Data Mining and Multidimensional Databases are the key to Data Warehouses and WWW

Michael Haller*, Georg Jenichl†, Josef Küng‡

8th July 1998

Abstract

With the permanent increase of data stored in databases and Data Warehouses it was necessary to invent new techniques for realizing queries on this data within an acceptable response time. The newly defined data model of **multidimensional databases** facilitates faster and easier analyses because of special designed structures. The ability to conceive the complex structures of the real world without any unnatural flattening makes the data modeling easier and offers much more performance for complex analysis.

Moreover **Data Mining** can be used to extract useful and previously unknown information from large databases and Data Warehouses. With Data Mining tools it is possible to analyze such data and mine interesting knowledge from this data.

Another revolutionary evolution in computer science was the **Internet**: the global net offers access to a huge amount of data and everybody can analyze the data. An integration of the three technologies, the multidimensional databases and Data Mining on the one hand and the internet on the other hand, is a logical conclusion. In this proceeding the first step is the combination of these technologies.

1 Introduction

Data Modeling for a Data Warehouse requires new modeling techniques and the creation of different types of schema than those used for traditional operational databases done with **OLTP**

*Institute for Applied Knowledge Processing (FAW), Johannes Kepler University of Linz, mhaller@faw.uni-linz.ac.at

†EURECOM Research Institute, Sophia Antipolis, jenichl@eurecom.fr

‡Institute for Applied Knowledge Processing (FAW), Johannes Kepler University of Linz, jkueng@faw.uni-linz.ac.at

(= Online Transaction Processing). In OLTP data was collected and stored for operations and control purposes. Nowadays, the Data Warehouse contains an application and a database designed and optimized for data analysis and reporting capabilities. An easy access to the enterprise data has to be supported for the guarantee that the user will not find it difficult to access and analyze the data correctly [McC97].
In Data Warehouses the user generally needs a multidimensional view for making decisions.
A multidimensional query tool allows multiple data views and it is easy to make ad hoc decisions and analyzes. Compared to operational, OLTP databases , there are important, fundamental differences in the way the Data Warehouse uses the enterprise data. OLAP databases require **read-only** access to the database and are used to create reports and query responses that assist in making important business decisions. Moreover the OLAP user is not interested in the moment to moment transactions, but in trends, summarizations and abstractions. Aggregations, summarizations and historical data are important and many can be calculated and stored in advance. The main goal of OLAP is a fast access, concurrence, and integrity.

2 Technologies for using a Data Warehouse

Data Mining is a set of techniques used in an automated approach to exhaustively explore and bring to the surface complex relationships in very large databases and Data Warehouses [B.96]. The goal is to uncover 'strategic competitive insight' to drive market share and profits, which is a cooperative effort between humans and computers. Humans describe goals, design databases and describe problems. Computers sift through this data and look for patterns that match these goals. After the computer has extracted the relationships the business analyst can select useful patterns. A pattern refers to any relation among elements of a database, i.e. records, attributes, and values. For instance a pattern can be an expression in a language that describes facts in the subset of data. E.g. if income $\leq$ t, then the person has defaulted on the loan would be an appropriate choice. The following figure illustrates this feature:
Each point on the graph represents a person who has been given a loan by a bank at a special time in the past. The dataset was classified into two classes: the rectangles represent persons who have defaulted on their loans, whereas the circles represent persons whose loans are in good status with the bank [UMF97]. This example demonstrates how difficult it is for a business analyst to decide whether or not to give a loan to a person. Therefore a business analyst has the following range of needs:
First he should understand what is happening in business. Then, he has understand the behavior of customers and markets. Finally there should be clear what can be done.

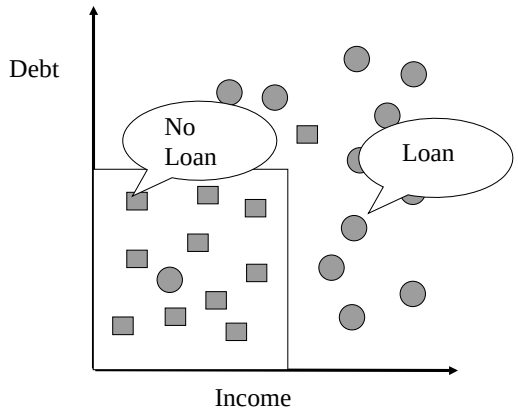

Figure 1: Threshold for classification

While the traditional query, reporting and multidimensional analysis focuses mainly on **what** is happening, the Data Mining technique is focusing the **why** part. It focuses the need to discover the why and then to predict and forecast possible actions. Of course, the prediction is performed with a certain confidence factor for each prediction. The following table will demonstrate how Data Mining differs from analytical processing [HS96]:

Feature	**Data Mining**	**Informational/Analytical Processing**
Analysis focus	Why is it happening ?	What is happening
Analysis technique	discover automatically	Slice and Dice
Analyst interaction	minimal guidance from analyst	business analyst initiated
Attribute number	about hundreds	Maximum 10
Confidence factor	derived from data	derived from business analyst
Dataset size	millions for each dimension	medium for each dimension
Dimensions	Many	limited
Focus	Transaction or detail data	summary data
Technology	Statistical Analysis/ knowledge discovery	mature

With the above mentioned techniques we can retrieve, manipulate, and analyze the data and then present the results with useful applications. these tools are used in two modes: **verification** and **discovery**.

With verification, the business user creates a hypothesis (e.g. business questions) and then tries to confirm his hypothesis by accessing the data in the Data Warehouse. Typical verification mode tools are query tools, reporting systems and multidimensional analysis tools. In the discovery mode, the tool tries to discover characteristics in the data, i.e. patterns and associations

that are not previously known or suspected by the business user. A typical discover mode tool is a **Data Mining tool**.
With **informational processing** we can perform tasks like data and statistical analysis, query and reporting. The data that is accessed and processed may be historical or fairly recent and lightly or heavily summarized. The result is shown with reports and charts. The scope of this technique is the 2- or 3-dimensional analysis of historical data for understanding the past. **Analytical processing** also supports the verification mode, but its goal is to make the data available to the business user in a business user's perspective. With **slice and dice**, **drill-down** and **roll-ups** we can answer difficult questions like e.g. how many Suzuki cars did we sell in Japan in the first quarter of 1998 that had a audio system with a price less than 100.000 Yen. These methods will be explained in the next section.
The core components of Data Mining technology have been under development for decades in research such as databases, machine learning , statistics and artificial intelligence.
Databases are critical to everyday commerce, and computers process and record massive amount of transactions. The fundamental concept of databases is the query model (e.g. ask a question to the database and the database is replying), whereas Data Mining uses another query form. A typical query is: What will happen with my product in the future? Unfortunately, the database systems of today offer little functionality to support such "mining" applications. At the same, machine learning techniques perform poorly when applied to such large datasets. This is a main reason that this huge amount of data is still unexplored.
With statistical analysis we can detect unusual patterns of data. These patterns are explained with statistical and mathematical models. Typical techniques are linear and nonlinear analysis, regression analysis, univariate, multivariate and time series analysis. With statistical tools we can successful reduce the analysis time and free scarce resources for other analysis activities. To use these tools a business user must select and extract the right data from the Data Warehouse or datamart. The key features for extraction are knowledge discovery algorithms for pattern and relationship recognition, which are derived from the artificial intelligence sector.

2.1 Data Mining Technology

The most important goals of Data Mining are prediction and description [UMF97]. Prediction uses some attributes from the database and tries to predict unknown or future values of other interesting attributes. Description focuses on human interpretable patterns, which describe the data. There are several Data Mining algorithms which are used to solve specific problems. They are categorized as associations, classifications and clustering algorithms.

2.1.1 Association

With association rules he can better understand the behaviour of customers. E.g.

`70 percent of customers who order butter and margarine also order bread.`

The number 70 refers to the confidence factor, a measure of the predictive power of the rule. The left hand side (LHS) item is butter whereas bread is the right hand side (RHS) item of the rule. The algorithm produces a large number of rules an the user can select a subset of rules that have a higher confidence level. The results can be analyzed with data visualization tools like in the next figure with SGIs Rule Visualizer.

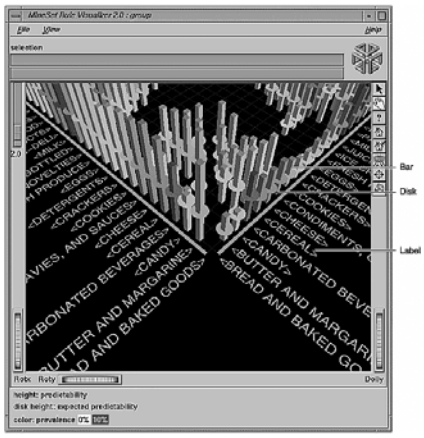

Figure 2: SGIs association rule visualizer

2.1.2 Dependency Analysis

This method presents an important class of discoverable knowledge. A dependency exists between two items if the value of one item can be used to predict the value of another. e.g. Smoking $\rightarrow$ Cancer. An item in a dependency can be an attribute or a relation between attributes. The exact dependency function may or may not be known, because it is probabilistic. e.g. A $\rightarrow$ B with 90 percent. With dependency analysis algorithms we can extract dependencies between items in the Data Warehouse by specifying the confidence factor, which is used to predict the value of one data object from the value of another.

2.1.3 Classification

With classification algorithms we can group data into meaningful classes like e.g. small, medium and large. With this approach we can assign records with a huge number of attributes into a relatively small set called segments. This assignment is performed automatically with clustering algorithms. With clustering algorithms we can discover these classes automatically. The goal of these algorithms is to generate segments of the input data according to a criteria. Note that different clustering algorithms generate different segments or classes of the same data. Some relevant clustering algorithms are pattern recognition, linear clustering, k-means clustering etc. Clustering is often one of the first step in Data Mining analysis.
While this section introduced Data Mining concepts, we will consider **M**ulti**d**imensional **d**ata**b**ases (MDDB) in the next section.

3 Multidimensional databases

Multidimensional analysis involves representing data as the user defines dimensions. Dimensions are almost related in hierarchies and a multidimensional database can have multiple hierarchies. Multidimensional database servers are designed to optimize the storage of dimensional data and provide flexible query and computational operations on dimensions with performance that dramatically exceeds what is possible with SQL and today's relational database implementations. Analysis requirements span a spectrum from statistics to simulation. The two forms of analysis most relevant to mainstream business users are commonly known as 'slice and dice' and 'drill-down' or 'roll-up'.

3.1 Slice and Dice

OLAP enables end users to slice and dice consolidated information in order to view data from many different perspectives. With this technique we can look at multidimensional correlations to uncover business behaviour and business rules. Users can 'cut' and 'rotate' particular pieces from the hypercube (the aggregated data) along any dimensions. The ease and speed with which a rotation can be performed is another example of the inherent advantages of manipulating data in a multidimensional array. Each rotation yields a different slice or a two dimensional table of data. Therefore the rotation is very often called **'data slice'**. The number of possible views increases exponentially with the number of dimensions. In a three dimensional cube with three dimensions like `model`, `color` and `dealership` has six different views.

3.2 Drill-Down and Roll-Up

OLAP allows users to **'drill'** or navigate through information to get more detail. Using the drill-down feature, users can view finer levels of detail within a dataset. In addition to drill-down, **'roll-up'** allows to look at a closer view of dataset.

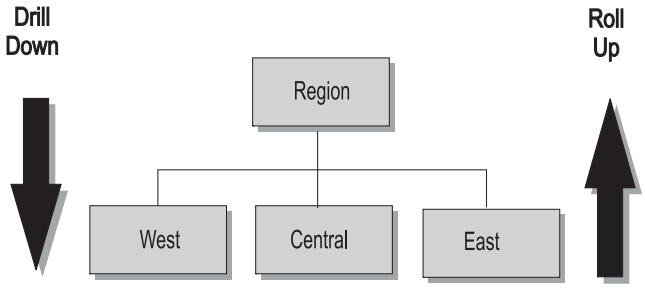

Figure 3: The Drill-Down- and Roll-Up-function

3.3 MSQL - Multidimensional Query Language

MSQL is a multidimensional query language, which provides a flexible access to Multidimensional Database Systems. It is based on the relational SQL and resembles CQL (Cube Query Language), which has its origin at the University of Erlangen [BL97]. In the most case of 'Multidimensional OLAP'-tools a query is reformulated from a graphical user interface. At the moment there is a missing of a standardized multidimensional query language for formulating textual complex decision support queries. Therefore the program WebMDDB has integrated this feature.
One of the primary goals of MSQL was to define a short and easy multidimensional grammar, so that the user can immediately create statements without any problems.
To get a valid result, the analyst has to use at least two dimensions. With the third dimension and the additional `where`-clause the drill-down and roll-up can be realized. With the operators `'=='` and `'!='` the user has the possibility to limitate dimensions and with the `and`-combination more complex queries can be created; of course, one of the biggest problem is the right choose of a view in a 50-dimensional data cube: selection of the dimensions.

- If you will select from the data-cube `Fruit` all products of all regions and from the hole period, the MSQL-statement has the following structure:

  ```
  select sal from fruit, product p, region r, time t
  ```

```
 where p.product == 'PRODUCT',
       r.region == 'REGION',
       t.time == 'TIME';
```

- In the next example the sale of products should be limitated to all eastern regions of 1997. Therefore you have to use a `Where`-clause as well for the dimension `region` (ranging-operation) as for the dimension `time` (drill-down-operation):

```
select sal from fruit, product p, region r, time t
 where p.product == 'PRODUCT',
       r.region == 'EAST',
       t.time == 'T1997';
```

- In the third example the sale should be limitated to all citrus-fruits and to the apples for all eastern regions produced in the 1997s. The `Where`-clause has to be modified including an `AND`-statement: Therefore the where clause was modified to the following statement:

```
select sal from fruit, product p, region r, time t
 where p.product == 'CITRUS' AND p.other == 'APPLE'
       r.region == 'EAST',
       t.time == 'T1997';
```

4 OLAP, Data Mining and the Web

In the article 'Data Warehouse And The Web' Neil Raden says, that in the past 12 months the two most pervasive themes in computer science have been the Internet and Data Warehousing [Rad96]. A marriage of these two giant technologies would be a logical conclusion and many database vendors are gearing up to offer online analytical processing through the World Wide Web. The Data Warehouse, with its underlying support for ROLAP and MOLAP technology enables users to slice and dice through data to crystallize information meaningful for making decisions. The Web is the technology that allows widely distributed users affordable, reliable and uncomplicated access to information all over the globe. Together, they create the underlying technological framework to extend astonishing analytical power to a world-wide electronic audience of millions. For several reasons, the logical platform for broader deployment of OLAP

and Data Mining is the Internet/Intranet WWW in combination with CORBA (Common Object Request broker architecture). Our approach was to implement a subset of functions for a distributed application using the JAVA language, a JAVA enabled Web browser, and a CORBA product. With CORBA we have a complete distributed object platform and with these distributed objects will be the next wave in Internet innovation. we can innovate the Internet because the objects encapsulate the inner workings and presents a well-defined interface. This means that an object's implementation does not affect other objects or applications because the object's interface stays the same.

4.1 CORBA Architecture

With an ORB (object request broker)-enabled browser a user can access objects from multiple servers and hosts [pap97] without regard to client and server operating systems and programming languages. The client object requests the services of other objects with the client ORB. Using an ORB, an object and its client may reside in the same process, or in different processes, which may execute on different hosts connected by a network. With the IIOP (Internet Inter ORB Protocol) the client ORB then looks for ORBs on other systems (Data Mining server, OLAP server) and these server provide objects that can provide the requested services. Each object has its unique name, according to the CORBA naming service, which identifies it and its services to other objects. Once the ORB has found the requested object/service, the two objects can communicate by using IIOP. Suppose a business user wants to analyze his Data Warehouse. From his ORB-enabled Web browser he enters a URL of a valid Web server, which contains a HTML page with an embedded JAVA applet. The Web server send the Web page via HTTP (Hyper Text Transfer Protocol) to the client. On the client side the user enters and selects interesting analysis tasks and with the client ORB, the IIOP messages are sent across the network via IIOP. The server ORB selects the corresponding server object and these objects can perform multiple tasks. One task could be to obtain the calculated rules from the rule induction server object or the retrieval of huge data from the database for further analysis. The server object routes the information with IIOP back to the client applet, which displays the obtained data.

Our basic premise was that it should be possible to design distributed applications in which all of the user-side client software is implemented as JAVA applets, which use CORBA for remote operations with the rest of the application's software components. The user can transparently download the applets when they are needed, thus this removes the need for manually distributing and installing any application specific software.

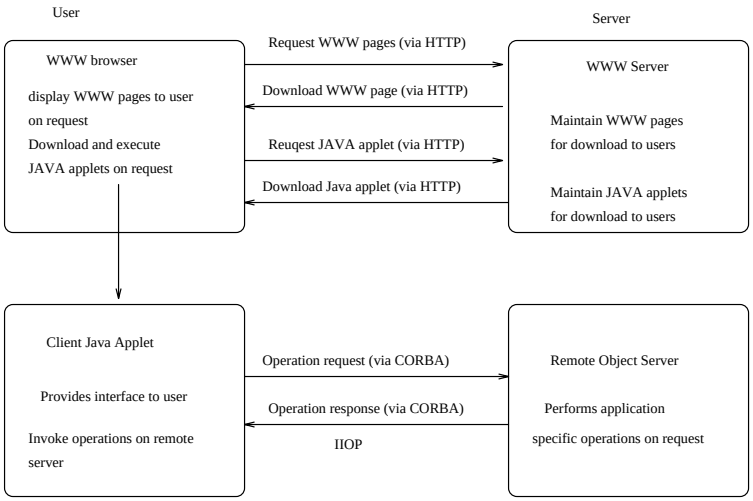

Figure 4: CORBA Architecture

- With the foundation of the JAVA programming language it was possible to make the Web more attractive and also platform independent on the client side.
- The most attractive elements of the Web is its complete platform independence.
- Everything is installed and maintained on the server.
- The minimal extravagance of installations: once you have installed a Web browser, additional applications do not to have to be installed.

  Of course there are multiple steps that we have to consider. E.g. should we download JAR (JAVA archive) files or store them locally on the client side. Of course the JAR concept has the advantage to pack many class files in one JAR file and we need only download these files, while the old approach was to download all class files separately.
- Training expenses are extremely low, because the skills and techniques used to navigate and select information are the same for all Web-based applications.

The Web interface is really powerful for the business manager who need structured navigational OLAP data. The look and feel of the client presentation can be enhanced with JAVA applets to build more powerful applications. With these applets it was possible to create better user interfaces and to be platform independent. The next section will give a better overview how to combine present tools that use these techniques.
At the moment the most common techniques are based on the CGI: it allows a thin client to request the services of a decision support system via the Web server. After receiving a request

for information via CGI, output from the decision support system is simply returned to the browser in HTML format. The disadvantage of this non JAVA solution is the data visualization, which is one of the most important key element of analytical processing. The standard method for displaying graphics with Web browsers would be to broadcast images from the server. In order to draw charts dynamically at the client need to be build into the products, which would be developed in JAVA. With JAVA it is also possible to transfer data with similar methods of transferring data from decision support systems to thin clients (with Microsoft's ODBC protocol and the JDBC protocol defined by Javasoft). Support of these two prominent delivery mechanism is a key factor in the acceptance of a decision support system in a Web-enabled environment, where JDBC and ODBC can be utilized to reach data with a direct access and more efficient.

5 Tools

5.1 Data Mining Tools

There are many tools used for Data Mining: neural networks, decision trees, rule induction and data visualization. These tools scan through detailed transactional data to detect hidden patterns. In this section we will present two important tools that can be used for Data Mining: neural networks and a GUI with parallel coordinates.

5.1.1 Neural networks

The neural network uses rules which it learns from patterns in the data to construct a hidden layer of logic, which is between the visible input and output layers. In general, it is difficult to determine the number of nodes in the hidden layer, but network optimization provides a criteria for establishing the number of nodes in the hidden layer. With this function we can display Akaike's Information Criterion (AIC) for the number of nodes in a specified hidden layer while learning takes place. AIC is commonly used in the field of statistics to determine the optimum number of hidden nodes in a hidden layer. Therefore we can reduce the prediction error during learning if we use a sufficient large number of nodes in the hidden layer. However, once a specific value is reached, the prediction error increases again.
The neural model has to train the network on a training dataset and then uses this model to make predictions. The previous figure shows the concept of neural networks. The problem with neural networks is that they can't be trained on very large databases, but they can be used in combination with special sampling methods.

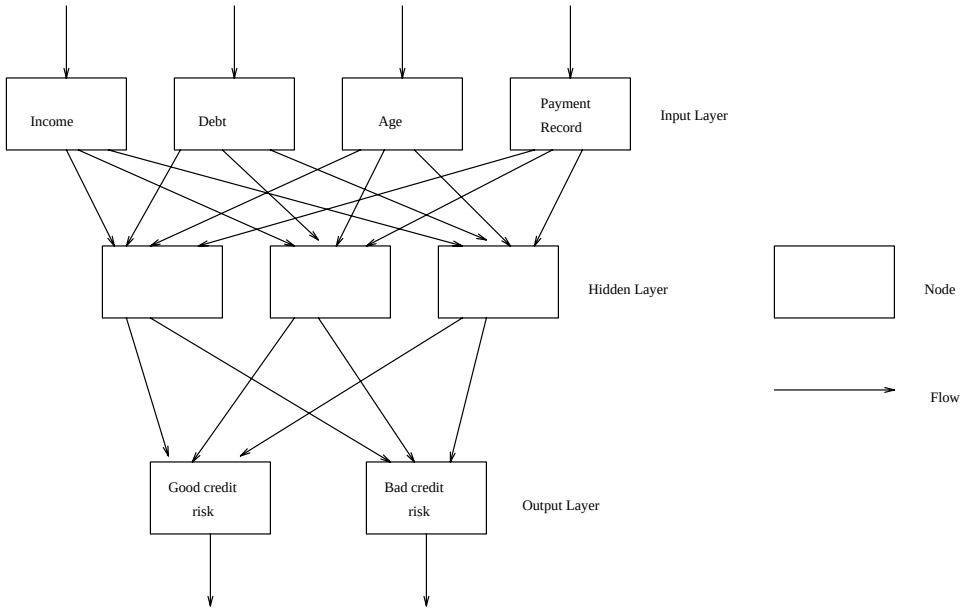

Figure 5: Neural network

5.1.2 Visual Data Mining

Human pattern recognition skills are in many situations better than automated mining algorithms because it is easier for the analyst to gain a deeper, intuitive understanding of the data by representing a graphical image, which presents a large amount of information in a concise manner. With the construction of visual interfaces we can detect many patterns. One famous visual Data Mining tool is the visualization of large datasets with parallel coordinates. With parallel coordinates data analysis we can visualize datasets with MANY (in principle unlimited) variables. We can use a set of visual queries to rapidly discover relations (if they exist) among the variables and how these relations effect various objectives. In parallel coordinate displays we have n variables (attributes) with n axes, which are plotted side by side. One record consists of n-1 lines which connect the n attributes of a record across the n axes. With this kind of visualization you can drag and drop axes, change axes, view the minimum and maximum of every attribute, see the categorization level and perform useful zoom ins. If we have a huge amount of records we will encounter that the display swamps beyond recognition. But even if it is completely overcrowded, we can use highlighting operations. This intelligibles subsets of records in an efficient way and can detect correlations across many variables. With this visualization the business analyst can display the relationships between many variables simultaneously by mapping the coordinates as parallel axes rather than the perpendicular axes

of traditional visualization tools. The representation of multivariate datasets preserves all the information and transforms multivariate relations into well defined 2D patterns. This helps us to manage large datasets.

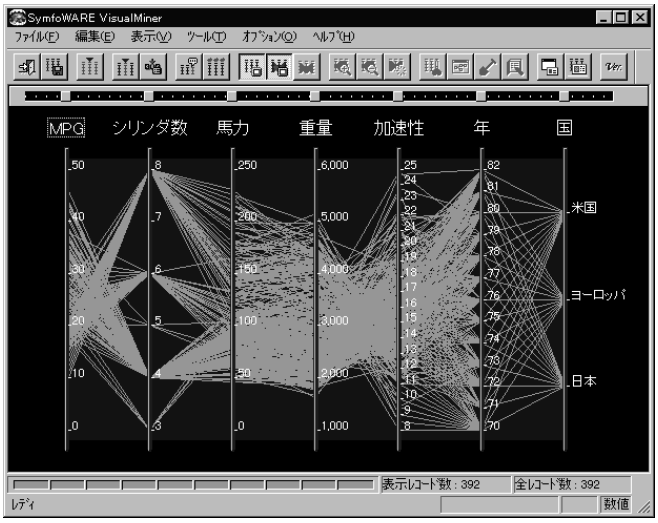

Figure 6: Parallel coordinates

5.2 MDDB Tools: WebMDDB - a web based analysis program

WebMDDB (= **W**eb **M**ulti**d**imensional **D**ata**b**ase) is a Web-based OLAP-tool developed at the Johannes Kepler University of Linz [Hal97]. The primary goal was to implement a tool, which allows to the end user to have an interactive access to huge amounts of data via a Web browser. A good representation of the data is one of the most striking features of the tool. It is one of the first tool that offers a complete end-to-end JAVA solution for full-featured, browser-based OLAP decision support. By using JAVA it was possible to realize a good interactive visualization, which allows making queries and analysis from a web browser. The following features are characteristic for WebMDDB:

1. 100% end-to-end thin-client JAVA solution.
2. Fully interactive real-time interaction between the user and the data.
3. Works with any standard Web server and browser.
4. Easy to use the (realization of drill-down, roll-up, ranging and rotation).

5. Interactive presentations of the data in 2-D and 3-D.

6. Integration of a multidimensional query language (MSQL), which is easy to use.

5.3 WebMDDB Benefits

WebMDDB is a full-featured OLAP query builder. Using the simple, point-and-click, OLAP Query Wizard, analysts and users can select their business information. After downloading the applet, he connects to the OLAP-database via JDBC. It is easy to integrate also other OLAP-databases by doing some modifiers. Once he is connected to the database, he can select information with dimensions, measures, levels, and much more.

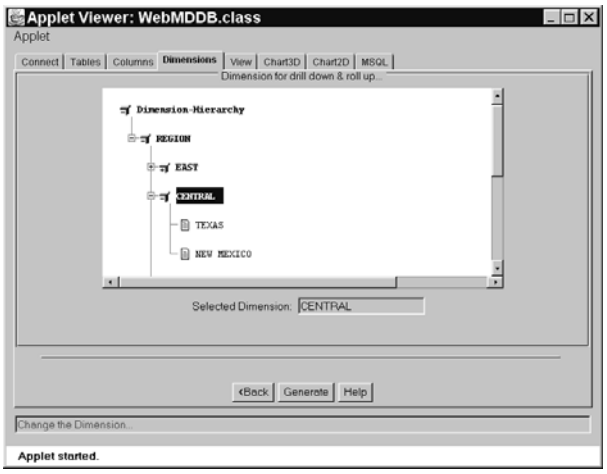

Figure 7: The hierarchical representation for making the drill-down- and roll-up-operations

5.3.1 Analysis

The analysis tool enables real-time, interactive data visualization with 2-D and 3-D JAVA based charts for more intuitive recognition of patterns and trends. Users can dynamically select data, resize and rotate charts. Using the 2-D visualization, the user can realize powerful tabular analysis. This allows him to **'slice and dice'** via pivoting and drill-down in a spreadsheet-like environment. Another feature of WebMDDB is the visualization of the distribution and the correlation between the data: each cell of the spreadsheet corresponds to a cube cell with its value. If there is a high value compared to the others, the cell has a deeper color. The last analyzing feature is the integration of a multidimensional text based query tool (MSQL).

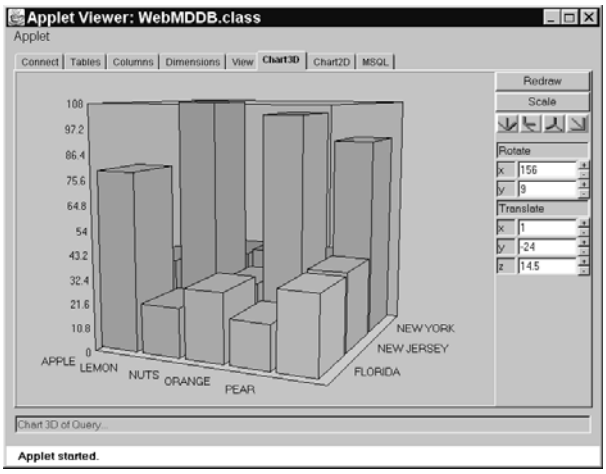

Figure 8: The 3-D visualization of WebMDDB

5.3.2 Architecture

Establishing and end-to-end java environment, WebMDDB maintains persistence and state between the client and the server with its connection to the database. No CGI is used and no unnecessary processes are spawned.
Of course there are many discussion about CGI but it was the first 3-tier client/server solution over the Internet. All together, WebMDDB's JAVA architecture allows you to quickly and easily provide data access and analysis capabilities to many users in a point-and-click environment.

6 Conclusion

We have seen that one of the most important things for Data Warehouses is a well defined visualization of the huge amount of data. Otherwise the end users loose the overview and are not able to make a good decision. Therefore in the future, there should be a direct one's attention to a well defined visualization on the data. One of imaginable possibilities could be realized with Virtual Reality and VRML: the user could navigate in his virtual environment and be involved directly in his data [HJS96, Arb97]. In this paper we have seen that Data Mining and MDDB are related facets of a new generation of intelligent information extraction and management tools. If the enterprise chooses the wrong tools, it can lead to business user frustration, lower productivity and loss of strategic understanding of the business.

References

[Arb97] Arbor. VRML: Visible Decision VRML Visualization. http://www.webgate.arborsoft.com/vdi/vrml.html, Juli 1997.

[B.96] Moxon B. "defining data mining. *DBMS Data Warehouse Supplement, August 1996*, 38:173–198, 1996.

[BL97] A. Bauer and W. Lehner. The Cube-Query-Language (CQL) For Multidimensional Statistical And Scientific Database Systems. In DASFAA'97, editor, *5th International Conference on Database*, Melbourne, Australia, April 1997.

[Hal97] Michael Haller. Multidimensionale Datenbanken und die Anbindung an das Internet mit Java, Dezember 1997.

[HJS96] Michael Haller, Georg Jenichl, and Gernot Schaufler. *Einführung in Virtual Reality*. GUP (Graphische und Parallele Datenverarbeitung), Universität Linz, 1 edition, Juni 1996.

[HS96] Prakash C. Rao Harijinder S.Gill. *The official Guide to Data Warehousing*. VCH, Weinheim-Basel-Cambridge-New York, 1996.

[McC97] David R. McClanahan. Data Modeling for OLAP. *Database Advisor*, (66–70), März 1997.

[pap97] Netscape White paper. Corba: Catching the next wave. *Netscape White Paper*, 31:1213–1228, 1997.

[Rad96] Neil Raden. Warehouses And The Web. *TechWeb*, Mai 1996.

[UMF97] Padhraic Smyth. Usama M. Fayyad, Gregory Piatetsky-Shapiro, editor. *From Data Mining to Knowledge Discovery: An overview*, University of Ulster, Northern Ireland, 1997. MIT Press.