Transactions on Information and Communications Technologies vol 19 © 1998 WIT Press, www.witpress.com, ISSN 1743-3517

Knowledge Discovery with Retail Basket Analysis

Maria de Fátima Rodrigues, Carlos Ramos & Pedro R. Henriques
Polytechnic Institute of Porto, School of Engineering
Rua São Tomé, 4200 Porto, Portugal
EMail: fr@dei.isep.ipp.pt

Abstract

Minute by minute the amount of data in the world databases is increasing inexorably. To support this growth of data the concept of data warehouse (DW) was created. DW when combined with On-Line Analytical Processing (OLAP) Codd[2] and Executive Information Systems (EIS) Buytendijk[1] tools, enable data access and visualization in a very flexible way. Features include very quick data exploration, vertical navigation (drill up/drill down), aggregation and graphical facilities.

However, the amount and the complexity of data in data warehouses is so big that it becomes difficult to the business analysts to recognise trends and relations in data even with multidimensional decision support systems. A new generation of tools and techniques for automated intelligent database analysis is needed. These tools and techniques are the subject of the rapidly emerging field of Knowledge Discovery in Data Bases (KDD). In this paper, we propose an integrated system - DECADIS - DEscoberta de Conhecimento em Armazéns de Dados de DIStribuição (Knowledge Discovery in Retail Data Warehouses), designed for understanding customer behaviour and consumption patterns in a Portuguese company in the retail industry.

1 Introduction

The introduction of bar codes and scanning of those for almost all commercial products and the computerisation of business transactions (e.g. credit card purchases) have generated an explosive growth in retail data warehouses. Simultaneously, the big competition in this business area, has created a significant need for a rigorous knowledge about sales versus clients. The treatment of such volumes of data and the need to understand customer behaviour can only be obtained with a new generation of tools to automatically discover information hidden in the collected data. The goal of identifying and utilising information hidden in data has three requirements:

- the captured data must be integrated from many applications
- the information contained in the integrated data must be extracted or mined
- the mined information must be organized in ways that enable decision-making.

These requirements imply that a data mining system must interact with a data warehouse which organises data from multiple sources in ways that facilitate analysis, and must interface with a multidimensional decision support system (DSS) based on OLAP/EIS technology, with two purposes: help to understand the relations among data and explore new questions as a result of data mining exercise. This is the structure of DECADIS system.

2 System Architecture

In this section the architecture of DECADIS system will be described. This project has started with the construction of a client data mart, because DECADIS is a specific system designed for few marketing analysts and because the data warehouse of this company has many data subjects with a great volume of data (more or less 270 GB) that are not relevant for this application.

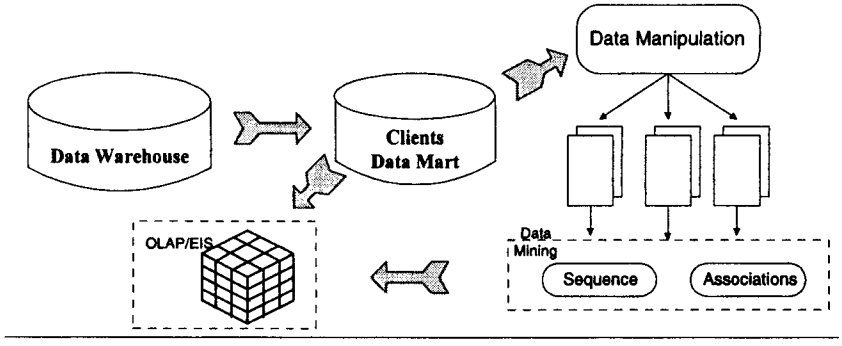

Figure 1: System Achitecture

After building a client data mart it was inevitable the development of an EIS application to have a first sensitivity and understanding of client data. This permits to define some client classes in terms of volume of purchases, variety of products, shopping frequency, client profession, etc, and have an economic evaluation of the project.

Next to this first phase, it was time to mine the client data mart. Our first goal was to do market basket analysis to give insight into who they are the costumers what they purchase and why they make certain purchases. This type of information is actionable because it can suggest new stores layouts; it can determine which products to put on special; it can indicate when to issue coupons, and so on.

3 Knowledge Discovery in Databases - An Overview

Knowledge Discovery in Data Bases (KDD) is the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data Fayyad[6].

This process uses data mining (DM) methods (algorithms) to extract (identify) knowledge according to specifications of measures and thresholds. This knowledge can provide new insights into relationships between data elements and facilitate more productive and sophisticated decision support applications.

Transactions on Information and Communications Technologies vol 19 © 1998 WIT Press, www.witpress.com, ISSN 1743-3517

Data mining (DM) involves fitting models to, or extracting patterns from data without the additional steps of the KDD process (such as incorporating appropriate prior knowledge and interpretation of the results). Different methodologies can be used, depending if data mining goals are discovery hidden associations (e.g., the fact that people who purchase milk also tend to purchase cereals is a pattern that relates product consumption habits), predictable sequences (e.g., 5 % rate reduction prices results in a 20 % increase sales), or accurate classifications (e.g. cars built on production line X using parts from suppliers A,B,C have significant more defects than average).

This section contains a brief description of the data mining techniques used in our system when doing retail basket analysis.

3.1.1 Association Rules

The problem of mining association rules over basket data, was introduced in Agrawal[4]. In its original form this task was defined for a special kind of data, often called basket data, where a tuple consists of a set of binary attributes called items. Each tuple corresponds to a customer transaction, where a given item has value true or false depending on whether or not the corresponding customer bought the item in that transaction. This kind of data is usually collected through bar-code technology, such as a supermarket scanner. Actually market basket analysis as other applications in fields like medicine, banking, insurance...

The application of the association rules algorithm relates to finding intra-transaction patterns and can be defined as follows: given a database of transactions, wherein each transaction represents a set of items (e.g. categories of items purchased), generate all associations such that the presence of some specific item(s) x in a transaction implies the presence of other item(s) y. The association rule $x \Rightarrow y$ will hold given a support and a confidence greater than a user-specified minimum support ($s_{\min}$) and a minimum confidence ($c_{\min}$), in which

support is the number (or fraction) of the transactions that contain a given item set; and

confidence measures the frequency that items in a multi-item set are found together.

The discovery of association rules is usually performed in two

steps. First an algorithm determines all the sets of items having s_{min} greater than or equal to s_{min} specified by the user. These sets are called frequent itemsets. Second, for each frequent itemset, all possible candidate rules are generated and tested with respect to c_{min}.

Recently the discovery of associaton rules has been extended to cope with atributes types other than strictly binary. For instance, Srikant[5] propose algorithms that, given a costumer transaction database and a generalization hierarchy (or taxonomy) on the items, discovery associations rules between items at any level of the hierarchy. In this paper we propose discovery association rules between items in different nivels of produt categorie but with traditional algorithms, our goal is that we can do the same, but with previous treatment of ticket data, such as, combined items at different nivels of hierarchy but with the same frequency.

3.1.2 Rule Induction

Rule induction induces a model - e.g. a rule set or a decision-tree based on statistical significance - and uses the induced model to classify new tuples. One strength of induction is that the process automatically include in its rule only the factors which really matter in making a decision; others will be discarded, and usually produces very comprehensible knowledge, unless the complexity of the relationship between classes and the nevel of noise in data are big.

4 Experiments

In this section we provide a description of a set of experiments conducted using different data mining techniques on *ticket data*. The results collect during the experimentation are presented.

All experiments were conducted using Clementine, version 4.0. This is an integrated data mining toolkit, wich uses a visual programming interface and supports the integration of all KDD stages. A data mining project is built by linking objects (referred to as "nodes") in the Clementine Workspace. Each node carries out a certain stage of the data mining process, such as data access or a particular model. Nodes can be combined flexibly to create various analysis paths and cover the whole data mining process. Clementine provides support

for decision tree analysis and rule induction, neuronal networks and Kohonen self-organising-maps, as well as association rule generation and regression analysis. All these models can be combined to maximise their efficiency and explanatory power.

Ticket data describes the contents of supermarket baskets (i.e. collections of items bought together), plus the associated loyalty card number of the purchaser. The goal is to discover groups of costumers who buy similar products and next characterised them only by its loyalty card number.

This data mining exercise contains two phases:

- firstly, links between items purchased are discovered using association rule modelling , and
- secondly, the purchasers of identified product groups are profiled using C5.0 rule induction.

This exercise does not make direct use of predictive modelling, so there is no accuracy measurement for the resulting models and no associated training/test distinction in the data mining process.

4.1 Choosing the Right Set of Items

Probably the most difficult problem when applying association rules is determining the right set of items to use in the analysis. The computations required to generate association rules grow exponentially with the number of items and the complexity of the rules being considered. The solution is to reduce the number of items by generalizing them. However, more general items are, usually they are less actionable. One compromise is to use more general items initially, then to repeat the rule generation to hone in on more specific items. As the analysis focuses on more specific items, we use only the subset of transactions containing those items.

Another problem with association rules is that it works best when all items have approximately the same frequency in the data, because items that rarely occur, are in very few transactions, may be pruned. The use of item taxonomies can ensure that rare items are rolled up, so they become more frequent and included in the analysis in some form. More common items may not have to be rolled up at all, prevent rules from being dominated by the most common items.

Transactions on Information and Communications Technologies vol 19 © 1998 WIT Press, www.witpress.com, ISSN 1743-3517

The purpose of complementary items is to enable the analysis to take advantage of information that goes beyond the taxonomy. Complementary items do not appear in the product taxonomy of the original items, because they cross product boundaries. This complementary items may include information about the transactions themselves, such as whether the purchase was made with cash, a credit card or check, the day of the week or the season. By including a complementary item for the month or the season when a transaction occurred, it's possible to detect differences between seasons and seasonal trends. However, it is not a good idea to put in data too many complementary items, because there is a danger, complementary items are a prime cause of redundant rules. *We must include only virtual items that could turn into actionable information if found in well-supported, high-confidence association rules* Linoff[3].

4.3 Data Transformations

We have a given database of costumer transactions. Each transaction consists of the following fields: loyalty card identifier, transaction-date, and the items versus the quantity purchased in the transaction. No costumer has more than one transaction with the same transaction time.

When applying market basket analysis, it is crucial to use a balanced taxonomy of the items being considered for analysis, that's why we must replace the items in the analysis with generalized items from different levels in the taxonomy. By judiciously choosing the right level of the taxonomy, these generalized items should occur about the same number of times in the data, improving the results of the analysis. This involves examine the data to determine the frequency of each item and with this information we can substitute the items purchased by it's right categorie. The result data set consists of a database of costumer transactions each of one with a set of items in different nivel categorie. We do not consider quantities of items bought, because they have no meaning in this data reduction performed, each item categorie is a binary variable representing wether an item was bought or not.

The dataset to be mine have N fields in each record, each representing a basket.

Basket summary:

Cardid	Loyalty card identifier for costumer purchasing this basket
Value	Total purchase price of basket
Pmethod	Method of payment for basket
Categ1	Flag for presence of product categorie 1
Categ2	Flag for presence of product categorie 2
....	
CategN	Flag for presence of product categorie N

4.4 Discovery Affinities in basket Contents

The first stage it was to acquire a broad brush picture of affinities (associations) in the basket contents using Generalised Rule Induction (GRI) to produce association rules. The result of this operation is an unrefined model containing the following association rules:

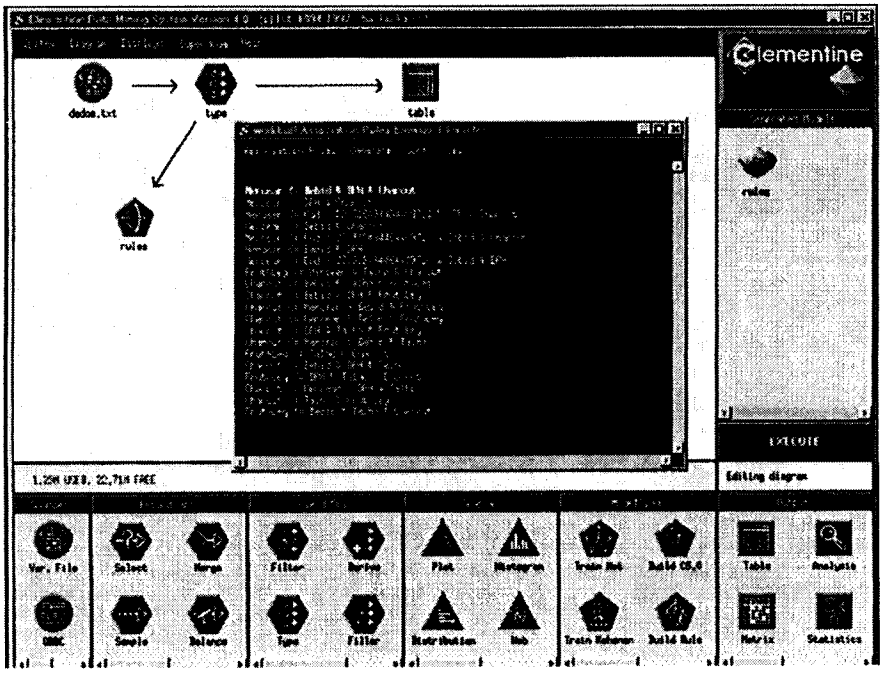

Figure 2: Association Rules Browse

and sausage; fruit&vegetables, grocery and sausage are also associated. It is possible highlight graphically this associations

using the *web display.* The Web Node is used to show the strength of connection between values of two or more symbolic fields. Connections are show graphically, with dotted, normal and heavy lines used to show connections of increasing strength.

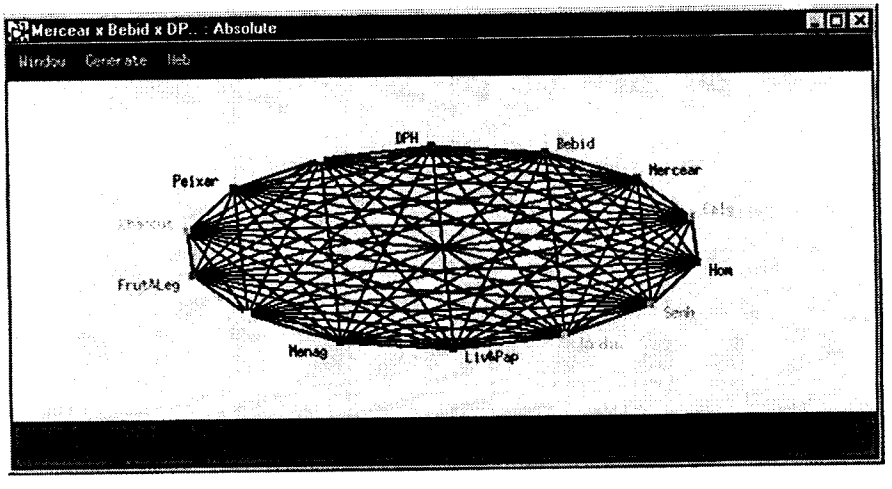

Figure 3: Web Node with all links

Because most of items categorie occur in several baskets, the strong links on this web are too numerous to show the groups of costumers suggested by the GRI model. We need to raise the thresholds used by the web so as to show only the strongest of these links.

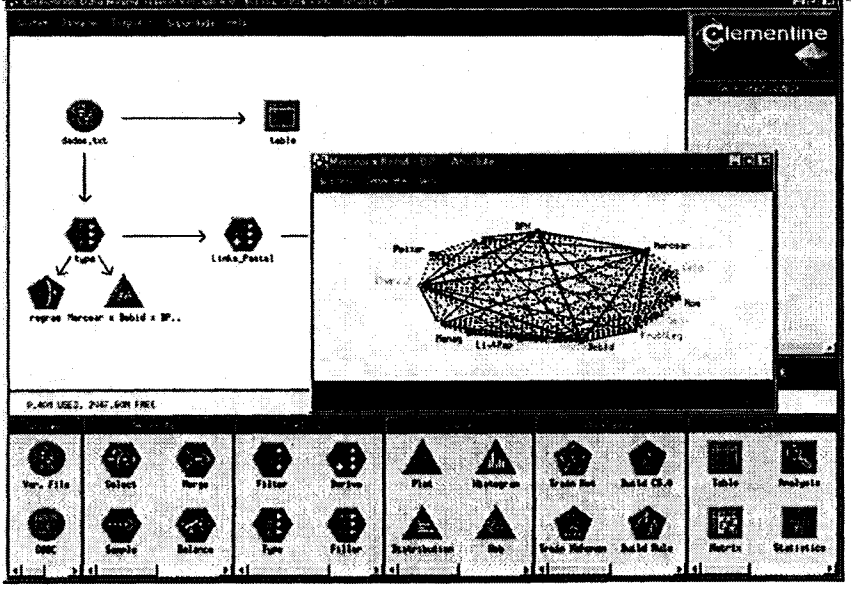

Figure 4: Web Node only with the strongest links

Here are the groups of costumers stand out:

- those who buy fruit & vegetables, grocery and sausage, wich we might call "healthy eaters"
- those who buy drinks, sausage and grocery

4.5 Profiling the Costumers Groupings

We have identified several groupings of costumers based on the types of products they buy, but we would also like to know who these costumers are, that is, their card number identifier. This can be achieved by "tagging" each costumer group with a flag, then using rule induction (C5.0) to build rule-based profiles of these flags. The result model contains rules that attach to each of one of the groups of products previous identified, various loyalty card number. This rules will support marketing experts in the business processes: defining

hypothesis adequate to each customer; designing promotions targeted to customers; and building predictive models of customer behaviour.

The next figure shows the stream constructed, which look like this:

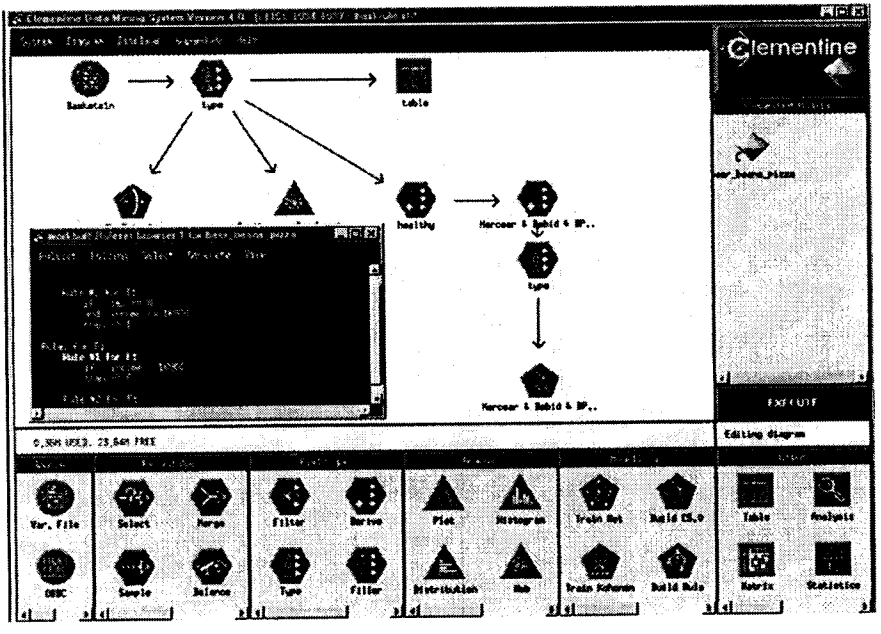

Figure 5: Stream constructed

5 Conclusions

We have proposed a system for discovering patterns of costumer behavior among purchases made by them in a retail company; the system is based on categorizing the items purchased according to it's buying frequency and the various nivels of categories existent in product codification of the company, and on applying a set of data pre-processing transformations and a set of data mining techniques to mine *ticket data.*

We have illustrated an application of the system describing some results obtained. Among others, behavioral patterns that include

affinities among types of products have been identified, and the combination of items at different nivels of hierarchy codification has been successfully.

From this experience we can envision that the proposed system can also be applied in other industries, such as, analyse medical patient histories to give indications of complications based on certain combinations of treatments, detect unusual combinations of insurance claims to detect frauds.

References

[1] Frank A. Buytendijk. Olap: Playing for keeps. Technical report, 1995.

[2] C.T. Salley E.F. Codd Associates E.F. Codd, S.B. Codd. Providing olap (on-line analitical processing) to user-analysts: An it mandate. Technical report, 1993.

[3] Gordon Linoff Michael J. A. Berry. *Data Mining Techniques for Marketing, Sales, and Customer Support.* John Wiley Sons, Inc, 1997.

[4] T. Imielinski R. Agrawal and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, May 1993.

[5] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings 1996 ACM SIGMOD Int. Conf. Management of Data*, pages 1–12, 1996.

[6] P. J. Smith U. Fayyad, G. Piatetsky-Shapiro and R. Uthurasamy. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI/MIT Press, 1996.