**Clustering**

* One of the first uses of clustering was an attempt to cluster items discussing the same subject.

**Manual Clustering:**

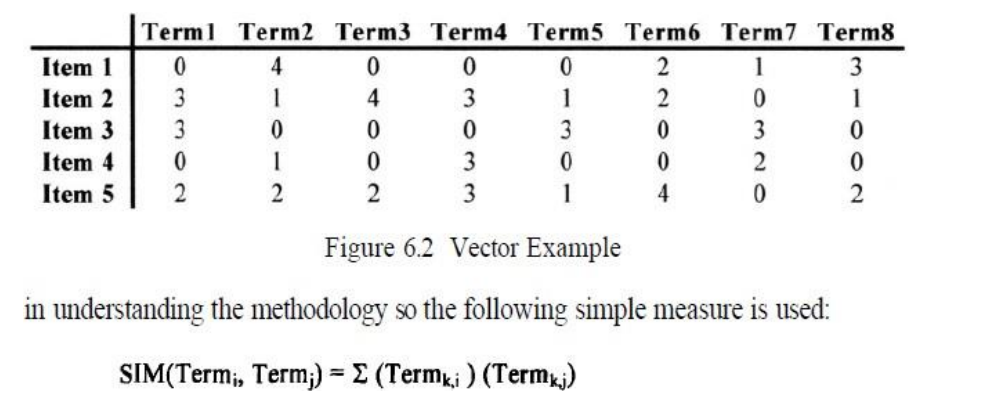
* The first step is to determine the domain for the clustering.
* Defining the domain assists in reducing ambiguities caused by homographs and helps focus the creator.
* The art of manual thesaurus construction resides in the selection of the set of words to be included.
* Care is taken to not include 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. Key Word And Context (KWAC) displays the keywords followed by their context.
* 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.

**Automatic Term Clustering:**

* 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.

1. **Complete Term Relation Method:**

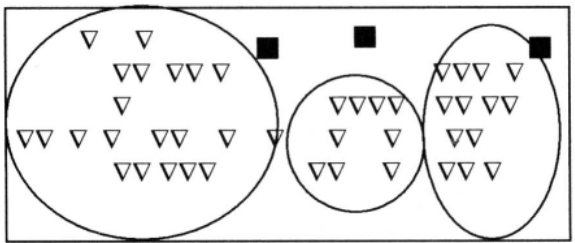
* In the complete term relation method, the similarity between every term pair is calculated as a basis for determining the clusters.
* The easiest way to understand this approach is to consider the vector model.
* The vector model is represented by a matrix where the rows are individual items and the 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.



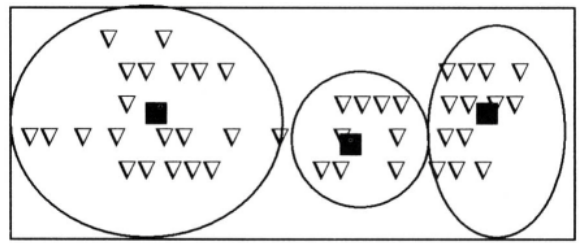
where “k” is summed across the set of all items.

1. **Clustering Using Existing Clusters:**

* An alternative methodology for creating clusters is to start with a set of existing clusters.
* This methodology reduces the number of similarity calculations required to determine the clusters.
* The initial assignment of terms to the clusters is revised by revalidating every term assignment to a cluster. The process stops when minimal movement between clusters is detected.
* To minimize calculations, centroids are calculated for each cluster. A centroid is viewed in Physics as the centre of mass of a set of objects. In the context of vectors, it will equate to the average of all of the vectors in a cluster.



Initial Centroids for Clusters



Centroids after Reassigning Terms

* The solid black box represents the centroid for each of the classes.

1. **One Pass Assignments**

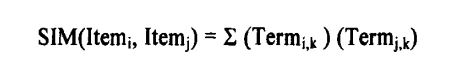
* This technique has the minimum overhead in that only one pass of all of the terms is used to assign terms to classes.
* The first term is assigned to the first class. Each additional term is compared to the centroids of the existing classes.
* A threshold is chosen. If the item is greater than the threshold, it is assigned to the class with the highest similarity.
* A new centroid has to be calculated for the modified class. If the similarity to all of the existing centroids is less than the threshold, the term is the first item in a new class. This process continues until all items are assigned to classes.

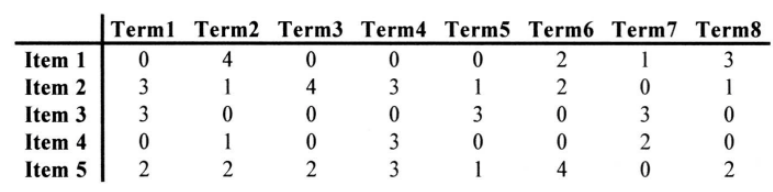
**Selective Dissemination of Information:**

* The Selective Dissemination of Information (Mail) Process provides the capability to dynamically compare newly received items in the information system against standing statements of interest of users and deliver the item to those users whose statement of interest matches the contents of the item.
* The Mail process is composed of the search process, user statements of interest (Profiles) and user mail files. As each item is received, it is processed against every user’s profile.
* A profile contains a typically broad search statement along with a list of user mail files that will receive the document if the search statement in the profile is satisfied.
* Selective Dissemination of Information has not yet been applied to multimedia sources.

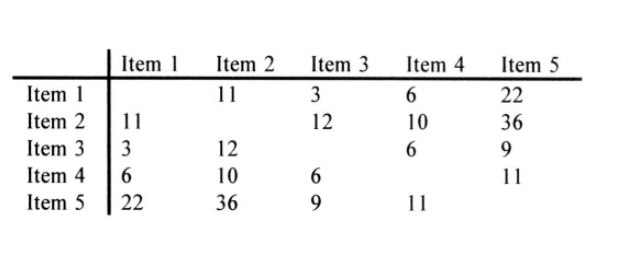
**Item Clustering:**

* Clustering of items is very similar to term clustering for the generation of thesauri.
* Manual item clustering is inherent in any library or filing system. In this case someone reads the item and determines the category or categories to which it belongs.
* When physical clustering occurs, each item is usually assigned to one category.
* With the advent of indexing, an item is physically stored in a primary category, but it can be found in other categories as defined by the index terms assigned to the item.
* Complete Term Relation Method, One Pass Assignments also applies to Item Clustering.
* Similarity between documents is based upon two items that have terms in common versus terms with items in common.
* Similarity equation:

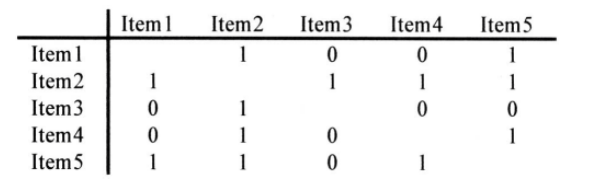




Item-Term Matrix



Item/Item Matrix



Item Relationship Matrix

All the items are in this one cluster, with Item 3 and Item 4 added because of their similarity to Item 2. The Star technique (i.e., always selecting the lowest nonassigned item) produces:

Class 1 - Item 1, Item 2, Item 5

Class 2 - Item 3, Item 2

Class 3 - Item4, Item2, Item5