47 USING DATA MINING FOR VIRTUAL ENTERPRISE MANAGEMENT

*L. Loss *, R. J. Rabelo #, D. Luz *, A. Pereira-Klen *, E. R. Klen **
Federal University of Santa Catarina – Florianópolis (SC), BRAZIL
** {loss, luz, klen, erklen}@gsigma.ufsc.br*
# rabelo@das.ufsc.br

This paper presents exploratory results on how a data-mining-based tool can be used to enhance the quality of decision-making in a Virtual Enterprise environment. The developed tool is based on the Clustering mining method and implements the K-Means algorithm. The algorithm is explained, its utilization in the proposed model is introduced and the implementation results are presented and stressed in the end of the paper.

1. INTRODUCTION

Data Mining (DM) has emerged as a very powerful technique to find out patterns and relationships in large information repositories. The application of DM on several domains (e.g. marketing, investment, fraud detection, manufacturing, financial services) has increased significantly in the last years. However, its application on more volatile scenarios, like the ones represented by Virtual Enterprises (VE) is still very incipient. A VE is here considered as a dynamic, temporary and logical aggregation of autonomous enterprises that interact with each other as a strategic answer to attend a given opportunity or to cope with a specific need, and whose operation is achieved by the coordinated sharing of skills, resources and information, enabled by computer networks (Rabelo et al., 04).

Recently, many investments have been made by enterprises to support inter-enterprises communication in order to improve the information exchange among suppliers and clients as well as to enable distributed information access facilitating and enhancing the Virtual Enterprise management. The downside of this success has been information overload: how should this amount of information be used in a value added way? The fact is that there is a mass of valuable information "hidden" in the enterprises' databases which are relevant for business (Chandra et al., 2000). Examples of this include patterns of clients' behaviors, seasonal or repetitive events, suppliers' performance per product, and many others. These qualitative and quantitative unknown information correlations can be used to improve both the quality of decision-making and the formulation of successful strategies among the VE partners.

Business Intelligence (BI), Competitive Intelligence and *Market Intelligence* are examples of techniques that have been used to better organize and to properly filter the information for decision-makers (Begg et al., 2002). In spite of their potentialities, some handicaps still have to be overcome such as their application on dynamic VEs, where new suppliers and clients can enter or quit along the operation process. Supporting this requirement is extremely important as the success of VE critically depends on recognizing partners′ expertise, tools and skills as marketable knowledge assets (Lavrac et al., 2002). Additionally, those techniques are not designed to be "active" tools, i.e. systems that go through information repositories in order to try to discover new information elements that can augment decision processes.

Based on that, this paper presents a hybrid approach which joins the fundamentals of DM and BI regarding the VE environment requirements. A preliminary validation of this approach was done by means of the development of an exploratory data mining tool that works together with the VE Cockpit, a BI-based VE management system (Rabelo et al., 2002).

This paper is organized as follows: Chapter 2 frames the global scenario in which the developed tool is inserted in. Chapter 3 describes the basic concepts of the data mining approach as well as explains the K-Means algorithm. Chapter 4 depicts the implemented prototype and results, and Chapter 5 provides the main conclusions.

2. GENERAL SCENARIO

In this work a VE is considered as a network of several enterprises where one of them – called VE Coordinator – has the role of managing the VE-related processes as well as of acting as the front-end with the end customer. The model presented in this section has been developed with the aim of extracting helpful information for the VE manager so that better decision-making can be taken during the VE Operation phase. The VE Manager interacts with the VE Cockpit system and is supported by its functionalities to operate the VE.

The information is stored in the VE Coordinator's database, which contains current and historical data about its suppliers, clients, and involved orders (production orders, shipment orders, sales orders and so forth). Figure 1 illustrates this model which is composed by:

- VE Cockpit system: having two main modules (Creation & Configuration, and Operation) which in turn feed the VE database during the course of the VE existence.
- Data mining tool (DM-Tool): its first module processes the database using a specific data mining algorithm (see next chapter) and sends its results to the DM Analyzer module. The second module processes these results and provides the VE manager with high-level conclusions, i.e. the envisaged information patterns.

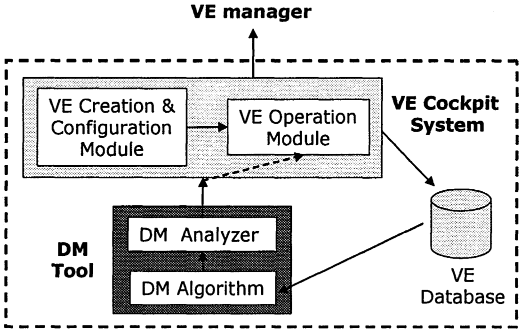


Figure 1 - General scenario of the data mining application

3. DATA MINING AND THE K-MEANS ALGORITHM

DM is the process related to the extraction of knowledge from data repositories with the aim of identify valid, new, potentially useful and understandable patterns (Fayyad et al., 1996a). DM is one of the main steps of Knowledge Discovery in Databases (KDD), generally defined as the process of automatically extracting useful knowledge from large collections of data (Adriaans et al., 1996). This is carried out by means of finding patterns in data, driven by some "rules of interest" that are defined by the user (Fayyad et al., 1996a) (Lavrac et al., 2002).

The KDD process attempts to develop technologies for automatic knowledge extraction by means of mapping low-level data (a large amount of "raw" data) into other forms that might be more compact (a short report), more abstract (a descriptive model of the process that generated the data), or more useful (a predictive model for estimating the value of future cases) (Fayyad et al., 1996a).

Based on the DM theory, two approaches can be used in the data processing. The first one is based on hypothesis tests, verifying or rejecting a hypothesis or previous ideas that are still undercover. The second one is based on the knowledge synthesis that is related to information discovering without any initial condition or supervision. Regarding the main characteristics of the VE domain and the requirements of a VE manager the second approach was chosen, as the main interest in this work is to find patterns which are not previously known.

Among a number of existing non-supervised method, *Clustering* was selected to be used considering its potentiality, simplicity and, at the same time, the facility to reach results quickly. Clustering is a common descriptive task where it is tried to identify a finite set of categories to describe data (Fayyad et al., 1996b). Examples of applications of clustering include discovering homogeneous subpopulations of potential consumers and identification of subcategories of suppliers according to some performance metrics. The clustering method is performed through an analysis of the relationships among the database's fields and tables. The similarities among attributes are in intrinsic property and it is not necessary to train pre-defined classes. Usually, it only requires an end-user to set up initial parameters and to refine them afterwards in the case a non-satisfactory result (i.e. a given configuration of data sets/patterns) is achieved. The existing clustering algorithms are based on several methods, such as (Berkhin, 2003): hierarchical methods, partitioning methods, grid-based methods, methods based on co-occurrence of categorical data, and constraint-based clustering.

The *K-Means* algorithm (MacQueen, 1967) is a widespread partitioning method that has been used in many works. In spite of some limitations, *K-Means* was the one selected to be used in this exploratory work since: it can be applied on several application domains; its implementation is relatively simple; and it works with information free of context, facilitating the search of data associations.

The K-Means Algorithm

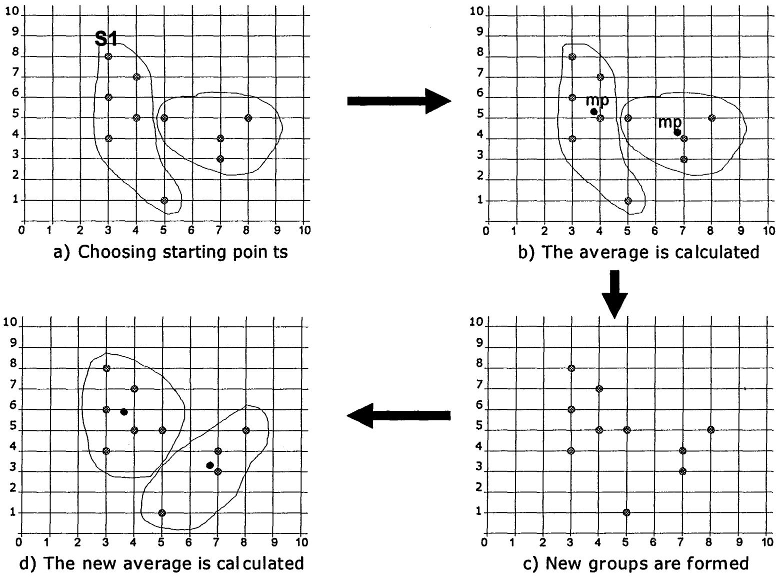
This section will briefly illustrate the functioning of the *K-means* algorithm. As an example, consider the simple database table illustrated in Table 1. It contains records related to ten suppliers about their production capacity level and their ranking (best-delivery ranking) from the VE Coordinator point of view.

Supplier	Capacity	Ranking
S1	3	8
S2	3	6
S3	3	4
S4	4	7
S5	4	5
S6	5	5
S7	5	1
S8	7	4
S9	7	3
S10	8	5

Table 1 – Database table

Firstly, the user should indicate the value of *k*, i.e. how many clusters (grouping criteria) (s)he is interested to find information about. Assuming that Table 1 would be the only one available, up to 10 clusters could be considered. In the example showed in figure 2, two clusters are used, trying to obtain some knowledge from the suppliers′ capacity and ranking. After that, a bidimensional vector / group is created to represent each supplier (Figure 2a), where, for instance, the Supplier 1 is fixed in the points (3,8).

Starting points are chosen for each group by a shuffle algorithm, after which medium points (*mps*) are calculated for each one (Figure 2b). All points are resettled according to the distance from the *mps*. Points will belong to the group that contains the closest distance to the *mp* so they can change from one group to another, i.e. new groups are created (Figure 2c). The *mps* are recalculated according to these new groups, and the process is repeated until that the new groups are equal to the previous ones, or the algorithm reach a (predefined) maximum number of iterations (Figure 2d). When a large number of database registers is involved, different final results can be reached by the algorithm. It means that different initial conditions (for instance, the number of clusters) lead to different results.



a) Choosing starting poin ts

b) The average is calculated

d) The new average is cal culated

c) New groups are formed

Figure 2 – Clustering steps

Summing up, on the one hand *K-Means* introduces some difficulties to decide which the most suitable solution is; on the other hand it provides other visions to the decision-maker that can enrich his/her insights.

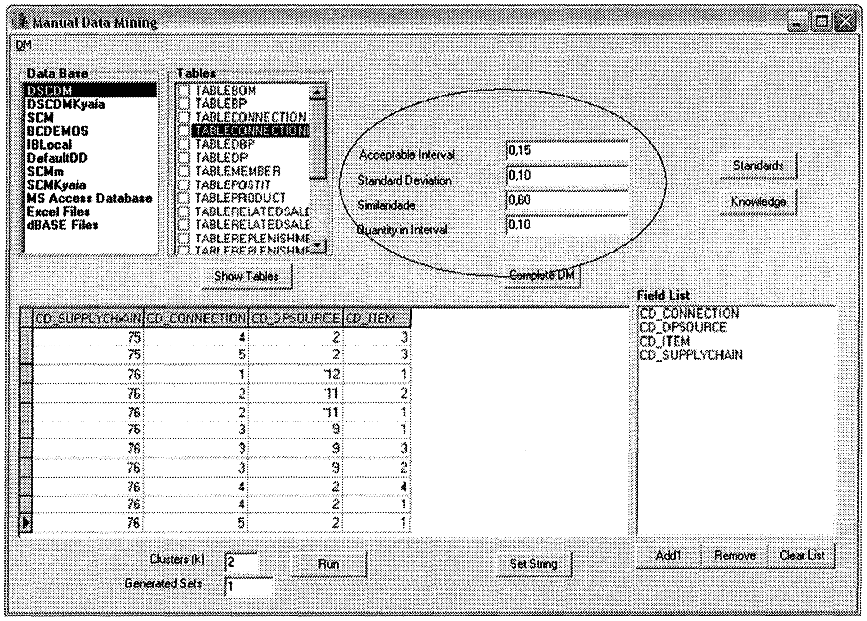
4. PROTOTYPE

The algorithm stressed in the previous chapter has been implemented in C++ in a PC / Windows XP platform and was integrated in the VE Cockpit system (see figure 1).

The first step in that module is to indicate the database to be used as well as to select the database's tables that should be mined by the system (Figure 3 / top left).

The second step is related to the configuration of the mining system. At this point two ways are provided: *manual* and *automatic* configuration. In the *manual* way the user should define the number of clusters as well as the number of data sets (i.e. possible different/final results) to be generated (Figure 3 / bottom left). It requires a certain level of experience from the user. In the *automatic* way, the mining system generates final results automatically, taking into account four pruning parameters that the user should specify. They are: acceptable interval, standard deviation, similarity, and quantity within the interval (Figure 3 / top / inside the circle). The final results are selected according to an internal value reached by the algorithm that is related to the sum of the Euclidean distances of the clusters.

In the third step, the user should define the fields of the selected database tables that will constitute the mining sample. In figure 3, the table *TableConnectionDetails* and the fields *CD_SupplyChain*, *CD_Connection*, *CD_DPSource* and *CD_Item* were chosen. This means that the user "thinks" that useful correlations between the supply chain id, the relations among companies per item type can be revealed. The K-Means algorithm then combines these four fields trying to identify relevant correlations among them. The data set used in this prototype come from a database fed with real information from industrial partners of a research project.



Acceptable Interval 0,15
Standard Deviation 0,10
Similaridade 0,60
Quantity in Interval 0,10

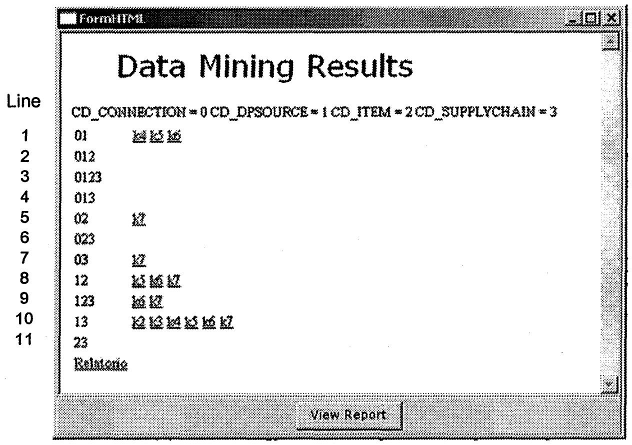
CD_SUPPLYCHAIN	CD_CONNECTION	CD_DPSOURCE	CD_ITEM
75	4	2	3
75	5	2	3
76	1	12	1
76	2	11	2
76	2	11	1
76	3	9	1
76	3	9	3
76	3	9	2
76	4	2	4
76	4	2	1
76	5	2	1

Field List: CD_CONNECTION, CD_DPSOURCE, CD_ITEM, CD_SUPPLYCHAIN

Clusters (k) 2
Generated Sets 1

Figure 3 – Setting standards for automatic search

In the fourth step the results generated by the algorithm is shown (Figure 4), also providing the numeric association with the alphanumeric fields. In this case, *0* means *CD_SupplyChain*, *1* means *CD_Connection*, *2* means *CD_DPSource* and *3* means *CD_Item*. Results can be saved in a database or be expressed as a HTML report.



Data Mining Results

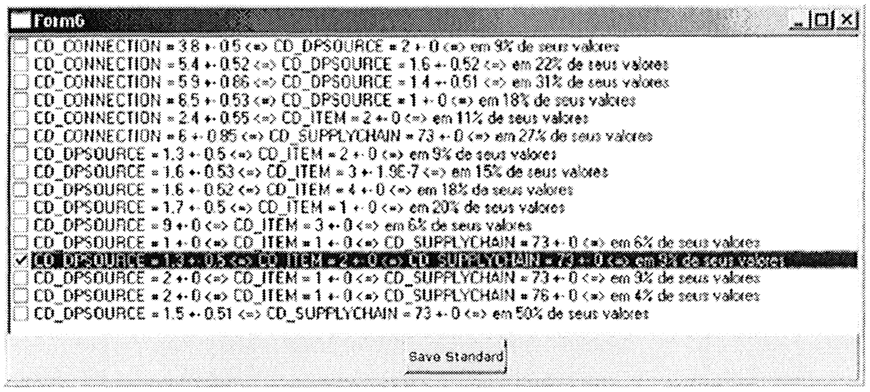
```
Line  CD_CONNECTION = 0 CD_DPSOURCE = 1 CD_ITEM = 2 CD_SUPPLYCHAIN = 3
 1    01      k4 k5 k6
 2    012
 3    0123
 4    013
 5    02      k7
 6    023
 7    03      k7
 8    12      k5 k6 k7
 9    123     k6 k7
10    13      k2 k3 k4 k5 k6 k7
11    23
      Relatorio
```

Figure 4 – Results using the automatic data mining search

From the six clusters found by the algorithm (the “lines” 1, 5, 7, 8, 9 and 10) the user would normally elect the ones which got the largest number of logical *k* occurrences (3, 3, and 6, respectively, in the lines 1, 8, and 10), i.e. the largest amount of correlations among those four fields.

The role of clustering algorithms uses to end here, at this stage and level of information processing. Nevertheless, this result is still expressed in a too low level, creating difficulties for an easier understanding and hence for an agile decision-making. For that purpose, the DM tool was extended with the DM Analyzer module in order to provide a clearer level of results description. Figure 5 shows the final results generated by the DM tool to the VE manager, i.e. the relevant correlations found out of those four database fields.

The interpretation of these results is done by the user, and each line represents *one* result. This means that *(s)he* has to elect which patterns make sense, which ones are indeed relevant and which ones deserve to be saved in a knowledge base for future use. For instance, consider that the user selects the pattern "*CD_DPSOURCE = 1.3 +- 0.5 <=> CD_ITEM = 2 +- 0 <=> CD_SUPPLYCHAIN = 73 +- 0 <=> in 9% of its values*". First of all, this can be considered a rich pattern as it comprises three of the four fields. Concretely, this means that a correlation among the fields *CD_DPSource*, *CD_Item* and *CD_SupplyChain* was found in the table's records, i.e. in *9%* of the cases the supply chain of id *73* had as a member the enterprise (source) of id *1* in the production of the item of id *2*.



```
□ CD_CONNECTION = 3.8 +- 0.5 <=> CD_DPSOURCE = 2 +- 0 <=> em 9% de seus valores
□ CD_CONNECTION = 5.4 +- 0.52 <=> CD_DPSOURCE = 1.6 +- 0.52 <=> em 22% de seus valores
□ CD_CONNECTION = 5.9 +- 0.86 <=> CD_DPSOURCE = 1.4 +- 0.51 <=> em 31% de seus valores
□ CD_CONNECTION = 6.5 +- 0.53 <=> CD_DPSOURCE = 1 +- 0 <=> em 18% de seus valores
□ CD_CONNECTION = 2.4 +- 0.55 <=> CD_ITEM = 2 +- 0 <=> em 11% de seus valores
□ CD_CONNECTION = 6 +- 0.85 <=> CD_SUPPLYCHAIN = 73 +- 0 <=> em 27% de seus valores
□ CD_DPSOURCE = 1.3 +- 0.5 <=> CD_ITEM = 2 +- 0 <=> em 9% de seus valores
□ CD_DPSOURCE = 1.6 +- 0.53 <=> CD_ITEM = 3 +- 1.9E-7 <=> em 15% de seus valores
□ CD_DPSOURCE = 1.6 +- 0.52 <=> CD_ITEM = 4 +- 0 <=> em 18% de seus valores
□ CD_DPSOURCE = 1.7 +- 0.5 <=> CD_ITEM = 1 +- 0 <=> em 20% de seus valores
□ CD_DPSOURCE = 9 +- 0 <=> CD_ITEM = 3 +- 0 <=> em 6% de seus valores
□ CD_DPSOURCE = 1 +- 0 <=> CD_ITEM = 1 +- 0 <=> CD_SUPPLYCHAIN = 73 +- 0 <=> em 6% de seus valores
☑ CD_DPSOURCE = 1.3 +- 0.5 <=> CD_ITEM = 2 +- 0 <=> CD_SUPPLYCHAIN = 73 +- 0 <=> em 9% de seus valores
□ CD_DPSOURCE = 2 +- 0 <=> CD_ITEM = 1 +- 0 <=> CD_SUPPLYCHAIN = 73 +- 0 <=> em 9% de seus valores
□ CD_DPSOURCE = 2 +- 0 <=> CD_ITEM = 1 +- 0 <=> CD_SUPPLYCHAIN = 76 +- 0 <=> em 4% de seus valores
□ CD_DPSOURCE = 1.5 +- 0.51 <=> CD_SUPPLYCHAIN = 73 +- 0 <=> em 50% de seus valores
```

Figure 5 – Report of the patterns

5. CONCLUSIONS

This paper presented exploratory results on the application of a data mining approach in a VE scenario, aiming at facilitating the decision-making process. *Clustering* technique / *K-Means* algorithm were used in this work.

Preliminary analysis from the results obtained with the software prototype have shown that data mining is a very powerful technique and can indeed support VE managers in decision-making, especially if the tool is integrated in a wider VE management system.

Three important conclusions brought up from this work:

1. Information and knowledge update: the VE data that come from the enterprises can be different from business to business, i.e. new patterns can be created and some

previous conclusions can become out of date along the time. Therefore, the user must run the system "periodically" and delete a pattern which is no longer valid. There is not a specific time to run the system. It is up to the users experience to decide when it is necessary to have a new set of patterns to be analyzed;

2 Interpretation of the results: Post-processed results (figure 5) are surely more interesting and understandable than not processed ones (figure 4). Even so, the interpretation of the patterns still remains a bit difficult, demanding an experienced VE manager to recognize their utility and validity;

3. K-Means algorithm: Databases are most usually composed by numeric and alphanumeric contents. The K-Means algorithm was designed to process only numeric data. It is then important to highlight that, in order to validate this exploratory work, only numeric data deriving from the VE scenario should be considered and used. This would enable the validation of the prototype as well as the developed approach. Efforts are currently being made in order to find a more adequate algorithm to suit the requirements of the envisaged databases and hence to better support decision-makers (VE Managers or even software agents).

Next steps refer also to a deeper validation of the results in a dynamic VE scenario where the system's knowledge base can be continuously updated with new data from the VE partners so that the VE Manager can also play the role of a Knowledge Manager.

ACKNOWLEDGEMENTS

This work was partially supported by CNPq – The Brazilian Council for Research and Scientific Development (www.cnpq.br). This work has been developed in the scope of the IST MyFashion.eu (www.myfashion.org) and IFM (www.ifm.org.br) projects. Special thanks to Solange Rezende and her team at USP-EESC, São Carlos-Brazil.

6. REFERENCES

1. Adriaans, P.; Zantige, D., Data Mining, Addison-Wesley, 1996.
2. Begg, C.; Connolly T., Database Systems: A Practical Guide to Design, Implementation, and Management. 3ed. Addison-Wesley, 2002.
3. Berkhin, P., Survey of Clustering Data Mining Techniques. Accure Software Inc. Website Accessed in November 21, 2003. http://www.accrue.com/products/rp_cluster_review.pdf
4. Chandra, C.; Smirnov, A. V.; Sheremetov, L. B.; Agent-Based Infrastructure of Supply Chain Network Management, Proc. PRO-VE'2000 Conference, pp.221-232, 2000.
5. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth P.; 1996a. From Data Mining to Knowledge Discovery: an overview. In: Advances in Knowledge Discovery & Data Mining, pp. 1-34.
6. Fayyad, U.; Shapiro, G. P.; Smyth, P., From Data Mining to Knowledge Discovery in Databases. AAAIMIT Press, pp.37-54, 1996b.
7. Lavrac N.; Urbancic, T.; Orel, A., Virtual Enterprises for Data Mining and Decision Support: A Model for Networking Academia and Business, Proceedings PRO-VE'2002 Conference, pp.389-396, 2002.
8. MacQueen, J. - "Some methods for classification and analysis of multivariate observations", in LeCam, L. and Neyman, J. eds., Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability, vol 1, pp. 281-297. Univ. of California Press, 1967.
9. Rabelo, R. J.; Pereira-Klen, A. A., Business Intelligence Support for Supply Chain Management, Proc. BASYS'2002, Kluwer Academic Publishers, pp.437-444, 2002.
10. Rabelo, R. J.; Baldo, F.; Tramontin, R. J. Jr., Pereira-Klen, A. A.; Klen, E. R.; Smart Configuration of Dynamic Virtual Enterprises, to be presented in PRO-VE'2004 Conference, Toulouse, Aug 2004.