International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:5, 2007

A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data

H. Baazaoui Zghal, S. Faiz, and H. Ben Ghezala

Abstract**—** Data mining is an extraordinarily demanding field referring to extraction of implicit knowledge and relationships, which are not explicitly stored in databases. A wide variety of methods of data mining have been introduced (classification, characterization, generalization…). Each one of these methods includes more than algorithm. A system of data mining implies different user categories, which mean that the user's behavior must be a component of the system. The problem at this level is to know which algorithm of which method to employ for an exploratory end, which one for a decisional end, and how can they collaborate and communicate. Agent paradigm presents a new way of conception and realizing of data mining system.

The purpose is to combine different algorithms of data mining to prepare elements for decision-makers, benefiting from the possibilities offered by the multi-agent systems. In this paper the agent framework for data mining is introduced, and its overall architecture and functionality are presented. The validation is made on spatial data. Principal results will be presented.

Keywords**—**Databases, data mining, multi-agent, spatial data mart.

I. Introduction

DECISIONAL data processing has been developed since the beginning of the 90s in order to provide, to the decision makers, systems dedicated to the analysis of the data. The decisional databases thus emerged in order to answer the specific needs for multidimensional analysis and the Knowledge Discovery from Databases. The databases and data warehouses become more and more popular and imply huge amount of data which need to be efficiently analyzed. Knowledge Discovery in Databases can be defined as the discovery of interesting, implicit, and previously unknown knowledge from large databases [4][7]. The collected data exceeds human's capacity to analyze and extract interesting knowledge from large databases. Many methods of data mining have been introduced (methods using classification, discovery of association, characterization…). Different user profiles can be implied in data mining system. Agents technology have proven to be particularly useful in several applications particularly when they imply managing user profiles. The Data Mining approach requires some steps to prepare data. One of these steps consists to choose between the different techniques of data mining. The goal of this paper is to present our approach which improves the data mining process by using agent paradigm. The complexity of the process and the importance of the algorithms number and diversity of user profiles incited us to resort to multi-agent systems.

The paper proposes a new approach of data mining system construction. The paper is structured as follows. Section II presents the proposed agent framework for data mining. Section III describes validation and experimentation of the framework. Finally, Section IV concludes the paper and outlines future work.

II. Framework for Data Mining Based Multi-Agent

In the recent years agent technology has found many interesting applications in different domains, principally decision support systems and Internet applications… Several agent system building tools and frameworks have been developed.

A Multi-Agents System (MAS) consists of processes proceeding at the same time, therefore several agents living at the same time, sharing common resources and communicating between them [5]. The MAS must respect the standards of programming defined by the FIPA (Foundation for Intelligent Physical Agents). Our development framework acts as an integrated GUI-based environment that facilitates the design process of a data mining system. It also supports the extraction of decision models from data and the insertion of these models into newly created agents. We present first a generic architecture of data mining framework based multi-agent. After, the architecture of the proposed framework is exposed. This architecture is modular. The main modules will be presented.

A. Generic Architecture of Data Mining Framwork

The figure 1 describes our proposition. The MAS will be present, first, in the level of concepts hierarchy definition. After, when the resulting models from the data mining process are elaborated, the MAS is used to present the best adapted decision to the user. The proposition of decision-aid is stored in knowledge base, which could be useful in a later decision-making. Thus, it must be available among the agents which

H. Baazaoui Zghal is with the Riadi-GDL Laboratory, National School of Computer Science, Manouba, Tunisia (hajer.baazaouizghal@riadi.rnu.tn).

S. Faiz is with National Institute of Sciences and Technology, Tunis, Tunisia (sami.faiz@insat.rnu.tn)

H. Ben Ghezala is with Riadi-GDL Laboratory, National School of Computer Science, Manouba, Tunisia (henda.bg@cck.rnu.tn).

are in the entry of MAS. The identification of the agents will be done in the following section.

Different methods of data mining have been proposed. Each method could have more than one algorithm. During the KDD process applied on geographic data, the user can be opposed to a problem of choosing between methods or even between algorithms. The conception of a system of data mining based multi-agent begin by the description of agents and the interaction between them.

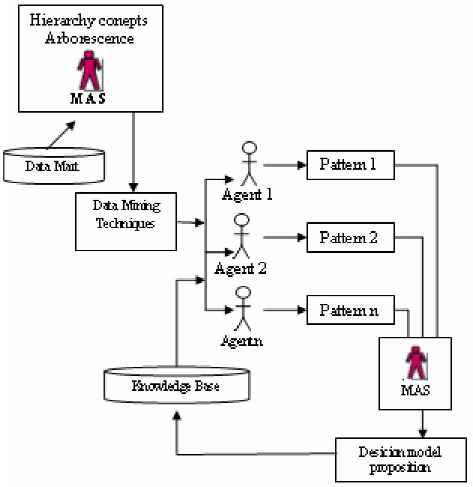

Fig. 1 Generic architecture of data mining framework based multi-agent

B. *Identified Modules*

The proposed framework is composed of modules which is an advantageous architecture because of its evolutivity. The management of user profiles passes by an adaptive system. The cooperation between the system modules is known in advance. The system supports the extraction of the decisional data models and the insertion of these models in newly created agents.

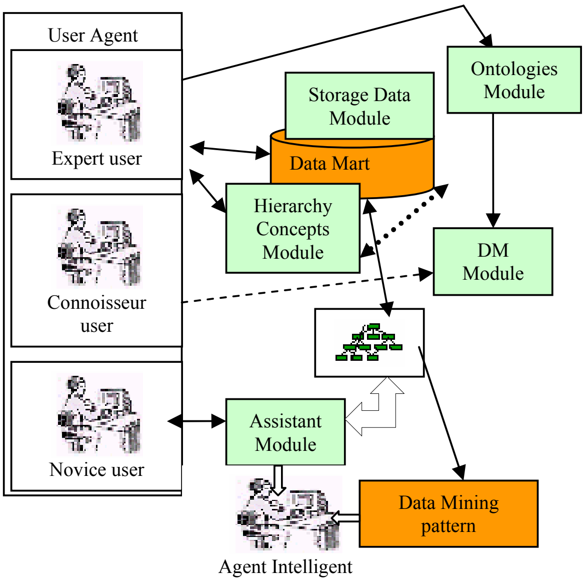

Fig. 2 Organisation modulaire du système proposé

This architecture can be thus enriched progressively by other modules. Identified modules are given by figure 2.

- The Storage Data Module (SDM) : data is prealably stored in data marts. This module permits loading of data [2].
- The Hierarchy Concepts Module (HCM) : recover the hierarchies of concepts already built.
- The Data Mining Module (DMM) : it can be subdivided into two sub-modules : the first for decisional extraction models and the second for recover constructed models.
- The Assistant Module (AM) : dedicated to user assistance essentially the novice user in the data mining phase.
- Ontologies Module : a tool of ontolgy conception, help the user to define these ontologies and communicate the results to the ontology agents.

C. *General Architecture of Data Mining System Based Multi-Agent (DAMSA)*

It is a question of modelling the characteristics of the system and the various possible behaviours of the user according to its profiles. This type of model makes it possible to take into account the various elements and to estimate the consequences of the various actions undertaken by the simulation or the construction of scenarios. Through the interactivity which the decision maker. A system based on such a model allows the evaluation of the strategic actions and consequently the anticipation of the phenomenon and the determination of the adequate strategy. To reach these results, it is difficult to consider a single model of expert system type for example. The field of data mining shows characteristics favourable to a modelling multi-agent: a system of which the browsing by traditional methods leads to too large combinative. Thus the approach multi-agent is adopted. Various methods of dated mining were defined in literature. For each one, there is more than one algorithm. During a process of extraction of knowledge starting from a database the user can be confronted with a problem of choice between methods or even in algorithms. The design of a SMA starts with the definition of the agents which compose it and the modes of interactions between them. The various agents identified on this level are the following:

- an evaluator agent: the user needs to choose the best result which meets its needs
- a comparison agent: who deals with comparing the various results obtained starting from the various launched algorithms
- a coordinator agent: the data have been stored in data marts. This agent will extract data to provide them to the agents
- one or more agents for data mining: carry out the algorithms of data mining according to the choice which was made. There will be thus an agent generalization, an agent classification, an agent characterization...these various agents will be carried out at the same time to allow the agent appraiser to do his work
- an interface agent: who according to profiles of the user will play the role of interface with on the one hand the

agent dated mining and on the other hand the softwares that will be used to display obtained results. It proposes to the user according to his profiles the possibility of formulating his requests and of displaying the result in the form of rules.

We distinguish three profiles of users:

- an assistant agent: which guides the user in his choices. Distinguished user profiles are: novice, connoisseur and expert:
- a novice user: is a user who reaches the system for the first time and on which no history does exist. The system must guide him and provide him prepared solutions by calling upon the agent assisting ·
- an expert user: can be the administrator. He could build a new Data Mart starting from the data sources, to prepare the requests and results of extraction of knowledge for a later use.
- a connoisseur user has an intermediary profile. It indicates that this user profile has knowledge of the field and can identify his needs. The system must give him the possibility of formulating its requests. If he could not be satisfied, an assistant is called.

The following stage of the design of a MAS consists in defining the proportions of the cognitive and reactive characters constituting the agents of this system. In the case of cognitive agents, the cooperation between several agents supposes that they take an active part in the realization of common goals. The systems of reactive agents are based on an emergent design of the intelligence.

The general architecture is given by figure 3 which illustrates the interaction between the described agents.

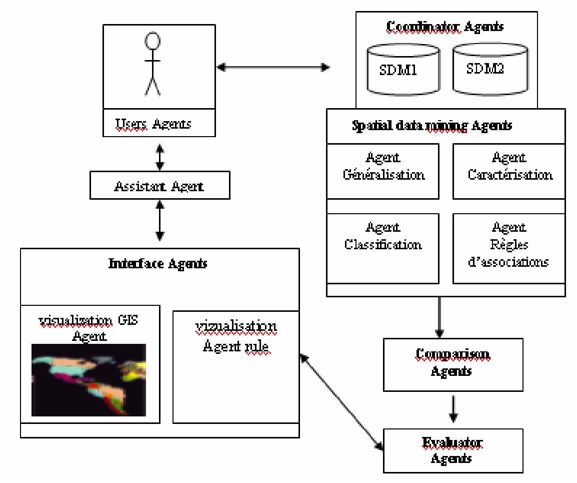

Fig. 3 General architecture of DAMSA

D. *Organisational Design of the Framework*

As organisational conception we use AALAADIN, which proposes a methodological framework and a multi-agent platform [5][6]. According to method AALAADIN it is necessary to identify the various agents, groups and roles, as well as the interactions between them.

The agent groups are the followings. Figure 4 shows the distribution of the various agents in their respective group is indicated by the diagram of classes.

- Group coordinators agents
- Group of the data mining agents
- Group of comparison agents
- Group of evaluator agents ·
- Group of user agents
- Group of interface agents

The agents able to play roles in the bellows groups are identified by the class diagram given in figure 4. Each agent corresponds to a class. The agents exchange messages to translate the interactions between them. These interactions between agents are modelled with the help of diagrams of sequences.

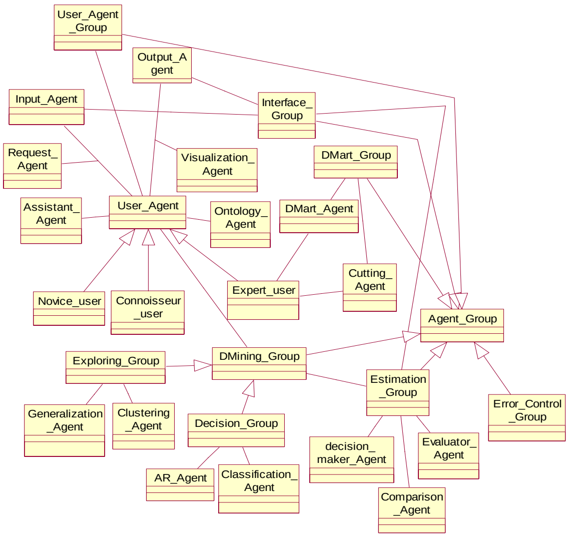

Fig. 4 Class Diagram of DAMSA

III. EXPERIMENTATION IN SPATIAL DATA CONTEXT

The experimentation of our approach has been made using spatial data. This section, first briefly presents characteristics of spatial information. After, we present implemented data mining algorithms.

This section briefly presents characteristics of spatial information. Geographic data concerns the spatial object aspect (the coordinates of the object) and non-spatial object (the different descriptive attributes of the object). Non-spatial object defines the description of the geographical entities which can be stored in a traditional relational databases where the attributes point to the spatial description of this entity.

A. *Spatial Data*

Spatial data concerns the spatial object aspect (the coordinates of the object) and non-spatial object (the different descriptive attributes of the object). Non-spatial object defines the description of the geographical entities which can be stored in a traditional relational databases where the attributes point to the spatial description of this entity.

The spatial data represents a phenomenon on a territory at various times of observation, the problems of integration of this data are at the same time semantic, temporal and spatial especially if the various data sources were not acquired in the objective to supply a spatial data warehouse [9]. The geographical data show particular characteristics which are necessary to take into account during the modeling and the integration of these data : semantic richness, precision of the procedures and the multiplicity of the geometrical representations. These spatial data often overlapping and are connected to the alphanumeric items. The overlap comes owing to the fact that a city can be included in a region, roads and rivers cross this region. Data such as the number of inhabitants of a city or the width of a road must be connected to the corresponding geometrical data. Thus, the geographical types of data are varied and complex to model.

[10] was the first to propose a framework for spatial data warehouses. Few prototypes were developped to support spatial information [3][11][12][13][14].

B. Implemented Algorithms

The concept of association rules introduced by Agrawal and al [1] was extended to the spatial databases by [8].

Definition: A spatial association rule has the following form:

P1 -> ... ->Pm =>Q1 ->... ->Qn (s%, c%)

Where at least one of the predicates is a spatial predicate. Various kinds of spatial predicates can constitute a spatial association rule. We take here some spatial predicates.

- Predicate including distance information: close_to and far_away
- Predicate including topological relations: intersect, overlap and disjoint
- Predicate including spatial orientation: left_of and west_of.

S% is the support, is the probability of the objects satisfying the antecedent of the rule, and the confidence C% of the rule, indicates that c% of the objects satisfying the antecedent of the rule will also satisfy the consequent of the rule. Obviously, most people are only interested in the patterns that occur relatively frequently (with large supports) and the rules that have strong implications (with high confidence). The rules with large supports and high confidence are strong rules. Based on this, two kinds of thresholds: minimum support and minimum confidence can be introduced, by a users or experts. Moreover, since many predicates and concepts may have strong association relationships at a relatively high concept level, the thresholds defined should be at different concept levels. Therefore, it is expected that many spatial association rules be expressed at a relatively high concept level.

We present here the algorithm for discovery of spatial association rules based on the definition of spatial association rules. The discovery of spatial association rules algorithm presents the descriptive attributes as being an organization of concept hierarchies. These hierarchies then guide the extraction of the association rules. He takes in input:

1. A database composed of three parts: a Spatial DataBase containing a set of spatial objects; a Relational DataBase describing non-spatial properties of spatial objects; and a set of concept hierarchies.
2. A query consists of three parts: a reference class; a set of task-relevant classes for spatial objects; a set of task-relevant spatial relations.
3. Two thresholds: minimum support and minimum confidence for each level of description.

The presented algorithm starts with an extraction of the interesting objects based on the user request. The second step is the extraction of the primary predicates. Which will be filtered by respecting the minimum support carried is out. The fourth step is consisting on refining the filtering predicates by an iterative way, while going top-down in the hierarchy to the desired level. Finally, the spatial association rules are extracted. The result of the algorithm will be a set of spatial association rules at various levels.

C. Experimentation and Evaluation

In object of experimentation the proposed approach we developed a prototype called CASAMME for Computer Aided Spatial Agent Mart Mining Environment, which integrates the process of design and implementation of a Spatial Data Mart. The library of models includes essentially the meta model, the exploration models and the sequence model. The validation was made through the instantiation of the different models by using road accidents databases [14].

A comparison was carried out between a manual approach and an approach using CASAMME environment, offering an automatic generation of the MDG and a prepared process for the data mining algorithms. The phase of development of multidimensional conceptual model facilitates considerably the phase of data mining. The cost of modeling in a number of days of the application is more significant for the approach suggested, but makes it possible to gain considerably on the costs and the quality of the phases of generation and exploration of the geographical data. The differences in costs are amply absorbed by the automated phases (generation, extraction for the MDG and preparation of the results of the algorithms for the excavation of geographical data). The phases of updates and maintenance will accentuate considerably the variations in favour of the approach suggested. The graphics of figure 5 represents an estimate of the cost in a number of day of the whole of the phases compared to the data file studied.

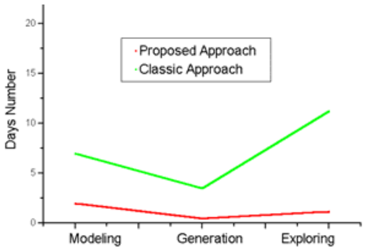

Fig. 5 Comparison between proposed approach and classic approach

IV. Conclusion and Future Work

Motivated by the increasing of complexity and numbers of data mining techniques, we present in this paper the decision support and data mining process to identify the main difficulties. After we describe the proposed architecture of DAMSA to improve the data mining in this sense we present our Computer Aided Spatial Agent Mining Mart Environment (CASAMME). We improved the exploration process and the knowledge extraction in CASAMME by using multi-agents system in order to take into account the user profile and style.

In [2] we presented a case tool for spatial data marts design and generation, which we extend to agent mining environment.

In this paper, we presented a new approach of data mining based on multi-agent system. In this approach, we used a multi-agent system to improve the execution time at different levels. In addition, we presented experimentation for this approach. The discovery of spatial knowledge is a process that needs a long execution. On the other hand, the improvements made in the spatial field are to improve the quality of extracted rules. A perspective to our framework consists to format the extracted knowledge in Predictive Modeling Markup Language (PMML) format.

REFERENCES

[1] Agrawal R., Gupta A., Sarawagi A., *"Modeling Multidimensional Database*s", ICDE'97 pages 232-243. IEEE Press, 1997.

[2] Baazaoui H., Faiz S. et Ben Ghezala H. (2003), CASME : A CASE Tool for Spatial Data Marts Design and Generation, 5th International Workshop on Design and Management of Data Warehouses (DMDW'2003), Septembre, 2003, Berlin.

[3] Bedard, Y., 2002, Geospatial Data Warehousing, Datamart and SOLAP for Geographic Knowledge Discovery, Université de Muenster, Germany, 2002.

[4] Fayyad U.M., Piatetsky-Shapiro G., Smyth P. (1996), « From Data Mining to KDD : an overview », AAAI/MIT Press, 1996.

[5] Ferber J. (1995), Les Systèmes multi-agents vers une intelligence collective, interEditions, France.

[6] Gutknecht (O), Ferber (J) and Michel(F), Integrating tools and infrastructures for generic multi-agent systems, Proceedings of the Fifth International Conference on Autonomous Agents, 2001.

[7] Han J. et Kamber M. (2002), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Canada, 2002.

[8] Koperski (K.) et Han (J.). (1995) Discovery of spatial association rules in geographic information databases. 4th Int. Symp. Advances in Spatial Databases, SSD, vol.951,pp. 47{66. Springer-Verlag, 1995.

[9] Laurini R. et Thompson D. (1994), *Fundamentals of Spatial Information Systems*, Academic Press, London.

[10] Lu W. et Han J. (1993), Discovery of general knowledge in large spatial databases, *Far East Workshop on GIS*, Singapore, Juin 1993.

[11] Marchand P., Brisebois A., Bédard Y. et Edwards G. (2004), Implementation and evaluation of a hypercube-based method for spatio-temporal exploration and analysis, Journal of the International Society of Photogrammetry and Remote Sensing 2004.

[12] Rao F., L. Yu Z., Li Y. et Chen Y. (2003), Spatial hierarchy and olap-favored search in spatial data warehouse, DOLAP, 2003.

[13] Wang, F. Pan, D. Ren, Y. Cui, D. Ding, et W. Perrizo (2003), Efficient olap operations for spatial data using peano trees, 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003.

[14] Zeitouni K.et Yeh L., Les bases de données spatiales et le data mining spatial, actes des Journées sur le Data Mining spatial et l'analyse du risque, Versailles (2000).