A Survey On Various Approaches In Data Clustering

Ms. R. Suganya,
Asst.Professor,
Dept of Computer Science,
Dr. Sns Rajalakshmi College of Arts and Science,
Coimbatore-49
India,

S. Sindhu,
Mphil Scholar,
Dept of Computer Science,
Dr. Sns Rajalakshmi College of Arts and Science,
Coimbatore-49,
India,

ABSTRACT

Fast retrieval of the relevant information from the databases has always been a significant issue. Different techniques have been developed for this purpose, one of them is Data Clustering.

In this paper Data Clustering is discussed along with its two traditional approaches and their analysis. Some applications of Data Clustering like Data Mining using Data Clustering and Similarity Searching in Medial Image Databases are also discussed along with a Microsoft Windows NT Operating system.

Keywords: data clustering, clustering method, clustering applications, clustering in Microsoft Windows NT.

1. INTRODUCTION

Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are also to be defined. Precisely, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the numbers of disk accesses are to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

Example to Elaborate the Idea of Clustering
In order to elaborate the concept a little bit, let us take the example of the library system. In a library books concerning to a large variety of topics are available. They are always kept in form of clusters. The books that have some kind of similarities among them are placed in one cluster.

For example, books on the database are kept in one shelf and books on operating systems are kept in another cupboard, and so on. To further reduce the complexity, the books that cover same kind of topics are placed in same shelf. And then the shelf and the cupboards are labeled with the relative name.

Now when a user wants a book of specific kind on specific topic, he or she would only have to go to that particular shelf and check for the book rather than checking in the entire library.

A cluster is an ordered list of objects, which have some common characteristics. The clustering method determines how the distance should be computed.

A similarity measure SIMILAR (Di, D_j) can be used to represent the similarity between the documents. Intermediate values are obtained for cases of partial agreement.

The lowest possible input value of similarity required to join two objects in one cluster. Similarity between objects calculated by the function SIMILAR ($D_{i,}$ Dj), represented in the form of a matrix is called a similarity matrix.

The dissimilarity coefficient of two clusters is defined to be the distance between them. The smaller the value of dissimilarity coefficient, the more similar two clusters are.

First document or object of a cluster is defined as the initiator of that cluster i.e. every incoming object's similarity is compared with the initiator. The initiator is called the cluster seed.

2. CLUSTERING METHODS

There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap.

These methods are sometimes divided into partitioning methods, in which the classes are mutually exclusive, and the less common clumping method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar, however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains.

The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods, the hierarchy is build up in a series of N-1 agglomerations, or Fusion, of pairs of objects, beginning with the un-clustered dataset. The less common divisive methods begin with all objects in a single cluster and at each of N-1 steps divides some clusters into two smaller clusters, until each object resides in its own cluster.

2.1 Partitioning Methods

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered.

In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset. Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run

2.2Hierarchical Agglomerative Methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster.
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains, return to step 2.

Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged.

2.3 The Single Link Method (SLINK)

The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

2.4 The Complete Link Method (CLINK)

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

2. 5 The Group Average Method

The group average method relies on the average value of the pair wise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter – cluster similarity, each object is , on average more like every other member of its own cluster then the objects in any other cluster.

2.6 Text Based Documents

In the text based documents, the clusters may be made by considering the similarity as some of the key words that are found for a minimum number of times in a document. Now when a query comes regarding a typical word then instead of checking the entire database, only that cluster is scanned which has that word in the list of its key words and the result is given. The order of the documents received in the result is dependent on the number of times that key word appears in the document.

3. CLUSTERING APPLICATIONS

Data clustering has immense number of applications in every field of life. So the history of data clustering is old as the history of mankind.

In computer field also, use of data clustering has its own value. Specially in the field of information retrieval data clustering plays an important role. Some of the applications are listed below.

3.1 Similarity searching in Medical Image Database

This is a major application of the clustering technique. In order to detect many diseases like Tumor etc, the scanned pictures or the x-rays are compared with the existing ones and the dissimilarities are recognized.

We have clusters of images of different parts of the body. For example, the images of the CT scan of brain are kept in one cluster. To further arrange things, the images in which the right side of the brain is damaged are kept in one cluster. The hierarchical clustering is used. The stored images have already been analyzed and a record is associated with each image. In this form a large database of images is maintained using the hierarchical clustering.

Now when a new query image comes, it is firstly recognized that what particular cluster this image belongs, and then by similarity matching with a healthy image of that specific cluster the main damaged portion or the diseased portion is recognized. Then the image is sent to that specific cluster and matched with all the images in that particular cluster. Now the image, with which the query image has the most similarities, is retrieved and the record associated to that image is also associated to the query image. This means that now the disease of the query image has been detected.

Using this technique and some really precise methods for the pattern matching, diseases like really fine tumor can also be detected.

So by using clustering an enormous amount of time in finding the exact match from the database is reduced.

3.2 Data Mining

Another important application of clustering is in the field of data mining. Data mining is defined as follows.

Definition1: "Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques."

Definition2: Data mining is a "knowledge discovery process of extracting previously unknown, actionable information from very large databases."

Use of Clustering in Data Mining: Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships.

This technique supports the development of population segmentation models, such as demographic-based customer segmentation.

Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign.

For example, a company that sale a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

4 .CLUSTERING IN WINDOWS NT

Pfister defines a cluster as "a parallel or distributed system that consists of a collection of interconnected whole computers that is utilized as a single, unified computing resource".

In general, the goal of a cluster is to make it possible to share a computing load over several systems without either the users or system administrators needing to know that more than one system is involved.

We describe the architecture of the clustering extensions to the Windows NT operating system. Windows NT clusters provide some user visible advantages: improved availability by continuing to provide a service even during hardware or software failure.

If any component in the system, hardware or software fails the user may see degraded performance, but will not lose access to the service. Increased scalability by allowing new components to be added as system load increases. Lastly, clusters simplify the management of groups of systems and their applications by allowing the administrator to manage the entire group as a single system.

Windows NT Clusters are, in general, shared nothing clusters. This means that while several systems in the cluster may have access to a device or resource, it is effectively owned and managed by a single system at a time.

Members of a cluster are referred to as nodes or systems. The Cluster Service is the collection of software on each node that manages all cluster specific activity. Cluster service is a separate, isolated set of components.

The following things work on top of clustering in Windows NT Environment.

- The Node Manager handles cluster membership, watches the health of other cluster systems.
- Configuration Database Manager maintains the cluster configuration database.

Creating a Cluster

When a system administrator wishes to create a new cluster, the administrator will run a cluster installation utility on the system to become the first member of the cluster. For a new cluster, the database is created and the initial cluster member is added.

The administrator will then configure any devices that are to be managed by the cluster software. We now have a cluster with a single member. In the next step of clustering each node is added to the cluster by means of similarity on the basis of the resources used. The new node automatically receives a copy of the existing cluster database.

5. CONCLUSION

In this paper, we try to give the basic concept of clustering by first providing the definition and clustering and then the definition of some related terms. We give some examples to elaborate the concept. Then we give different approaches to data clustering and also discussed some algorithms to implement that approaches. The partitioning method and hierarchical method of clustering were explained. The applications of clustering are also discussed with the examples of medical images database, data mining using data clustering.

So we try to prove the importance of clustering in every area of computer science. We also try to prove that clustering is not something really typical to databases but it has its applications in the fields like networking.

6. REFERENCES

[1]Cluster analysis - wikipedia, the free encyclopedia
En.wikipedia.org/wiki/cluster analysis

[2] Survey of clustering data mining techniques
Www.cs.iastate.edu/~honavar/clustering-survey.pdf

[3] rob short, rod gamache, john vert and mike massa "windows nt clusters for availability and scalability" microsoft online research papers, microsoft corporation.

[4] jim gray "qqjim gray’s nt clusters research agenda" microsoft online research papers, microsoft corporation.

[5] Clustering: a survey - upload & share power point presentations and Www.slideshare.net/rcapaldo/cluster-analysis-presentation

[6] Clustering survey - #2fishygirl on scribd scribd
Www.scribd.com/doc/24442369/clustering-survey

- [7] Getting started: clustering ideas Grammar.ccc.commnet.edu/grammar/composition/brainstorm_clustering.htm
- [8] Clustering Www.vi.virginia.gov/search97/doc/user/10_is3.htm
- [9] Different techniques of data clustering Members.tripod.com/asim_saeed/paper.htm
- [10] Soft clustering: an overview Www.interscience.in/spiss_ijcct_accta_2010vol1_nol2/wddw_paper1.pdf