
Visual and analytical mining of transactions data for production planning for production planning and marketing

Gurdal Ertek, Can Kuruca, Cenk Aydin,

Besim Ferit Erel, Harun Dogan, Mustafa Duman,

Mete Ocal, Zeynep Damla Ok

Sabanci University

Istanbul, Turkey

Proceedings of 4th International Symposium on
Intelligent Manufacturing Systems, September 6-8, 2004: 848-859

Visual and analytical mining of sales transaction data for production planning and marketing

Gurdal Ertek*, Can Kuruca, Cenk Aydin, Besim Ferit Erel, Harun Dogan, Mustafa Duman, Mete Ocal, Zeynep Damla Ok

(*) Corresponding author

**Sabanci University, Faculty of Engineering and Natural Sciences,
Orhanli, Tuzla, 34956, Istanbul, Turkey**

**Tel: +90(216)483-9568
Fax: +90(216)483-9550
Email: ertekg@sabanciuniv.edu**

Visual and analytical mining of transactions data for production planning and marketing

Gurdal Ertek, Can Kuruca, Cenk Aydin, Besim Ferit Erel, Harun Dogan, Mustafa Duman, Mete Ocal, Zeynep Damla Ok

Abstract

Recent developments in information technology paved the way for the collection of large amounts of data pertaining to various aspects of an enterprise. The greatest challenge faced in processing these massive amounts of raw data gathered turns out to be the effective management of data with the ultimate purpose of deriving necessary and meaningful information out of it. The following paper presents an attempt to illustrate the combination of visual and analytical data mining techniques for planning of marketing and production activities. The primary phases of the proposed framework consist of filtering, clustering and comparison steps implemented using interactive pie charts, K-Means algorithm and parallel coordinate plots respectively. A prototype decision support system is developed and a sample analysis session is conducted to demonstrate the applicability of the framework.

Submission areas: Decision support systems, data mining, information technologies

Introduction

Widespread use of information technology has resulted in massive collections of data regarding most aspects of an enterprise. The amount of data on the marketing side has exploded due to widespread usage of barcode systems, accounting and Enterprise Resorce Planning (ERP) software and also due to collection of Business-to-Consumer (B2C) and Business-to-Business (B2B) electronic commerce data. The amount of data that comes from manufacturing processes has also exploded, due to application of Computer Integrated Manufacturing (CIM) systems, barcode and radio frequency technology, which provide bulky amounts of real time data.

Effective collection, management, reporting, interpretive analysis and mining of enterprise data can help in establishing effective control of manufacturing activities, achieving effective production planning and increased sales, and consequently increasing the firm's profitability. Data mining can also serve the purpose of increasing customer satisfaction by offering and timely delivering them products that they are willing to purchase. Keeping existing customers is typically much more profitable than trying to acquire new customers. This observation is indeed the underlying concept of Customer Relationship Management (CRM) systems (Shaw et al., 2001).

As suggested earlier, data from various areas of an enterprise is now widely available; however, in this paper, only sales transactions data is considered. The reason for choosing this particular data type is that sales transactions data is collected and archived in almost every firm and it is actually the essential input to two very critical aspects of enterprise planning, namely marketing and production planning. A framework for the analysis of this type of data is proposed and implemented in the software developed, namely CuReMa. The main contribution of the framework and the prototype software is the integration of visual and analytical data mining techniques for marketing and production planning. Kreuseler and Schumann (2002) present a similar approach combining visual and analytical data mining techniques without focusing on particular enterprise data.

The paper starts by offering a brief review of the literature concerned with visual and analytical data mining techniques. In the following sections, we present the framework we propose and explain how it is implemented in CuReMa. Before concluding, the applicability of the proposed framework is demonstrated with a sample analysis session with the software.

Literature Review

Analytical data mining techniques are widely used and implemented (Han and Kamber, 2001). In recent years, visual mining of data has also gained importance. The traditional exploratory graphical data analysis methods such as scatter plots, box plots (Chambers et al., 1983) have been enriched with a wide array of new visual representations and methods (de Oliveira and Levkowitz, 2003). The field of computer science involved in such representations and methods is referred to as *information visualization*. The number of journal articles in information visualization has shown significant increase from 1990's to present (Chen, 2002), indicating that the field is a promising branch of computer science.

Work related to this paper can be grouped in two categories, based on their scope:

a) *Data mining for marketing:*

Shaw et al. (2001) provide a recent survey of data mining applications for marketing. They present a taxonomy of data mining tasks and provide five broad categories: Dependency analysis, class identification, concept description, deviation detection, and data visualization. Our study suggests a framework that relates to all but the first of these categories.

Keim et al. (2002) develop a visualization technique named “pixel bar charts” for analysis of very large multi-attribute data sets and demonstrate applicability of their approach by analyzing real-world e-commerce data sets.

b) *Data mining for manufacturing:*

Applications of information visualization in the domain of manufacturing include data representation for engineering design and data analysis for predicting product failure rates (Spence, 2001, p28—30 and p60—61, respectively).

Dabbas and Chen (2001) present an integrated relational database approach for semiconductor wafer fabrication, which resulted in improved manufacturing performance. They describe how they combined multiple data sources and various reports under the integrated approach.

Data mining tools that can handle fairly large amounts of data and allow derivation of insights are numerous. Spotfire, Miner3D and XLMiner[1] are among successful commercial products that can be used to analyze data from a variety of domains. Software libraries which allow building customized visual interfaces to domain-specific applications are also available: For instance, Eick (2000) presents ADVIZOR as “a flexible software environment for visual information discovery” that allows creating visual query and analysis applications. Our implementation has some similarities to these products: For example, one of the similarities between CuReMa, Spotfire and Miner3D is that all of them allow the user to conduct a query on products or customers within a given range through the use of sliders.

Proposed Framework

An approach that integrates visual data mining methods with analytical methods for mining sales transaction data is proposed to perform the three critical functions listed below, and to answer the given questions as well as many other unlisted ones:

1) Filtering

This initial step includes filtering of data in different dimensions and displaying and analyzing the filtered data. In this stage of the analysis, the products or the customers within a cluster or a group are represented with pie charts. This representation allows the identification of significant items and outliers in the cluster or the group. The answers to the following questions can be obtained as a result of filtering:

- What are the total sales of top-selling *n* products for the top-purchasing *c* customers within time interval (t_1, t_2) and what is the share of each product?
- What are the total sales of *n* slowest-moving products for a given set of customers (selected from a list sorted based on sales) within time interval (t_1, t_2) and what is the share of each product?
- How much sales were generated for each product represented in the pie chart?

[1] The information regarding these products can be found on http://www.spotfire.com, http://www.miner3d.com, and http://www.resample.com/xlminer/ respectively.

- What are the total purchases of top-purchasing *c* customers from a given set of products (selected from a list sorted based on sales) within time interval (t_1, t_2), and what is the share of each customer?
- How many purchases were made by each of the customers represented in the pie chart?

2) Clustering

The second step involves clustering products and customers with respect to selected measures and generating related reports. This phase provides the answer for the questions such as:

- How can the products be grouped with respect to seasonal sales patterns?
- How can the market be segmented with respect to seasonal purchasing patterns of customers?

Many other useful statistics about each cluster may be acquired after clustering. These statistics include the revenue generated by each product, the product with the highest sales level, and the product purchased by the greatest number of customers.

3) Comparison

Finally, clusters are compared with respect to selected measures. One of the questions answered through comparison would be the following:

- How do the products differ from each other with respect to seasonal sales patterns? Are there significant differences among given clusters?

The questions above focus on customer and product clusters, yet CuReMa can answer the same questions based on customer and product groups, which could come with the dataset. The sample analysis session will demonstrate how some of these questions are answered using CuReMa.

An Implementation of the Framework: CuReMa

The data mining framework has been implemented as a prototype decision support system to demonstrate the viability of the proposed approach. Real world data from a regional distributor of automotive spare parts, covering approximately 18 months of sales transactions, has been analyzed with the developed software. The software, named CuReMa after Customer Relationship Management, allows mining of the dataset for marketing and production planning purposes.

The implementation has been done in Java programming language (Doke et al., 2002) using the Eclipse Integrated Development Environment (IDE)[2] under Microsoft Windows operating system. CuReMa was built based on a 3-tier design, separating data access, interface and business classes (Doke et at. 2002). MySQL is employed as the database server and MySQL Control Center[3] is used in constructing and maintaining the database. Java and MySQL were selected primarily due to their platform independence, which allows porting the developed system to various operating systems, such as Linux and MacOS. Java language also has the advantages of being purely object oriented and having extensive libraries that allow rapid prototyping. Being interpreted at runtime -as opposed to being executed as native code- makes Java programs run

[2] The information on Eclipse IDE can be retrieved from http://www.eclipse.org
[3] MySQL Control Center can be accessed through http://www.mysql.com/products/mysqlcc/

slower than programs written in some other languages. However, implementation of distributed shared memory and remote method invocation (not implemented in CuReMa) can enable programmers to build scalable Java programs (Kielmann et al., 2001).

The relational database in MySQL contains a number of tables. The table “transactions”, (shown in Figure 1) contains the sales transactions data and forms the source of all the other tables.

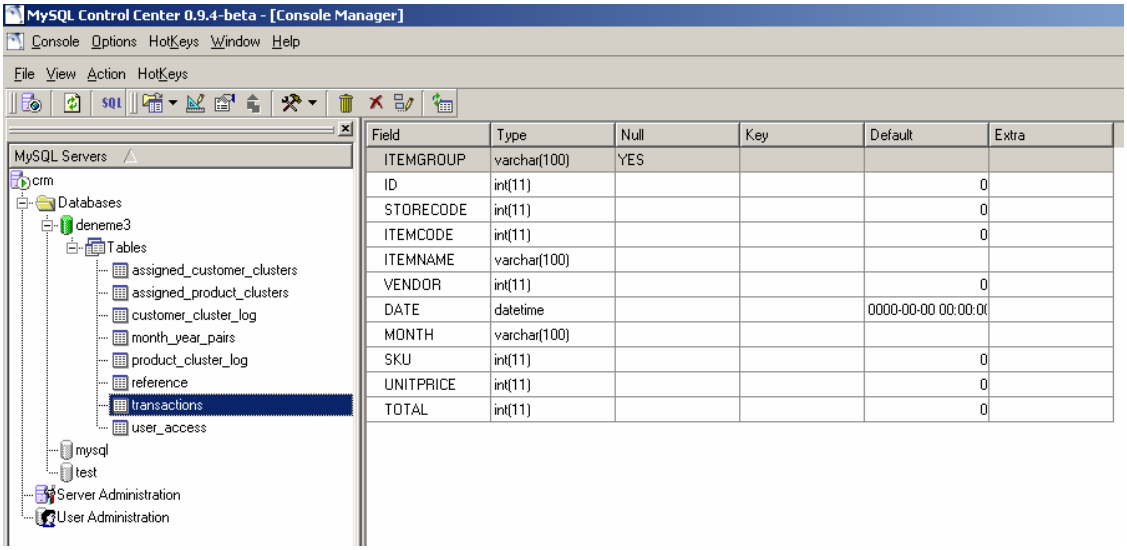

Figure 1. Snapshot of the database, illustrating the fields of the “transactions” table

We now present how each function proposed in the framework is implemented:

Filtering: Interactive Visual Querying

In CuReMa, the user can perform visual query of the dataset using pull down menus and sliders. The sliders enable filtering by time, by customer and by product name. The customers and the products are sorted in decreasing order of total sales from the left to the right. Positions of the sliders are translated into SQL (Structured Query Language) (Elmasri and Navathe, 1994, Ch7) statements and reflected on the pie charts in real time. One such statement (that corresponds to the filtering illustrated in Figure 2) is the following:

```
SELECT ItemName AS NAME,
SUM(TOTAL) AS TOT FROM deneme3.reference
WHERE DATE <='2003.6.6' AND DATE >= '2003.3.5'
AND CustomerNo>= 1 AND 173 >= CustomerNo
AND 1 <= ItemNo AND 259 >= ItemNo
GROUP BY NAME ORDER BY TOT DESC
```

Pie charts display the share of each customer/product within the selected ranges.

Clustering: Analytical Data Mining

The analytical data mining method implemented in CuReMa is K-Means clustering (Han and Kamber, 2001, Ch8). This method starts with a set of entities, accompanied with given attribute

values, and proceeds by clustering the entities into a specified number of clusters. The attribute values are the average sales in each month of the year.

K- Means algorithm initially selects random values as the K-cluster means. Then, it examines each object pertaining to the dataset and assigns the object to the closest cluster possible, whose proximity is measured by the aggregate distance between the cluster mean and the object's actual values. An epoch consists of N iterations where N is the number of elements in the dataset. As the new objects are added to the clusters for various epochs, the means are dynamically updated. If the total number of moves of all data objects from a cluster to another is zero at the end of an epoch, the algorithm ceases and the partitioning is completed.

The algorithm may end up with fewer clusters than specified: For instance, while exploration of the automotive spare parts dataset in CuReMa, it has been observed that a customer clustering run with a desired number of 7 clusters resulted in only 3 clusters.

Comparison: Visual Data Mining

Parallel coordinate plots (Inselberg and Dimsdale, 1990) are used for comparison of product/customer clusters, and are implemented in Mondrian and XMDV[4] software. Parallel coordinate plot “maps a k-dimensional data or object space onto the 2D display by drawing k equally spaced axes in parallel” (de Oliveira and Levkowitz, 2003). In the plot, each axis corresponds to an attribute, and each line corresponds to an element of the dataset. In CuReMa, customer purchases and product sales can be compared in a parallel coordinate plot based on monthly sales.

Sample Analysis Session

In this section we present how the user can interact with CuReMa in an analysis session.

[4] Mondrian and XMDV can be accessed from http://www.theusrus.de/Mondrian/index.html and http://davis.wpi.edu/~xmdv/ respectively.

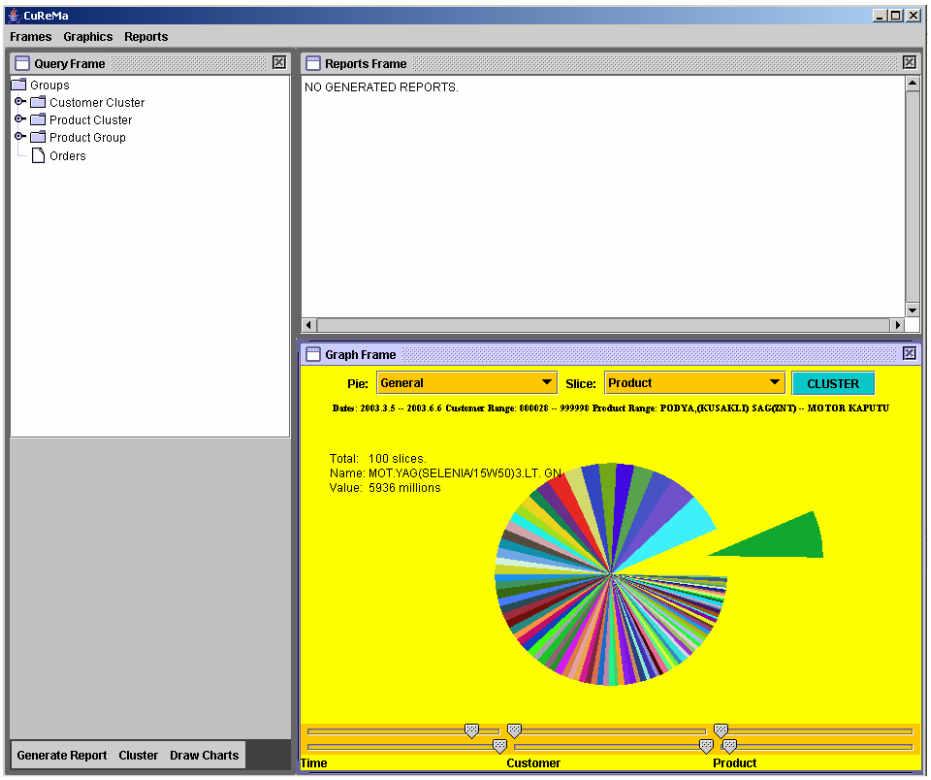

Figure 2. Filtering of products

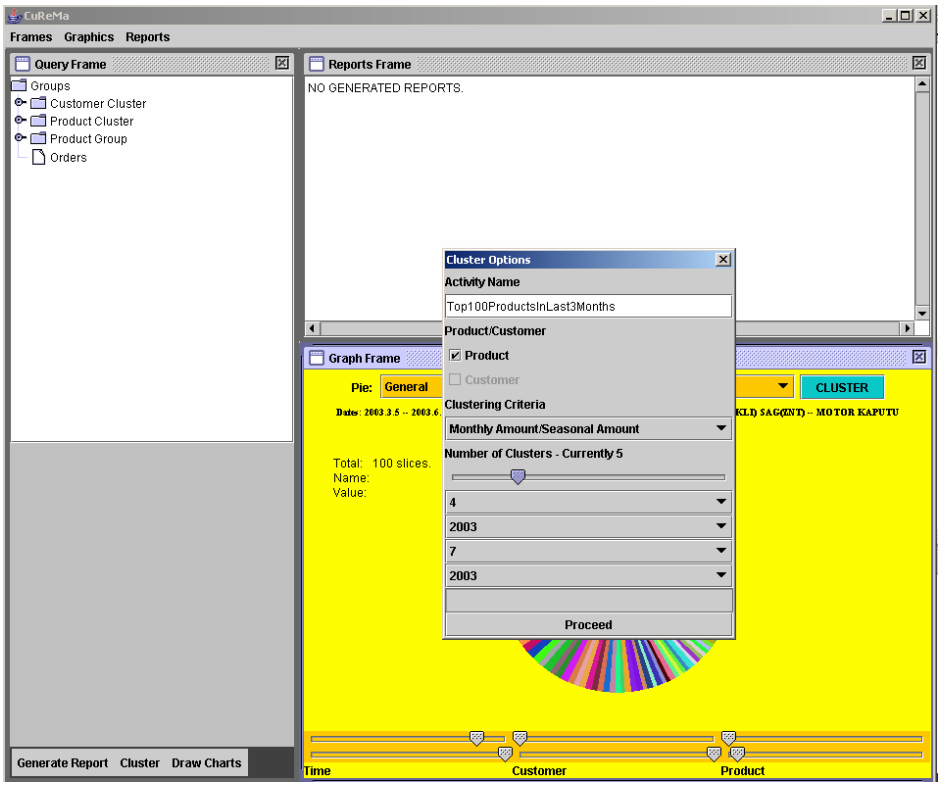

Figure 3. Clustering of products

The analysis begins with Filtering, where the user selects slicing based on products from the pull-down menus in the Graph Frame (Figure 2). The user selects the data for the last 3 months for all customers (with sales in the last 3 months) and top-selling 100 products from the sliders. The top-selling product is observed to be “MOT.YAG” (motor oil) by selecting the largest slice in the pie chart.

Next, the user performs clustering of the products from the previous analysis by clicking the “Cluster” button on the Graph Frame (Figure 3). Since products are displayed, the check box next to Product is selected. The user also selects clustering criteria, number of clusters, and the time range.

Clustering is completed in approximately 2 minutes on a PC with Intel Pentium IIIE 933 MHz processor and 384 MB of RAM. Then the user selects the item named “Top100ProductsInLast3Months” in the Query Frame, “Parallel Coordinate Plot” from Graphics menu and clicks “Draw Charts” button to obtain the plot in Figure 4. In this plot, products in clusters 2 and 3 are observed to exhibit almost constant sales throughout the selected months whereas products in cluster 1 have the largest amount of sales, but also a greater variance as compared to others. A report can be generated inside Reports Frame by clicking “Generate Report” button in the left bottom corner.

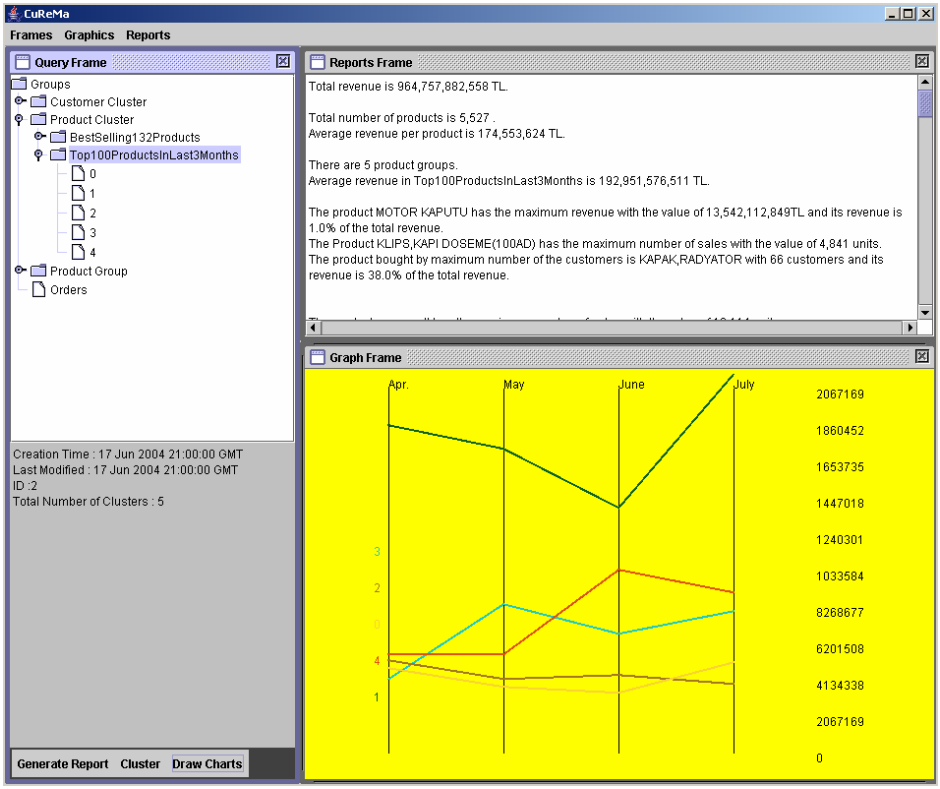

Figure 4. Comparison of product clusters

The user can also analyze customers, clustering them into similar-behaving groups. Figure 5 offers the result of one such clustering. In this stage of the analysis, all the customers are selected at all the times, and clustering is performed. It can be observed that cluster 1 shows

stable sales, with minimal variability. Trying to increase the sales volume to these customers can be a profitable strategy, since these customers do not cause big fluctuations in production schedules. Clusters 2 and 3 are observed to have very high levels of sales in March, while clusters 0 and 4 exhibit high sales volumes in August and September respectively. These fluctuations should be further investigated by examining the reports generated and by filtering the sales transaction data for these months.

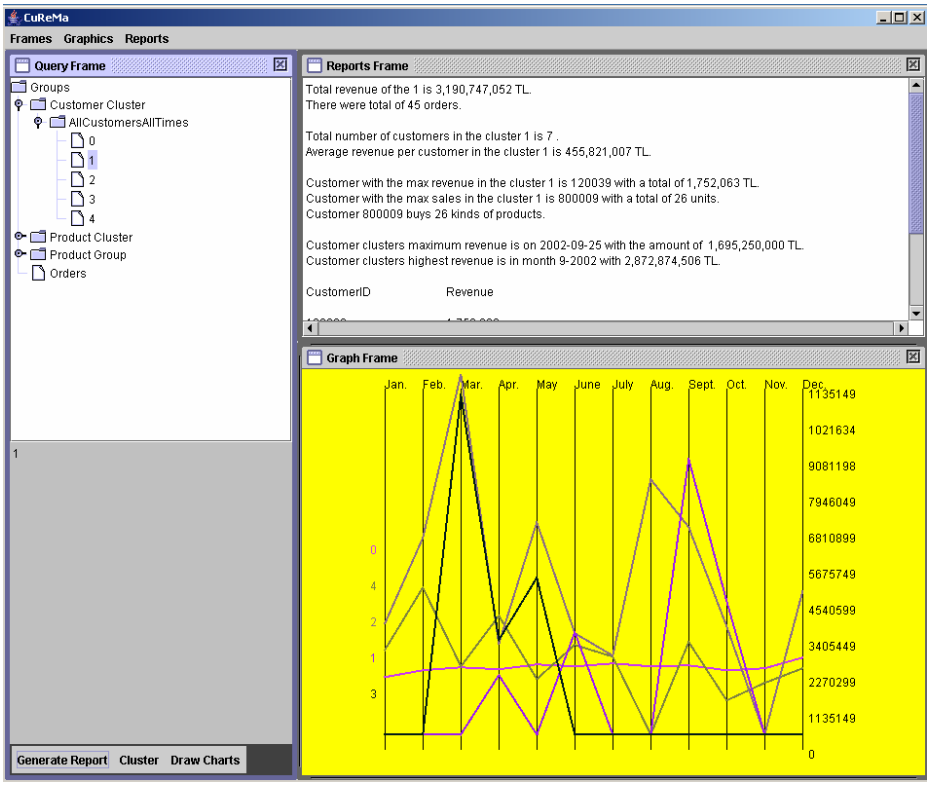

Figure 5. Comparison of customer clusters

Conclusions and Future Work

In this paper a framework is presented for the analysis of sales transaction data using visual and analytical data mining techniques. The framework suggests applying filtering, clustering and comparison through interactive pie charts, K-Means method and parallel coordinate plots, respectively. The framework is implemented in a software to demonstrate how the analysis can be carried out. A sample session is exhibited.

The framework proposed in this paper can be extended to answer a greater range of questions. Visual metaphors reported in information visualization survey papers (e.g. de Oliveira and Levkowitz, 2003) can be adapted to the current framework, allowing observation and query of other aspects regarding the data. From an analytical point of view, other data mining techniques such as deriving association rules for revealing patterns (Changchien and Lu, 2001) and performing cross-tabulation analysis (Verhoef et al., 2002) can be made part of the framework and implemented in CuReMa.

CuReMa can be tested and used with a variety of transactional datasets coming from different industries (e.g. from an e-commerce website) and the scope of the program can be broadened to include other data fields, such as ages and income levels of customers, and marginal profits of the products. On the methodological side, clustering algorithms other than K-Means can be built into the software and their performance can be compared to the performance of K-Means. One such method is SOM (Self Organized Maps), which is implemented in coherence within a visual data mining framework by Kreuseler and Schumann (2002). The clustering function in the software can be extended by taking into consideration other attributes besides monthly sales averages. One such attribute can be recency of purchases by customers and by products: Verhoef et al. (2002) report that recency and frequency of purchases are among the most popular variables used for clustering while carrying out market segmentation analysis.

Data analyzed can have domain-specific considerations. For example, in many countries around the world, inflation is a real-world fact that cannot be neglected if the analyst attempts to perform a valid and insightful analysis. Adding tables into the database that keep track of the inflation indices on a monthly basis and adjusting calculations based on these indices would be pretty valuable for the dataset analyzed in such a study. Another domain-specific consideration would be the effects of holidays. One could take into account religious holidays in Turkey which turn out to be around ten days earlier from the previous year. This and similar considerations pose significant challenges.

One possible area of future work is to investigate the implemented software from the point of view of human-computer interaction (Shneiderman, 1998). One can investigate the performance of various visual components and their arrangements by formal usability tests on groups of users and analyze results.

Acknowledgement

The authors would like to thank Ender Yalcin for providing the sales transaction data and for explaining the processes involved in decision making. The authors would also like to thank Selim Balcisoy, Pelin Gulsah Canbolat and the anonymous referees for their valuable suggestions and remarks.

References

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole, Pacific Grove, California.

Changchien, S. W. and Lu, T. (2001) Mining association rules procedure to support on-line recommendation by customers and products fragmentation. *Expert Systems with Applications*, **20**, 325--335.

Chen, C. (2002) Information visualization. *Information Visualization*, **1**, 1--4.

Dabbas, R. M. and Chen, H. (2001) Mining semiconductor manufacturing data for productivity improvement – an integrated relational database approach. *Computers in Industry*, **45**, 29--44.

De Oliveira, M. C. F. and Levkowitz, H. (2003) From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, **9**, 378--394.

Doke, E. R., Satzinger, J. W. and Williams, S. R. (2002) *Object-oriented application development using Java*, Course Technology, Boston, Massachusetts.

Eick, S. G. (2000) Visual discovery and analysis. *IEEE Transactions on Visualization and Computer Graphics*, **6**, 44--58.

Elmasri, R. and Navathe, S. (1994) *Fundamentals of Database Systems*, Addison-Wesley, Menlo Park, California.

Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco.

Inselberg, A. and Dimsdale, B. (1990) Parallel Coordinates: A tool for visualizing multidimensional geometry. *Proc. IEEE Visualization '90*, 361--375.

Keim, D. A., Hao, M. C., Dayal, U. and Hsu, M. (2002) Pixel bar charts: a visualization technique for very large multi-attribute data sets. *Information Visualization*, **1**, 20—34.

Kielmann, T. K., Hatcher, P., Bouge, L. and Bal, H. E. (2001) Enabling Java for high-performance computing. *Communications of the ACM*, **44**, 110--117.

Kreuseler, M. and Schumann, H. (2002) A flexible approach for visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, **8**, 39--51.

Shaw, M. J., Subramaniam, C., Tan, G. W. and Welge, M. E. (2001) Knowledge management and data mining for marketing. *Decision Support Systems*, **31**, 127--137.

Shneiderman, B. (1998) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Addison Wesley Longman, Reading, Massachusetts.

Spence, R. (2001) *Information Visualization*, ACM Press, London.

Verhoef, P. C., Spring, P. N., Hoekstra, J. C. and Leeflang, P. S.H. (2002) The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision Support Systems*, **34**, 471--481.