**Assignment No 2**

**AIM: To implement single pass algorithm for clustering**

**OBJECTIVE:** To study

* What is Clustering?
* Single pass algorithm for clustering.
* Measure of association
* The graphical representation of clustering.

**THEORY:**

**Clustering**

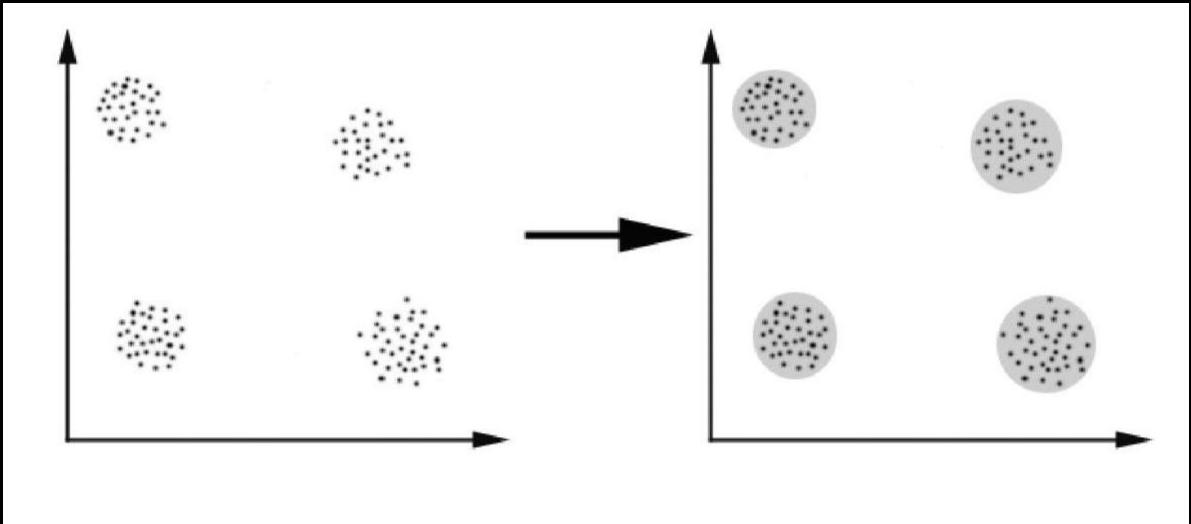
Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

A definition of clustering could be ―the process of organizing objects into groups whose members are similar in some way‖ . A *cluster* is therefore a collection of objects which are ―similar‖ between them and are ―dissimilar‖ to the objects belonging to other clusters.

Clustering is the process of grouping the documents which are relevant. It can be shown by a graph with nodes connected if they are relevant to the same request.

A basic assumption is that documents relevant to a request are separated from those which are not relevant i.e. the relevant documents are more like one another than they are like non relevant one.

**A simple graphical example:**



To identify the 4 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are ―close‖ according to a given distance (in this case geometrical distance). This is called *distance-based clustering*.

Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

**The Goals of Clustering**

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute ―best‖ criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, user could be interested in finding representatives for homogeneous groups (*data reduction*), in finding ―natural clusters‖ and describe their unknown properties (*“natural” data types*), in finding useful and suitable groupings (*“useful” data classes*) or in finding unusual data objects (*outlier detection*).

**Clustering Requirements**

The main requirements that a clustering algorithm should satisfy are:

* Scalability;
* Dealing with different types of attributes;
* Discovering clusters with arbitrary shape;
* Minimal requirements for domain knowledge to determine input parameters;
* Ability to deal with noise and outliers;
* Insensitivity to order of input records;
* High dimensionality;
* Interpretability and usability.

**SINGLE PASS CLUSTERING**

Single Pass clustering quickly by which we make incremental clustering to stream data. This clustering technique provides us with a simple yet flexible technique for stream data1. Given a collection of clusters and a threshold value h, if a new document n has the highest similarity more than h to some cluster, the document n is appended to the cluster, and if there exists no cluster, a new cluster is generated which contains only the document n. Clearly Single Pass Clustering is suitable for incremental clustering to temporal data (or data stream) since, once a document is assigned to a cluster, itis not changed in the future.

**The algorithm is as follows:**

1. Let h be a threshold value.
2. Let S be an empty set and d1 be the first document. We generate a new cluster C1 consisting of d1.
3. When a new document di(i > 1) comes in, calculate the similarity values to all the clusters C.
4. Let simmax be the highest value and Cdi the most similar cluster. If simmax > h, add di to Cdi and adjust the center of Cdi . Otherwise, we generate a new cluster Cdi that contains only di.
5. Repeat the process above until no data comes. In (4) we define simmax = MAX(sim(~di, ~C)). Also we define similarity of a document d and a cluster C where the center is VC as below (called cosine similarity):

sim(~d, ~C ) = d~ · V~C / |~d| | V~C|

**Measures of association**

Association is the similarity between objects characterized by discrete state attributes. The measure of similarity or association is designed to quantify likeness between the objects in such a way that an object in a group is more like the other members of the group that is like any object outside the group then a cluster method enables such a group structure to be discovered.

There are five commonly used measures of association in IR.

|  |  |
| --- | --- |
| 1.| X ∩ Y | | Simple matching coefficient |
| 2.| X ∩ Y | / | X | + | Y | | Dices coefficient |
| 3.| X ∩ Y | / | X U Y | | Jaccard’s coefficient |
| 4.| X ∩ Y | / | X |1/2 \* | Y | ½ | Cosine coefficient |
| 5.| X ∩ Y | / min ( | X |, | Y | ) | Overlap coefficient |

In short, measure of association is calculated by this program by taking into account frequency of occurrence of words in both the documents i.e least value of frequency of occurrence of a common word in both documents is considered for finding out the measure of association.

**Classification methods**

1. Multi state attribute (E.g. : Colour)
2. Binary state (E.g. :Keyword)
3. Numerical (E.g.: Hardness scale or weighted keyword)
4. Probability distribution

**Cluster hypothesis**

The hypothesis can be simply stated as closely associated document tend to be relevant to the same request. This hypothesis is referred as Cluster hypothesis.

The basic assumption in retrieval system is that documents relevant to a request are separated from those which have not relevant. Compute the association between all pairs of documents.

1. Both of which are relevant to a request and
2. One of which is relevant and the other is not

**Criteria to be satisfied by the methods of clustering**

Two criteria have frequently been used: 1. Theoretical soundness of the method

a.The method produces a clustering which is unlikely to be altered drastically when further objects are incorporated, i.e. it is stable under growth.

b.The method is stable in the sense that small errors in the description of the object lead to small changes in the clustering.

c.The method is independent of the initial of the objects.

1. The second criterion for choice is the efficiency of the clustering process in terms of speed and storage requirements.

**Two approaches of clustering:**

1. The clustering is based on a measure of similarities between the objects to be clustered.

The example of first approach is graph theoretic method which can be used to define clusters in terms of graphs derived from measure of similarity.

2. The cluster method proceeds directly from the object description.

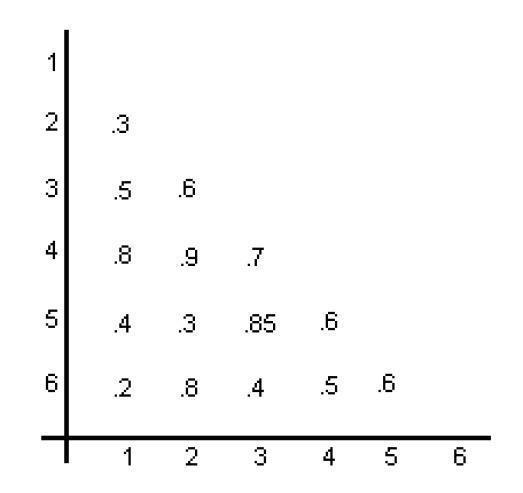
**Graphical representation of clustering**

Here similarity matrix is used in order to draw the graph; the documents having measure of association greater than threshold value can be represented by the edge in the graph. This is an identical cluster which can shown as connected graph.

From the graph below, it can be easily understood that all documents are associated. But documents like 2 & 5 are not directly associated & same is the case for documents 4 & 5.In this way clusters can be depicted.

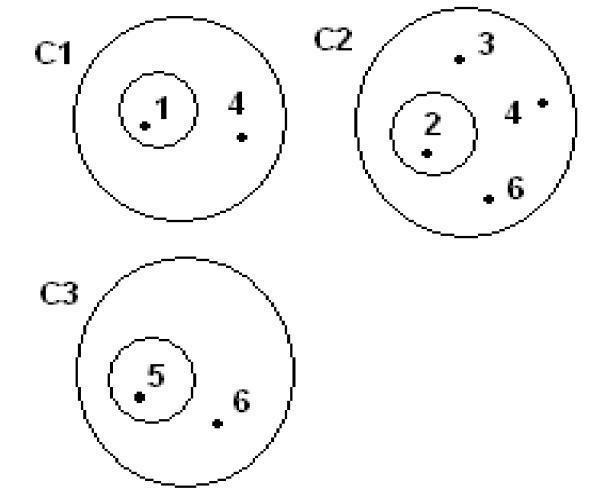
**Example:-**

**Objects {1, 2, 3, 4, 5, 6}**



**Threshold: 0.59**

**Clusters are :**



**CONCLUSION:** Thus, we have implemented the single pass algorithm for clustering.