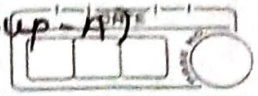
Experiment No:- 02 (Group-A)



Problem Statement:
Implement single-pass algorithm for clustering of files

objectives:
To study:
1. What is clustering?
2. Single pass algorithm for 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 organising objects into groups whose members with nodes connected are similar in some way". A cluster is therefore a collection of objects which are "similar" beth them and are "disimilar" to the objects belonging to other clusters.

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

- In choosing a cluster method for use in experiment IR, two criteria have frequently been used. The first of these is the theoretical soundness of the method. It means first of these certain criteria of adequacy. To list some of the more important of these:

(1) In choosing The method produces a clustering which is unlikely to be altered drastically when further details are incorporated.
2) The method is independent of the initial ordering of the objects.

The algorithm also use, a no. of empirically determined parameters such as:
1) The number of clusters desired.
2) A minimum and max size of each cluster.
3) A threshold value on the matching function, which an object will not be included in a cluster.
4) The control of overlap betn clusters.
5) An arbitrarily chosen objective fun which is optimized

Cluster hypothesis:-

- closely associated documents tend to be relavent to the same requests.

DATE

Single pass Algorithm:

1. The object descriptions are preprocessed serially.
2. The first object becomes the cluster representative of the first cluster.
3. Each subsequent object is matched against all cluster representatives existing at its processing time.
4. A given object is assigned to one cluster according to some condition on the matching fun^n.
5. When an object is assigned to a cluster the representative for that cluster is recomputed.
6. If an object fails a certain test it becomes the cluster representative of a new cluster.

Algorithm:

1. The object description are processed serially.
2. The first object becomes the cluster representative of the first cluster.
3. Each subsequent object is matched against all cluster representatives existing at its processing time.
4. A given object is assigned to one cluster (or more) according to some condition on the matching fun^n.
5. When an object is assigned to a cluster the representative for that cluster is recomputed.
6. If an object fails a certain test it beromes the cluster representative of a new cluster.

Conclusion:

Implementation is concluded by stating analysis of single pass algorithm for clustering with the benefits & limitations.

24-10-2024