows linkages between clusters to be specified. The term class is frequently used as a synonym for the term cluster.

The process of clustering follows the following steps:

- Define the domain for the clustering effort. Defining the domain for the clustering identifies those objects to be used in the clustering process. Ex: Medicine, Education, Finance etc.
- Once the domain is determined, determine the attributes of the objects to be clustered. (Ex: Title, Place, job etc zones)
- Determine the strength of the relationships between the attributes whose co-occurrence in objects suggest those objects should be in the same class.
- Apply some algorithm to determine the class(s) to which each item will be assigned.

Class rules:

- ➢ A well-defined semantic definition should exist for each class.
- ➢ The size of the classes should be less.
- ➢ Within a class, one object should not dominate the class. For example, assume a thesaurus class called “computer” exists and it contains the objects (words/word phrases) “microprocessor,” “286-processor,” “386-processor” and “pentium.” If the term “microprocessor” is found 85 per cent of the time and the other terms are used 5 per cent each, there is a strong possibility that using “microprocessor” as a synonym for “286- processor” will introduce too many errors. It may be better to place “microprocessor” into its own class.
- ➢ Whether an object can be assigned to multiple classes or just one must be decided at creation time.

There are additional important decisions associated with the generation of thesauri that are not part of item clustering. They are

1) Word coordination approach: specifies if phrases as well as individual terms are to be clustered
2) Word relationships: Aitchison and Gilchrist specified three types of relationships: equivalence, hierarchical and nonhierarchical. Equivalence relationships are the most common and represent synonyms. Hierarchical relationships where the class name is a

general term and the entries are specific examples of the general term. The previous example of “computer” class name and “microprocessor,” “pentium,” etc Non-hierarchical relationships cover other types of relationships such as “object”-“attribute” that would contain “employee” and “job title.”

3) Homograph resolution: a 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.
4) Vocabulary constraints: this includes guidelines on the normalization and specificity of the vocabulary. Normalization may constrain the thesaurus to stems versus complete words.

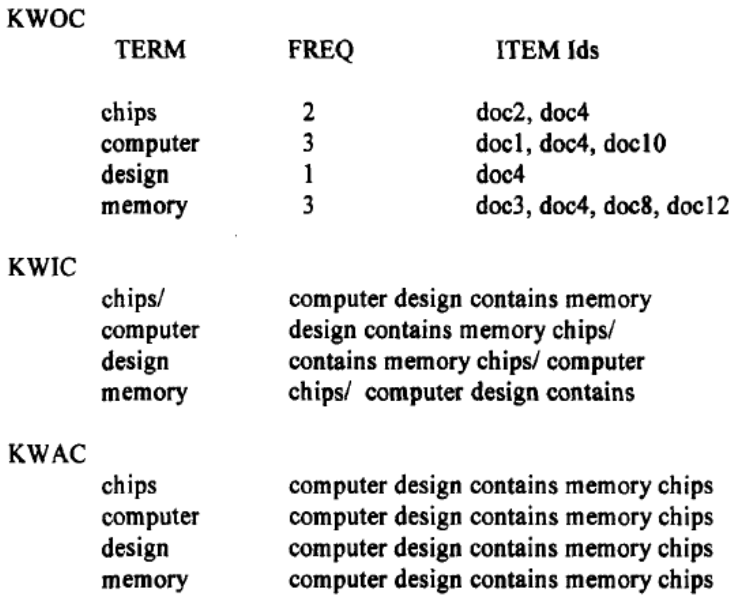
6.2 Thesaurus Generation

There are three basic methods for generation of a thesaurus; hand crafted, co- occurrence, and header-modifier based. In header-modifier based thesauri term relationships are found based upon linguistic relationships. Words appearing in similar grammatical contexts are assumed to be similar. The linguistic parsing of the document discovers the following syntactical structures: Subject-Verb, Verb- Object, Adjective-Noun, and Noun-Noun. Each noun has a set of verbs, adjectives and nouns that it co-occurs with, and a mutual information value is calculated for each using typically a log function.

6.2.1 Manual Clustering

The art of manual thesaurus construction resides in the selection of the set of words to be included. . Care is taken to not include words that are unrelated to the domain of the thesaurus. If a

concordance is used, other tools such as KWOC, KWIC or KWAC may help in determining useful words. A Key Word Out of Context (KWOC) is another name for a concordance. Key Word In Context (KWIC) displays a possible term in its phrase context. It is structured to identify easily the location of the term under consideration in the sentence. Key Word And Context (KWAC) displays the keywords followed by their context.



KWOC

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

KWIC

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

KWAC

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

Figure 6.1 Example of KWOC, KWIC and KWAC

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

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

6.2.2 Automatic Term Clustering

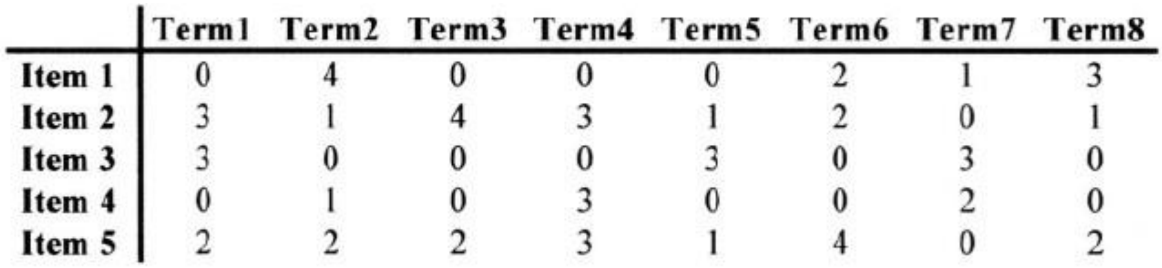
There are many techniques for the automatic generation of term clusters to create statistical thesauri. When the number of clusters created is very large, the initial clusters may be used as a starting point to generate more abstract clusters creating a hierarchy. The basis for automatic generation of a thesaurus is a set of items that represents the vocabulary to be included in the thesaurus. Selection of this set of items is the first step of determining the domain for the thesaurus. The processing tokens (words) in the set of items are the attributes to be used to create the clusters.

Implementation of the other steps differs based upon the algorithms being applied. The automated method of clustering documents is based upon the polythetic clustering where each cluster is defined by a set of words and phrases. Inclusion of an item in a cluster is based upon the similarity of the item's words and phrases to those of other items in the cluster.

6.2.2.1 Complete Term Relation Method

In the complete term relation method, the similarity between every term pair is calculated as a basis for determining the clusters. The 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 the unique words (processing tokens) in the items. The values in the matrix represent how strongly that particular word represents concepts in the item.

Figure 6.2 provides an example of a database with 5 items and 8 terms. To determine the relationship between terms, a similarity measure is required. The measure calculates the similarity between two terms. In Chapter 7 a number of similarity measures are presented. The similarity measure is not critical



	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Item 1	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
Item 4	0	1	0	3	0	0	2	0
Item 5	2	2	2	3	1	4	0	2

Figure 6.2 Vector Example

in understanding the methodology so the following simple measure is used:

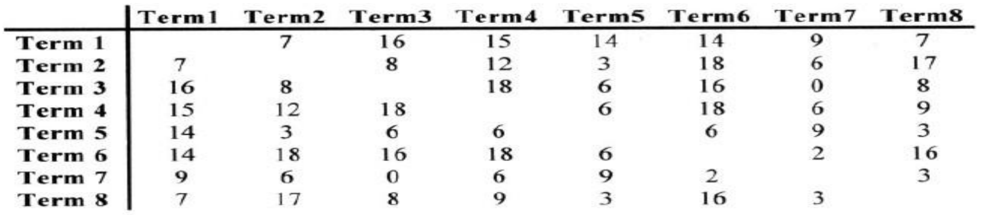
$$\mathbf{SIM(Term_i, Term_j) = \Sigma\ (Term_{k,i}\ )\ (Term_{k,j})}$$

where “k” is summed across the set of all items. In effect the formula takes the two columns of the two terms being analyzed, multiplying and accumulating the values in each row. The results can be paced in a resultant “m” by “m” matrix, called a Term-Term Matrix (Salton-83), where “m” is the number of columns (terms) in the original matrix. This simple formula is reflexive so that the matrix that is generated is symmetric. Other similarity formulas could produce a non- symmetric matrix.

Using the data in Figure 6.2, the Term-Term matrix produced is shown in Figure 6.3. There are no values on the diagonal since that represents the auto correlation of a word to itself. The 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. In this example, the threshold value of 10 is used. Thus two terms are considered similar if the similarity value between them is 10 or greater. This produces a new binary matrix called the Term Relationship matrix (Figure 6.4) that defines which terms are similar.

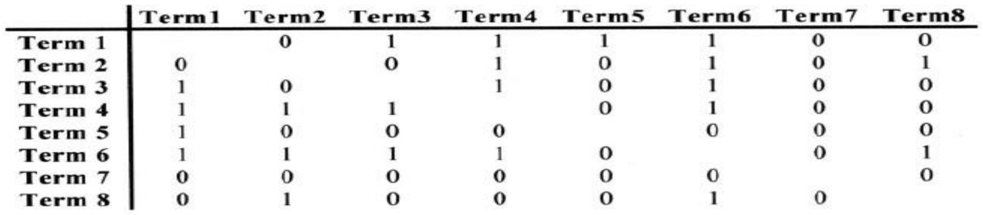
A one in the matrix indicates that the terms specified by the column and the row are similar enough to be in the same class. Term 7 demonstrates that a term may exist on its own with no other similar terms identified. In any of the clustering processes described below this term will always migrate to a class by itself.

The final step in creating clusters is to determine when two objects (words) are in the same cluster. There are many different algorithms available. The following algorithms are the most common: cliques, single link, stars and connected components.



	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
Term 1		7	16	15	14	14	9	7
Term 2	7		8	12	3	18	6	17
Term 3	16	8		18	6	16	0	8
Term 4	15	12	18		6	18	6	9
Term 5	14	3	6	6		6	9	3
Term 6	14	18	16	18	6		2	16
Term 7	9	6	0	6	9	2		3
Term 8	7	17	8	9	3	16	3	

Figure 6.3 Term-Term Matrix



	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8
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	

Figure 6.4 Term Relationship Matrix

Applying the algorithm to Figure 6.4, the following classes are created: Class 1 (Term 1, Term 3, Term 4, Term 6)

Class 2 (Term 1, Term 5)

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

Class 4 (Term 2, Term 6, Term 8)

Class 5 (Term 7)

Notice that Term 1 and Term 6 are in more than one class. A characteristic of this approach is that terms can be found in multiple classes. In single link clustering the strong constraint that every term in a class is similar to every other term is relaxed. The rule to generate single link clusters is that any term that is similar to any term in the cluster can be added to the cluster. It is impossible for a term to be in two different clusters. This in effect partitions the set of terms into the clusters. The algorithm is:

1. Select a term that is not in a class and place it in a new class
2. Place in that class all other terms that are related to it
3. For each term entered into the class, perform step 2
4. When no new terms can be identified in step 2, go to step 1.

Applying the algorithm for creating clusters using single link to the Term Relationship Matrix, Figure 6.4, the following classes are created:

Class 1 (Term 1, Term 3, Term 4, Term 5, Term 6, Term 2, Term 8)

Class 2 (Term 7)

There are many other conditions that can be placed on the selection of terms to be clustered.