Clustering Analysis of Library System Using Divisive Method

Htwe Ei Hlaing, Than Nwet Aung
Computer University (Meiktila)
hlaing.cf@gmail.com

ABSTRACT

Data mining is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses and other massive information repositories. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning. Nowadays, library has been constrained by its large amount of books. To overcome this problem, this paper presents a system which is used for the analysis of library system by using hierarchical clustering, divisive methods. In divisive methods of hierarchical cluster analysis, the clusters obtained at the previous step are subdivided into smaller clusters. So, the system provides the large books into several groups with their relevant field. This system provides the finding relevant and desired books with quickly.

1. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amounts of data. Data mining or knowledge discovery is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either shifting through an immense amount of material, or intelligently probing it to find where the value resides [1].

Clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the interclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together [1].

This system is used for the analysis of library system by using hierarchical clustering. In hierarchical clustering, the data are not partitioned into a particular cluster in a single step. Hierarchical clustering is subdivided into agglomerative method, which proceeds by series of fusion of the n object into groups, and divisive methods, which separate n objects successively into finer grouping. Agglomerative techniques are more commonly used.

The idea behind clustering was that if certain document match a user query, the documents in the same cluster also are likely to be relevant. In the library system, there are many kinds of books, journals, news, business, and papers etc. This system is provided for fining relevant and desired books with quickly. The rest of the paper is organized as follow: section 2 shows related work, section 3 explains hierarchical clustering. Section

4 describes use case diagram. Section 5 shows experimental results. Finally, section 6 highlights the conclusion.

2. RELATED WORK

University of North Carolina at Chapel Hill has applied a number of different clustering schemes to their data, but has focused on agglomerative and divisive hierarchical methods. These methods have proven to be effective and their graphical representation as trees provides a useful way of identifying and analyzing groups of related communication patterns. In clustering source level communication patterns, they extract from each connection vector a number of numerical features that are designed to capture important aspects of the two-way data transfer it describes. They initially tested their approach by clustering training data sets with a small number of connections. They analyzed this data set using divisive hierarchical clustering, after converting each connection vector into a feature vector that included all of the statistical features. Ten of the connection in the data set carried Telnet traffic (i.e., interactive remote shell), while the other ten carried web traffic. The communication patterns used by these two protocols are quite different, so appropriate clustering should be able to split the data set into two subpopulations. The clustering algorithm accurately separated two different communication patterns. Each connection was first transformed into a connection vector, and then summarized into a feature vector. Feature vectors were clustered using the average-linkage agglomerative method. The methodology for the study of data exchange patterns in transport connections provides an effective way of visualizing and clustering the behavior of internet sources. The traffic workloads for testing and simulation should reflect the clearly distinguished patterns of communication uncovered by clustering data set of internet connections [2].

Segmentation is a broad term, covering a wide variety of problems and of techniques. Clustering is a process where by a data set is replaced by clusters, which are collections of data points that "belong together". It is natural to think of image segmentation as clustering: they represent an image in terms of clusters of pixels that "belong together". The specific criterion to used depends on the application. Pixels may belong together because it has the same color and/or it has the same texture and /or it is nearby, etc. Divisive clustering uses it to split insufficiently "coherent" clusters. Segmentation using simple clustering methods is relatively to take a clustering method and build an image segmental from it [3].

3. HIERARCHICAL CLUSTERING

There are two approaches to improving the quality of hierarchical clustering: (1) perform careful analysis of object "linkages" at each hierarchical partitioning method, (2) integrate hierarchical agglomeration method and iterative relocation by first using a hierarchical agglomerative algorithm and then refining the result using iterative relocation.

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is called Hierarchical Agglomerative Clustering (HAC). Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached.

Modern visualized techniques available for library books by applying them, a user does not need to wade through all the documents but may grasp a high-level picture instead. This system implements clustering in a collection one-by-one techniques of divisive methods and evaluate how closely cluster produced by a computer resemble those created by human experts [4]**.**

3.1. Agglomerative Methods

An agglomerative hierarchical clustering procedure

produce a series of partitions of the data, P_n, P_{n-1}, ... , P_1. The first P_n consists of n single object 'clusters', the last P_1, consists of single group containing all n cases. At each particular stage the method joins together the two clusters which are closest together (most similar). Differences between methods arise because of the different ways of defining distance (or similarity) between clusters. Several agglomerative techniques will now be described in detail [4].

3.2. Divisive Analysis (Diana)

DIANA is a hierarchical clustering technique, but its main difference with the agglomerative method (AGNES) is that it constructs the hierarchy in the inverse order.

Initially (Step 0), there is one large cluster consisting of all n objects. At each subsequent step, the largest available cluster is split into two clusters until finally all clusters, comprise of single objects. Thus, the hierarchy is built in n-1 steps (n= number of objects).

In the first step of an agglomerative method, all possible fusions of two objects are considered leading to $\frac{n(n-1)}{2}$ combinations. In the divisive method based on the same principle, there are 2^{n-1}-1 possibilities to split the data into two clusters. This number is considerably larger than that in the case of an agglomerative method. To avoid considering all possibilities, the algorithm proceeds as follows:

1. Find the object, which has the highest average dissimilarity to all other objects. This object initiates a new cluster– a sort of a splinter group.
2. For each object i outside the splinter group compute (i=an object).
3. D_i = [average d(i,j) j $\notin$ $R_{splinter\ group}$] - [average d(i,j) j $\in$ $R_{splinter\ group}$] (D= dissimilarity between two objects i and j, j= an object, R= the splinter group).
4. Find an object h for which the difference D_h (h= an object) is the largest. If D_h is positive, then h is, on the average close to the splinter group.
5. Repeat Steps 2 and 3 until all differences D_h are negative. The data set is then split into two clusters.
6. Select the cluster with the largest diameter. The diameter of a cluster is the largest dissimilarity between any two of its objects. Then divide this cluster, following steps 1-4.
7. Repeat Step 5 until all clusters contain only a single object [5].

4. USE CASE DIAGRAM

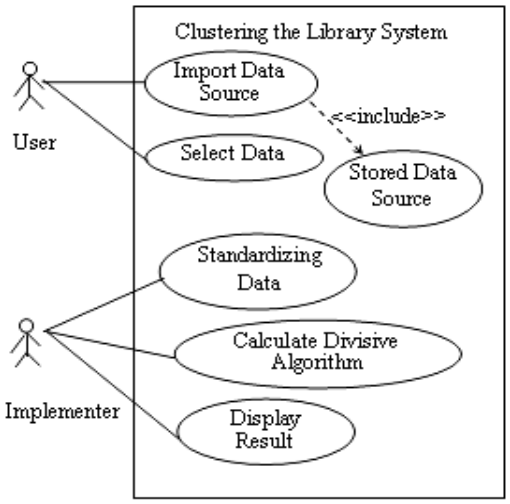

Figure 1. Use case diagram of the system

The figure 1 illustrates the system with UML notation of use case diagram. Use case diagram gives a graphical table of the system boundary and interaction between users and the system and shows human initiated functionality.

The system describes function requirements. It includes actor, use case and relationship. The user associates import data source. Data source depends on Import Data Source.

4.1. Training Dataset

In the system, there are large amount of variety of book stored in the database or data repository.

The table as shown in table 1 is the design view of the large amount of books. In this book table, there are eight attributes. They are ID,

ISBN, Author, Title, Edition, Publisher, Price and Year. The book table from the selected data source is imported by the application user. If the user calculates the desired number of cluster group, the clustered group of books will display with weight data or source data.

Table 1. Book table design

Fields	Data Type	Description
ID	Number	Primary Key
ISBN	Text	
Author	Text	
Title	Text	
Edition	Text	
Publisher	Text	
Price	Number	
Year	Date	

5. EXPERIMENTAL RESULTS

This system contains import and clustering. When user click the import button, the import data set form as shown in figure 2 is appeared.

To import data set into the system, type the database name into the textbox or click "Browse" button to choose the MS Access database. And type the table name into the table name text box. Then double click the table name into the above table lists to view the content of the table.

Data table will show in the "Import table view" grid. After all, click "Import" button to import database.

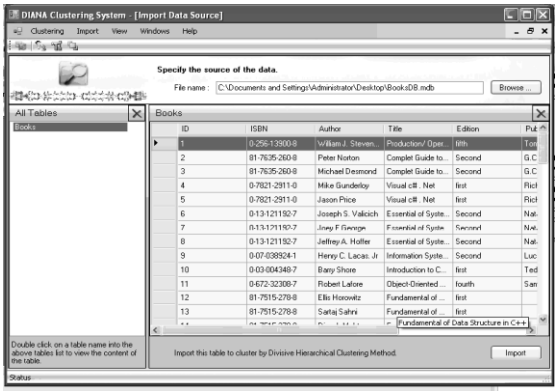

Figure 2. Import dataset form

Clustering consists of data source, standardizing and DIANA clustering. When user click the data source button, select data source form as shown in figure 3 is appeared.

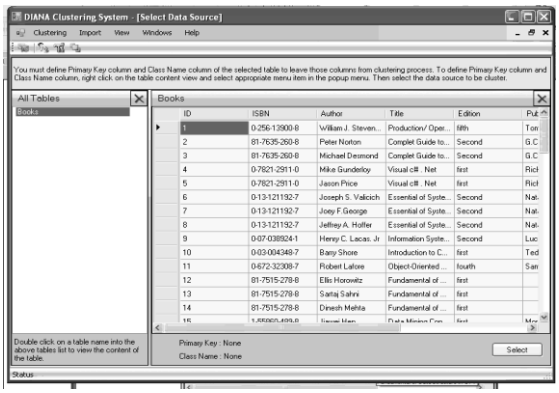

Figure 3. Select data source form

User must define Primary column and class name column of the selected table to leave these columns form clustering process. To define Primary key column and class name column, right click on the table content view and select appropriate menu item in the popup menu. Then select data source to be cluster.

When user clicks the standardizing button, standardizing data source form (ISBN) as shown in figure 4 is appeared.

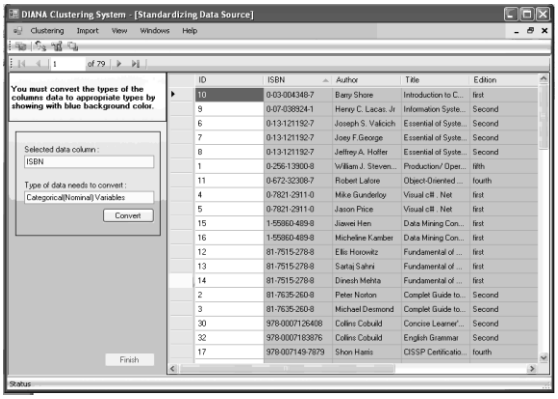

Figure 4. Standardizing data source form (ISBN)

If the important data table contains categorized data, user needs to convert it, to weighted numeric data format. The types of the columns data must be converted to appropriate type by showing with blue color. The user need to select data column and then press "Convert" button. When user click the DIANA clustering button, DIANA clustering form (a) as shown in figure 4 is appeared.

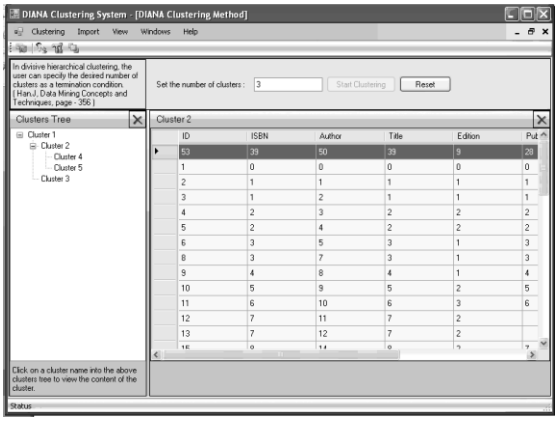

Figure 4. DIANA clustering form (a)

To start clustering, set the number of cluster into the text box. The maximum value is limited by the system as the number of rows of the important data table. Then click “Start Cluster” to start cluster analysis process. After the clustering process, the system shows the contents of the first cluster of the value into the grid view. In divisive hierarchical clustering, the user can specify the desired number of clusters as a termination condition.

If the user specifies the number of clusters 3 as a termination condition, the user can view the cluster group .The user can also view the cluster group by clicking on a cluster name into the above clusters tree to view the content of the cluster. If the user specifies the number of clusters 7 as a termination condition, the user can view the cluster group as shown in figure 5. The user can also view the cluster group by clicking on a cluster name into the above clusters tree to view the content of the cluster with weight data format.

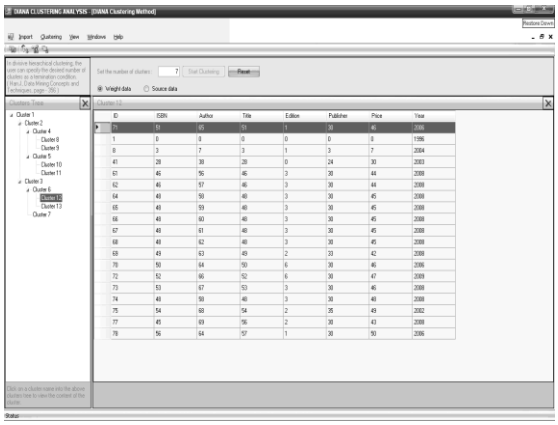

Figure 5. DIANA clustering form (b)

6. CONCLUSION

Today, using HCA methods have become very popular. HCA methods are discussed and used to produces the mining technique of clustering.

Many advantages can be received by the computer system. By using the hierarchical clustering divisive methods, user can find relevant and desired books quickly. The user imports data in database and then imported data is selected to cluster. Data are standardized to measure. Data are calculated to cluster by using Divisive algorithm the cluster tree come out. The concept is easy to understand and apply. The resulting clusters are less sphere-shaped than partitioning methods, but still have that tendency because distance is used as similarity measure. The number of clusters is also chosen at a later stage, which is better than

partitioning methods. In this paper, the system is used for only divisive approach in Hierarchical Clustering Method. This system can be extended the comparison of the other methods such as partitioning methods, density-based methods, grid-based methods and model-based methods and then test the various data sets in case study example.

REFERENCES

[1] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufman, 2001.

[2] "Statistical Clustering of Internet Communication Patterns", Felix Hernandez-Campos, F. Donelson Smith.

[3] "Segmentation Using Clustering Methods", decsai.ugr.es.

[4] http://www.scribd.com/doc/6569475/rbookonlinerading

[5] www.unesco.org/webworld/idams/advguide/Chapt7_1_5.htm