A Review of C-TREND Using Complete-Link Clustering for Transactional Data

Arna Prabha Jena
Dept. of Computer Science Engineering
Centurion University
Bhubaneswar, India
arna.prabha@gmail.com

Annan Naidu
Dept. of Computer Science Engineering
Centurion University
Bhubaneswar, India
annannaidu@cutm.ac.in

Abstract: Data mining has made broad and significant progress since its early beginnings. Today data mining is used in a vast array of areas, and numerous commercial data mining system that are available. There are many data mining systems and research prototypes to choose from. When selecting a data mining product that is appropriate for one's task, it is important to consider various features of data mining systems from a multidimensional point of view. Researchers have been striving to build theoretical foundations for data mining. Various clustering techniques have been used for identifying and visualizing trends in multi-attribute transactional data (e.g., hierarchical clustering techniques). In this paper, in order to compute distances (similarities) between the new cluster and each of the old clusters, complete-link clustering has been used.

Key Word: Data mining, cluster, clustering, hierarchical clustering, complete-link clustering.

I. INTRODUCTION

The diversity of data, data mining tasks, and data mining approaches possess many challenging research issues in the field of data mining. The development of effective and efficient data mining methods, system and services, and interactive and integrated data mining environments is the key area of study. Identifying temporal relationships (e.g., trends) in data constitutes an important problem that is relevant in many business environments, and the data mining literature has provided analytical techniques for some specialized types of temporal data. The research field of data mining has developed enormous methods for identifying patterns in data in order to provide insights to users.

The ability to identify trends in general temporal data [1] can provide significant benefits, such as competitive advantages to a firm performing forecasts or making decisions on future investments and strategies however, current temporal analytical techniques do not provide rich visualization of such trends.

Early data mining applications put a lot of effort into helping businesses gain a competitive edge. The exploration of data mining for businesses continues to expand as e-commerce and e-marketing have became main stream in the retail industry. Data mining is increasingly used for the exploration of applications in other areas such as web and text analysis, financial analysis, industry, government, biomedicine, and science. Because generic data mining system may have limitations in dealing with application-specific problems, we may see a trend towards the development of more application-specific data mining systems and tools, as well as invisible data mining function embedded in various kinds of services. In the business environment, data which are collected by firms and organization are multi-attribute and they are temporal in nature. Data mining techniques are often used to mine temporal data, which is a very complex task. Many new data analysis and visualization techniques have been proposed for representation of multi-attribute temporal data in a graphical manner. Cluster based approaches has been undertaken to implement the temporal cluster graph.

Visualizing and analyzing the multi-attribute data are extremely difficult, because it consists of numerous attribute. We can overcome this issue by mining the data according to specific time periods and then compare the data mining results across time periods to discover similarities. It is a clustering approach for discovering temporal patterns, which builds on temporal clustering methods and complement existing temporal data mining approaches. This technique can be applied in wide variety of data analysis.

C-TREND, *Cluster-based Temporal Representation of Event Data*, is a new method for discovering and visualizing, trends and temporal patterns in transactional attribute data that are build on the basis of standard data mining clustering techniques. Particularly, C-TREND separates data into user-defined partitions based on time periods and then identifies clusters of the dominant transaction types occurring within each partition.

Clusters are then compared to the clusters in adjacent time periods to identify cross-period similarities and, over many time periods, trends are identified. Trends are presented in an output graph that uses nodes to represent dominant transaction types and edges to represent cross-time relationships.

C-TREND provides three advantages over existing techniques. First, C-TREND presents temporal data in a unique and intuitive manner that emphasizes trends between dominant transaction types over time, and its output graphs resemble evolutionary diagrams and naturally portray the changes in data characteristics over time. Second, C-TREND is a meta-analysis tool for data mining results (specifically, hierarchical clustering) and, therefore, is designed to provide the domain expert with substantial control over the data presentation. In particular, CTREND provides the user with the ability to adjust all key parameters for creating output trend graphs, which allows a domain expert to visualize the data in a manner that provides the most value. Third, C-TREND presents a set of graph statistics.

II. LITERATURE REVIEW

A. C-Trend is useful for transactional data by considering

Clustering: A computer cluster is a group of linked computers, working together closely so that in many respects they form a single computer. The components of a cluster are commonly, but not always, connected to each other through fast local area networks. Clusters are usually deployed to improve performance and availability over that provided by a single computer, while typically being much more cost-effective than single computers of comparable speed or availability.

Data mining: Data mining is the process of sorting through large amounts of data and picking out relevant information. It is usually used by business intelligence organizations, and financial analysts, but is increasingly being used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data and "the science of extracting useful information from large data sets or databases " Data mining in relation to enterprise resource planning is the statistical and logical analysis of large sets of transaction data, looking for patterns that can aid decision making.

Temporal data: Temporal data mining is a single step in the process of knowledge discovery in temporal databases that enumerates structures (temporal patterns or models) over the temporal data. Temporal data mining is concerned with the analysis of temporal data and for finding temporal patterns and regularities in sets of temporal data. Also temporal data mining techniques allow for the possibility of computer-driven, automatic exploration of the data.

Trend analysis: The term "trend analysis" refers to the concept of collecting information and attempting to spot a pattern, or trend, in the information. In some fields of study, the term "trend analysis" has more formally-defined meanings. The data mining techniques uses clusters identified in multiple time periods and identifies trends based on similarities between clusters over time. It is a clustering approach for discovering temporal patterns, which builds on temporal clustering methods. Trend analysis decomposes time-series data into trend (long-term) movements, cyclic movements, seasonal movements(which are systematic or calendar related), and irregular movements (due to random or chance events).

Data Visualization: Visual data mining integrates data mining and data visualization to discover implicit and useful knowledge from large data sets. Visual data mining includes data visualization, data mining results visualization, data mining process visualization and interactive visual data mining. Data visualization is the process of presenting data in some visual form and allowing the human to interact with the data.

Visual data mining: It integrates data mining and data visualization to discover implicit and useful knowledge from large data sets. Visual data mining includes data visualization, data mining result visualization, data mining process visualization and interactive visual data mining.

B. *Hierarchical clustering***:** Hierarchical clustering is one of the visualization techniques partition all dimensions into subsets (i.e subspaces). The subspaces are visualized in a hierarchical manner.

C. *Dendrogram***:** A Dendrogram is a tree-structured graph used in heat maps to visualize the result of a hierarchical clustering calculation. The result of a clustering is presented either as the distance or the similarity between the clustered rows or columns depending on the selected distance measure.

D. *D.Temporal cluster graph***:** To proceed with the temporal cluster graph firstly, transactional data set is to be partitioned with respect to time and then clustering technique is applied. The temporal cluster graph is a directed graph that consists of set of nodes and directed edges.

III. RELATED WORKS

In [6] the authors have focused on the data mining capabilities of the various visualization techniques. Visual data mining techniques are useful for the exploration and analysis of large databases to find interesting data cluster and their properties. Here, database query and information retrieval techniques are combined with new types of visualization techniques. It uses five visualization techniques-

Pixel -oriented techniques which map each data value to a colored pixel and present the data values belonging to one attribute in separate windows and are divided into two types- Query independent pixel-oriented technique which sort the data according to some attributes and uses a screen filling pattern to arrange the data values on the display and Query dependent pixel-oriented techniques which visualizes the relevance of the data items with respect to the query, to give the users the feedback on their queries and to direct them to their search.

Geometric projection technique which is the second technique aiming at finding interesting projection of multidimensional data set and the central challenge it try to address is how to visualize a high-dimensional space on a two dimensional display. To visualize n-dimensional data points, the parallel coordinate techniques draws n equally spaced axes, one for each dimension, parallel to one of the display axes. The data record is represented by a polygonal line that intersects each axis at the point corresponding to the associated dimension value. It reveals a wide range of data characteristics i.e different functional dependencies and data distributions. The major limitation of the parallel coordinate technique is that it cannot effectively show a data set of many records. Even for a data set of several thousand records, visual cluster and overlap often reduce the readability of the visualization and make the pattern hard to find.

Icon based technique: Used to map each multi dimensional data item to an icon and subdivided into Chernoff face visualization technique which is the first approach of iconic display which gives two dimensional representations of properties of face icon which includes eyes, nose, mouth, the shape of face itself i.e it make use of the ability of the human mind to recognize small differences in facial characteristics and to assimilate many facial characteristics at once. So, by condensing the data, Chernoff faces make the data easier for users to digest. In this way, they facilitate visualization of regularities and irregularities present in the data, although their power in relating multiple relationship is limited.

Stick figure technique: It is somewhat similar to an icon display technique which maps multidimensional data to five piece stick figure, where each figure has four limbs and a body. Both the approaches show the common disadvantage that a number of data items that can be visualized are quite limited.

Hierarchical based technique: This technique subdivide the k-dimensional space and present the subspaces in hierarchical fashion. It focus on visualizing multivariate functions and therefore, not particularly interesting for data mining.

Graph based technique: It effectively present a large graph using layout algorithm, query language and abstraction techniques.

A central goal of paper [6] is to evaluate and compare the above visualizing techniques which may be used for visualizing large databases. Two demerits has arises i.e firstly, comparisons had been made by taking stick figure and parallel coordinate technique which is useful for visualizing a wide range of data at one point of time but overlapping of data may prevent visualization and secondly, secondary storage based version of the VisDB system is implemented as they support high transaction rates and a fast search of specific data items, but most of them do not provide a sufficient performance for range queries on multi attributes. So, solution is to use multidimensional data structure. But still there is a need to determine which visualization technique is most appropriate for specific types of correlations, clusters and functional dependencies and which multi dimensional data structure is best.

The research field of data mining has developed a number of methods for identifying patterns in data to provide insights and decision support to users [1][3][4]. Data mining and business intelligence approaches are often used for class identification and data visualization in knowledge management systems. . In the business environment, data which are collected by firms and organization are multi-attribute and they are temporal in nature. Data mining techniques are often used to mine temporal data, which is a very complex task. In this, a new data analysis and visualization technique has been proposed for representation of multi-attribute temporal data in a graphical manner. Cluster based approaches has been undertaken to implement the temporal cluster graph. Visualizing and analyzing the multi-attribute data are extremely difficult, because it consists of numerous attribute. We can overcome this issue by mining the data according to specific time periods and then compare the data mining results across time periods to discover similarities. Multidimensional technique is used as a type of data to show the visual graphic of multidimensional data.

Visualization approaches [1] are- Hierarchical techniques which subdivide the multidimensional space and create a hierarchical decomposition of set of data, Graph-based techniques which generates large graphs using layout algorithm, Abstraction techniques which gives relational meanings clearly and quickly, Interaction technique which provides the user with the ability to dynamically change visual representations, Interactive filtering which dynamically focuses on the partitioning of a data set into segments and focusing on interesting subset, Interactive zooming which provides the user with a variable display of data at different levels of analysis. Then to proceed with the temporal cluster graph firstly, transactional data set is to be partitioned with respect to time and then clustering technique is applied. The temporal cluster graph is a directed graph that consists of set of nodes and directed edges. Distance between each node can be calculated by

Euclidean Formulae:

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)} \quad (1)$$

Manhattan Formulae:

$$d(x,y) = \sum_{i=1}^{m} \left| x_i - y_i \right| \quad (2)$$

So far all the algorithms were using matrices to produce partition and distance between the matrices is used for calculations. But then, DENDROGRAM Data structure has been proposed for storing and extracting cluster solutions generated by hierarchical clustering algorithms. Now calculations are made using Tree structure. The main contribution of paper [1] is the development of a novel and useful approach for visualization and analysis of multi-attribute transactional data based on a new temporal cluster graph construct, as well as the implementation of this approach as the Cluster-based Temporal Representation of EveNt Data (C-TREND) system. C-TREND utilizes optimized Dendrogram data structures for storing and extracting cluster solutions generated by hierarchical clustering algorithms. C-TREND is the system implementation of the temporal cluster graph-based trend identification and visualization technique; it provides an end user with the ability to generate graphs from data and adjust the graph parameters. C-TREND consists of two main phases: 1) offline preprocessing of the data and 2) online interactive analysis and graph rendering. In the preprocessing phase, the data set is partitioned based on time periods, and each partition is clustered using one of many traditional clustering techniques such as a hierarchical approach. The results of the clustering for each partition are used to generate two data structures: the node list and the edge list. The time partition size is set exogenously by the user and stays constant throughout preprocessing and online interactive analysis.

Dendrogram tree constructed by collecting the data and by using Hierarchical clustering techniques [3] - Single-link clustering in which the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster is considered. In Complete-link clustering (also called the diameter or maximum method), the distance between one cluster to the longest distance from any member of one cluster to any member of the other cluster and in Average-link clustering the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. Hierarchical clustering is called agglomerative because it merges clustering iteratively. As in agglomerative procedure, the clusters are initially the singletons (single-member clusters). At each stage the individuals or groups of individuals that are closest according to the linkage criterion are joined to form a new, larger cluster (i.e it is a bottom-up strategy which starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects. At the last stage, a single group consisting of all individuals is formed. Paper [4] presents a cluster-based time-series representation of data that implements the time-series cluster graph which maps multi-attribute time-series data to two dimensional directed graphs. Here, trend discovery is addressed by unsupervised learning technique i.e learning without training data (i.e a sample from the data source with the correct classification).

The method of hierarchical clustering is using single-link clustering. As hierarchical clustering is classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. That is, if a particular merge or split decision later turns out to have been a poor choice, the method cannot backtrack and correct it. One promising direction for improving the clustering quality is to integrate hierarchical clustering with other clustering techniques, resulting in multiple-phase clustering.

IV. CONCLUSION

In previous papers, a C-TREND technique is used for identifying & visualizing multi-attribute transactional data using- Hierarchical clustering, Dendrogram tree, Cluster graph, Time-series with cluster graph. The above all techniques are using only single-link clustering for identifying & visualizing multi-attribute transactional data, but this paper uses complete-link clustering. In cluster analysis, complete linkage or farthest neighbour is a method of calculating distances between clusters in agglomerative hierarchical clustering. In complete linkage [3] the distance between two clusters is computed as the maximum distance between a pair of objects, one in one cluster, and one in the other. Complete linkage clustering avoids a drawback of the alternative single linkage [3] method - the so-called chaining phenomenon, where clusters formed via single linkage clustering may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant to each other. Complete linkage tends to find compact clusters of approximately equal diameters.

REFERENCES

[1] Adomavicius Gediminas, Bockstedt Jesse "*C-TREND: Temporal Cluster Graphs for Identifying and Visualizing Trends in Multiattribute Transactional Data*", IEEE transaction on knowledge and data engineering, vol. 20, no.6,pp.721-733, June 2008.

[2] Ahmed Riaz Syed, "*application of data mining in retail business*", Proc. International conference on information technology: coding and computing, April 2004.

[3] Rani Radha D., Bharati Vini A., Sravani A., "*Analysis of Dendrogram Tree for Identifying and Visualizing trends in Multi attribute Transactional Data*", International Journal of Engineering Trends and Technology, vol. 3, pp. 14-18, [2012].

[4] Sridath Phani V.R, Srinivas Kudipudi , Rao Srinivasa V., "*Visualization of Time-Series Cluster Graphs Using Hierarchical Clustering Techniques*", International Journal of Engineering Sciences and Technologies, vol.7, pp.65-69, [2011].

[5] Ruey-shun Chen, Ruey-chyi Wu and J. Y. Chen, "*Data Mining Application In Customer Relationship Management Of Credit Card Business*", Proceedings Of The 29th Annual International Computer Software And Applications Conference, Pp.1-2, Taiwan [2005].

[6] Daniel A. Keim and Hans-peter Kriegel ,"*Visualization Techniques For Mining Large Databases: A Comparison*", IEEE Transactions On Knowledge And Data Engineering, Vol. 8, No. 6, pp.923-938, December 1996.

[7] Jia-dong Ren, Jie Bao, Hui-w Huang, "*The Research On Spatio-temporal Data Model And Related Data Mining*", proceedings Of The Second International Conference On Machine Learning And Cybernetics,, P.37.-40,November 2003.

[8] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", Elsevier Inc, 2006.

[9] Ying Peng,Yongyi Ma, Huairong Shen,"Clustering Belief Functions using Agglomerative Algorithm",Information Engineering and Computer Science(ICIECS),pp. 1-4,[2010].

[10] Takumi, satoshi,miyamoto,sadaaki,"Top-Down Vs Bottom-Up methods of linkage for asymmetric Agglomerative Hierarchical clustering",Granular computing(GrC),pp. 459-464,[2012].

[11] Srinivas M, Mohan C.K,"Efficient clustering Approach using incremental and hierarchical clustering methods",neural Networks(IJ CNN),international Joint conference Publication,pp.1-7,[2010].

[12] Hakim R.B.F., Subanar, winarko E.,"Reducing Dendrogram Instability of Features using Rough Set indiscernibility Level",Distributed Framework and Applications,pp. 1-10,[2010]