CASME : A CASE Tool for Spatial Data Marts Design and Generation

Hajer Bâazaoui Zghal[1], Sami Faïz[2], Henda Ben Ghézala[1]

[1] RIADI-GDL Laboratory, ENSI Campus universitaire Manouba, Tunisia
hajer.baazaouizghal@isd.rnu.tn
henda.benghezala@isd.rnu.tn

[2] LTSIRS Laboratory, ENIT
INSAT, BP. 676-1080, Tunis,Tunisia
sami.faiz@insat.rnu.tn

Abstract. Geographic Information Systems (GIS) showed their insufficiencies in front of complex requests for decision-makers. Resulting of the association of the databases and the decision-making systems the decisional data processing was developed since the beginning of the 90th as a new way. The decisional databases thus emerged in order to answer the specific needs for On-Line Analytical Processing (OLAP) and data mining. Extensions were made to make appropriate the analysis and the algorithms to specificities of the handled spatial data. This paper describes the modeling and implementation of Spatial Data Mart (SDM). We define a formal framework for the progressive construction of spatial data warehouses by assembling these SDM. Our approach includes a meta model for SDM construction. The construction is done in accordance with the UML meta model. After the validation step, construction is followed by an automatic generation of the spatial data mart in Spatial Oracle. A CASE tool, called CASME (Computer Aided Spatial Mart Engineering), constitutes the interface through which the user will have to carry out the process.

1 Introduction

Spatial databases store information about the position of objects in space. The spatial information becomes more and more popular (maps, satellite images…) and creates huge amount of data which need to be efficiently analyzed. First GIS (Geographic Information System) appears to integrate spatial and non-spatial data to yield the spatial information that is used in decision-making. After decision-making becomes more exigent inefficient and appears the topic of data warehousing. [12][9] were the first to propose a framework for spatial data warehouses. In this paper, we concentrate on modeling and implementing spatial data warehouse by assembling spatial data marts. Design and implementation of spatial data warehouse is a highly complex engineering task that calls for methodological support. A tool environment helping the designer in its specification and implementation is proposed. A CASE tool, called CASME (Computer Aided Spatial Mart Engineering), is proposed too.

The paper is organized as follows. Section 2 presents briefly the spatial information. Section 3 describes the spatial decision support. In this section we first present related works, after we present extensions to spatial information field. Section 4 presents our system architecture and our approach for modeling and implementing SDM. Section 5 presents the suggested CASE tool for design and implementing SDM, its architecture and validation step. The paper concludes with a presentation of our current and future work.

2 Spatial information

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.

The geographical 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 [5]. 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. The modeling of the data is expressed by successive states of models symbolizing various levels of abstraction between the recorded and reality data. The modeling of the treatments often results in data diagrams representing the repository of data, flows and the procedures of organization. UML class diagram, represented by figure 1, is a complex object generalizing any type of geographical data.

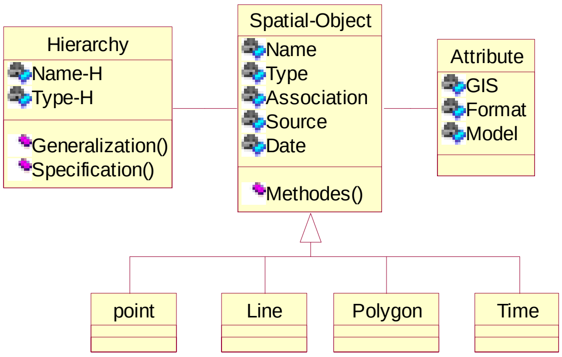

Figure 1. Composition of geographical object

- A geographical object is characterized by its name and its source, an attribute dates representing the notion of the successive versions in time. Each object can be composed by several other objects (components)
- A composing object is identified by its name, its type, its size and its site (physical addresses). The type of data of the under-object (point, line, polygon, image) determines the methods of analysis which will be used later on (ie. spatial OLAP or spatial data mining)

Considering the specifities related to this type of data, construction of spatial data warehouse must take account of them.

3 Spatial Decision Support

Decisional data processing have been developed since the beginning of the 90th 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 (OLAP) and the knowledge discovery from databases (data mining).

3.1 Related work

Different proposals have been made regarding how to represent a conceptual multidimensional schema to model data warehouses and data marts [8]. A data model is a set of concepts that can be used to describe the structure of database. A categorization of data models and a comparison of several multidimensional data models is proposed in [1]. Different levels can be identified : conceptual, logical, physical and formal.

- Research efforts at conceptual level : models contain concepts which are close to the user. [8] present a graphical conceptual model (DFM) for data warehousing, besides a methodology to obtain a multidimensional schema from the operational schemas (E/R or Relational). [16] describe an Object-Oriented conceptual model based on a subset of UML. A query notation is also presented. [17] define a set of user requirements for a « data warehouse conceptual model" starER model.
- Research efforts at logical level can resumed by [10] which describes the implementation of multidimensional model on a relational DBMS. [Vassiladis et al. 1999] give a survey of logical models for OLAP databases.
- Research efforts at physical level explain how a data cube could be implemented. explain the implementation of the cube (MOLAP database).

These models do not take into account the specificity of spatial information. In the next part, we present the extensions of these efforts to spatial information field.

3.2 Extensions to spatial information field

Extensions were made to make appropriate the analysis and the algorithms to specificities of the handled spatial data. These concepts were extended to the field of spatial information by [12]. This work was devoted to the application of KDD process to spatial data. Later, many works concerning the spatial OLAP were born [11][4] [15].

3.2.1 Spatial data mining

Spatial data mining is an extraordinarily demanding field referring to extraction of implicit knowledge, spatial relationships, which are not explicitly stored into geographical databases [7]. In fact, the collected data can exceed human's capacity to analyze and extract interesting knowledge from large spatial databases. Many methods of knowledge discovery in geographic database have been introduced (methods using classification, discovery of spatial association, characterization…). For knowledge extraction a method of *generalization* has been defined. Objects often contain detailed information at primitive concept levels. It is necessary to summarize a large set of data and present it at a high concept level. For example, one may like to summarize the detailed data of results of baccalaureate and diploma in a region and present its general results pattern. This method is an extension of spatial data of generalization based on attribute-oriented induction. It consists on substituting detailed values by less detailed values to obtain the desired level of detail after aggregate and count identical obtained occurrences. The generalization supposes the existence of background knowledge in the form of concept hierarchies. It first takes a data mining query and collects the set of relevant data in a database. Then, data generalization is performed by increasing the generalization hierarchies and summarizing the general relationships between spatial and non-spatial data at higher concept levels. [12] extended attribute-oriented induction to spatial databases and presented two algorithms, *spatial-data-dominant generalization* and *non-spatial-data-dominant generalization*. As methods of knowledge discovery in geographic databases, we can find Generalization-Based Data Mining [11][19], methods using Clustering [9][6], discovery of spatial association rule [2].

3.2.2 Spatial OLAP

Spatial OLAP (SOLAP) can be defined as a visual platform built especially to support rapid and easy spatio-temporal analysis and exploration of data following a multidimensional approach comprised of aggregation levels available in cartographic displays as well as in tabular and diagram displays [5]. SOLAP are meant to be client applications sitting on top of a multi-scale spatial data warehouse. However the non-expert can also see them as a new type of user interface for multi-scale GIS applications and web mapping.

By the association of the cartographic representation and OLAP navigation, the user moves in the multidimensional structure and obtains representations of the data via cartographic, tabular posting or statistical diagram which are a function of dimensions, measurements and the selected levels of hierarchy. The spatial data must be organized in *fact*, *dimensions* and *measurements*. Extensions were made to describe these characteristics when spatial data are handled. These SOLAP tools offer a number of operations making it possible to navigate in the data, such as the slicing and dicing, pivoting, roll-up or drill-up and drill-down. Generally, SOLAP applications have a geometric representation for each spatial dimension. These last will be visualized in a given scale.

3.3 Existing prototypes

Few prototypes were developed to support spatial information:

- *Geominer prototype* comprises five functional modules: characterizer, analyzer, comparator, associator and classifier. Queries formuled GMQL offers to the user interrogation of spatial data [9].
- *SOLAP prototype* is based on a multidimensional database structures as used in data warehouses, OLAP servers and data mining tools. Users can explore various levels of details: communal level at the regional level while passing by the level of the department, following various topics, various times and various elements of analysis. Display included different possibilities: the statistical thematic map, statistical diagrams and tables [15].
- *Prototype of analysis of the risk of accidents* integrates mainly the algorithms of data mining, adapted to the needs for the analysis of the road risk: generalization within two alternatives spatial and non spatial, and characterization [19].

4 Our Approach

We propose to improve the process of spatial data decision-making by using decisional system techniques. This system is based on the approach of the data warehouses and data marts which store the relevant data for the decision-makers. This process will include the storage, the analysis and the exploration within spatial data. With an aim of improving the process of decision-making we install the decisional information system by integrating modeling and exploration of spatial data marts.

The general architecture of the system comprises data sources from which the data mart is constructed in accordance with our meta model described in this paper. The object-oriented paradigm was besides selected for the different models and justifies ourselves by its adequacy for the integration of complex heterogeneous sources [14] especially when we manipulate spatial data. In fact, the result is closer to the user conception. Every object or class modeled will have a correspondence with some real entity. The data of the SDM are organized multidimensionally in order to support the processes of Spatial OLAP and Spatial Data Mining. We adopted a semi-formal specification using Unified Modeling Language (UML) [13], supporting a rich semantics and powerful concepts resulting with the approach object (such as the heritage, aggregation...). The conformity of the instances of models will be checked by a DTD (Document Type Definition) developed in XML. The exploitation of the data for the decision-making is carried out through tools of handling and interrogation. The process of exploitation is described by a uses case diagram. Figure 2 presents the global schema of the Spatial Data Mart (SDM).

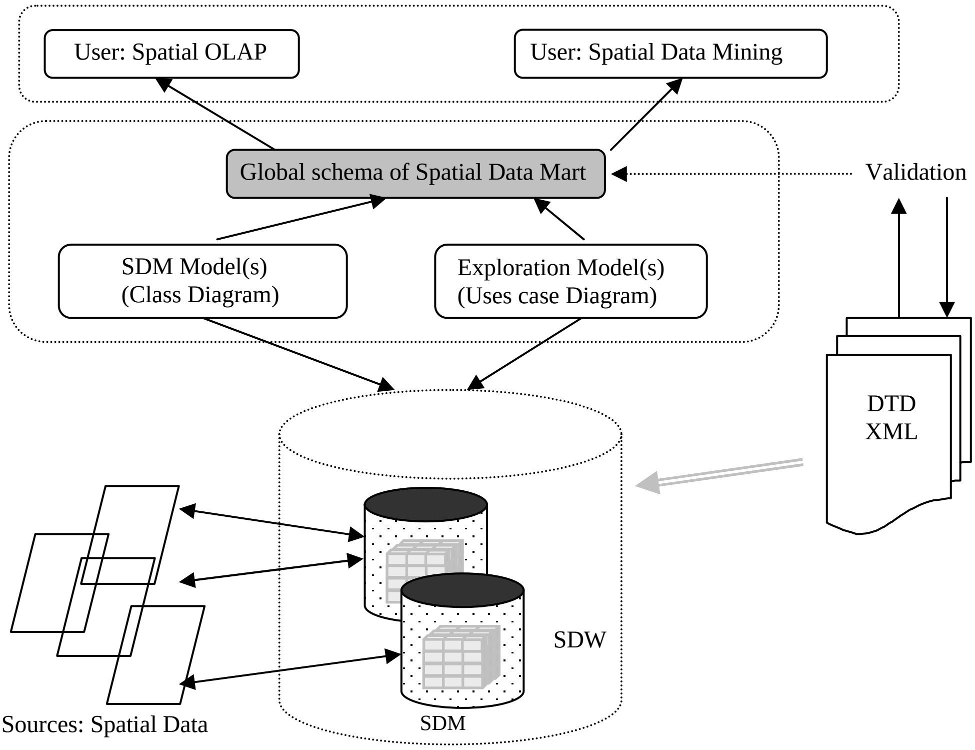

Figure 2. Structure of our system

5 Specification of Meta model for Spatial Data Mart Construction

The global schema of the SDM is the most important part of SDM model (cf. figure 2). We propose here a modeling framework into which generation of the SDM in *Spatial Oracle* and exploration techniques are being embedded. For details concerning exploration of spatial data, we refer to [3]. We proposed an object-oriented meta model of a SDM. Our approach has been successfully deployed for object oriented software engineering tools. The CASME user manipulates the model through graphical interfaces. Each interface represents a certain aspect of the SDM design.

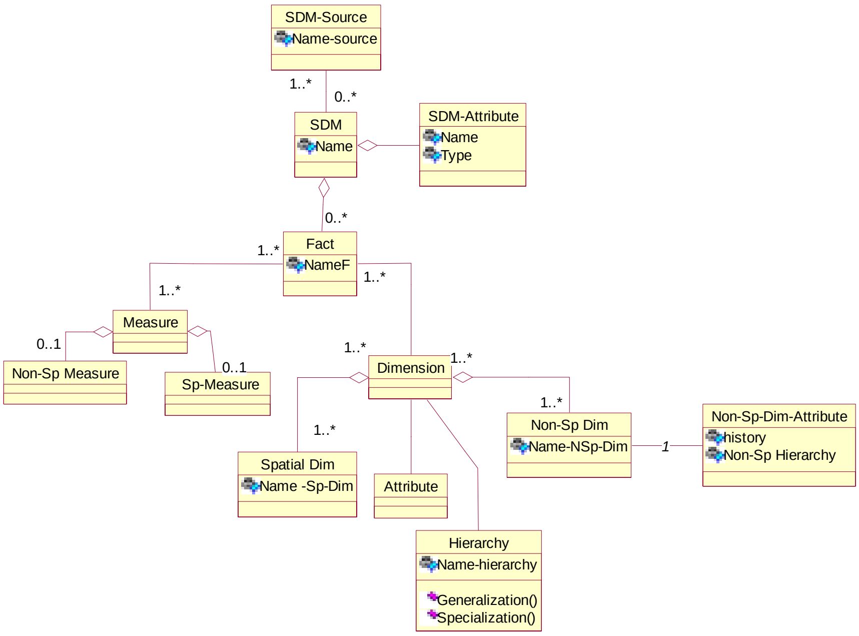

Figure 3. Meta Model for Spatial Data Mart Construction

We specify a meta model for the construction of spatial data mart (cf. figure 3). From the global source which can be represented through traditional model dedicated to the management of the spatial data we build the SDM represented in our model by a class SDM. This class models an entity extracted from the global source, made up of a source class (subject-oriented). The SDM is identified by a single identifier which does not change during its cycle of life. This SDM is also characterized by its Name. It is specialized for a type of activity, a type of analysis or a user group. It contains only information which is considered to be important to answer the objectives of the users. It is composed of one or more facts characterized by their respective names. Each SDM is characterized by its multidimensional schema formed by facts, dimensions and measurements. Dimensions and measurements can be spatial or non–spatial. The multidimensional analyses use hierarchies to handle the analyzed values at various levels of detail. We propose to model the hierarchy dependences a class called Hierarchy. It is significant to insist on the importance to keep a history of the various levels of hierarchy for dimensions and measurements being given the nature of the spatial data. A dimension can be visualized as being a simple non–spatial dimension whereas at a certain time in the hierarchy it was spatial and that following generalizations which it became non–spatial and the user must know the history of dimension. Same thing for measurements (we deal with granularity of measurements). A SDM consists of dimensions and measurements whose parameters are organized according to a certain hierarchy. The fact corresponds to the topic or subject of analysis. A SDM can include one or more facts presenting the various topics which interest the user.

In an object context a fact is represented by a class. It is indicated by its Name and the attributes which measuring the activities contained in the store. These measurements are consequently gathered by topic. A measurement can be numerical when it contains only numerical data such as the returned monthly one of an area. A measurement can also be spatial: a collection of pointers to spatial objects. For example the case of a generalization of areas having temperature and precipitations in the same cell. Thus, measurement forms a collection of pointers to the corresponding areas.

Bedard distinguishes three possible choices regarding the computation of spatial measures in spatial data cube construction [4]:

- Collect and store the corresponding spatial object pointers but do not perform precomputation of spatial measures in a spatial data cube.
- Precompute and store some rough estimation of the spatial measures in a spatial data cube.
- Selectively precompute some spatial measures in a spatial data cube.

Four types of dimensions can be identified: non-geometric spatial dimension, geometric to non–geometric spatial dimension, entirely geometric dimension and temporal dimension. Bedard distinguishes three types of dimensions: temporal dimension, spatial dimension and thematic dimension

- *Temporal dimension.* Describes the organization of the time of study in conformity with the user needs. It can be a simple hierarchy (day, month, period and year) or complex (in the form of temporal models describing the evolution of the phenomena in time) but its instance does not include specific members.
- *Spatial dimension.* Describes the representation of the territory surface, it could comprise specific members, but that would restrict the cartographic representation of the data at one moment given according to only valid territorial cutting to this moment.
- *Thematic dimensions.* Several logical models can be defined for these dimensions.
- *Non-geometric spatial dimension.* It is a dimension which contains only non–geometric data. The administrative units can for example be built for a spatial data warehouse like dimension containing only nominal data making it possible to locate a phenomenon in space. Such a dimension can start with the names of the municipalities and their generalization can, also be non-geometric (ie: country or department). Such a case can be implemented without having resorts to the concepts of spatial data warehouse as long as a cartographic representation is not necessary.
- *Geometric to non-geometric spatial dimension.* It is a geometric dimension which becomes non-geometric following a generalization.
- *Entirely geometric dimension.* It is a dimension of which all the levels and even close generalization are and remain geometric. Let us take the example

of the polygons of precipitation of department, they are geometric data and even following generalizations like 10-100 millimeters, 100-200 millimeters keep their geometric character.

- *Temporal dimension.* It is difficult to imagine information systems not including a temporal reference, such as the dates and durations of studied activities. Temporal dimension also constitutes strategic information to predict a future behavior or to explain the causes of the current state of the things. Some GIS manages the temporal data like descriptive data, which is not adapted to multidimensional representation having among its dimensions temporal.

6 CASME : Computer Aided Spatial Mart Engineering

Figure 4 describes the architecture of our CASE tool, called CASME for Computer Aided Spatial Mart Engineering. The process of design and implementation of a SDM includes the 5th following steps:

- *First step*. User have to choose a particular spatial topic.
- *Second step*. User have to specify facts, dimensions and measures.
- *Third step*. User obtains corresponding model which must been in conformity with our meta model (cf. figure 3).
- *Fourth step*. Generation of the SDM in Spatial Oracle.
- *Last step*. Development of the exploration functions.

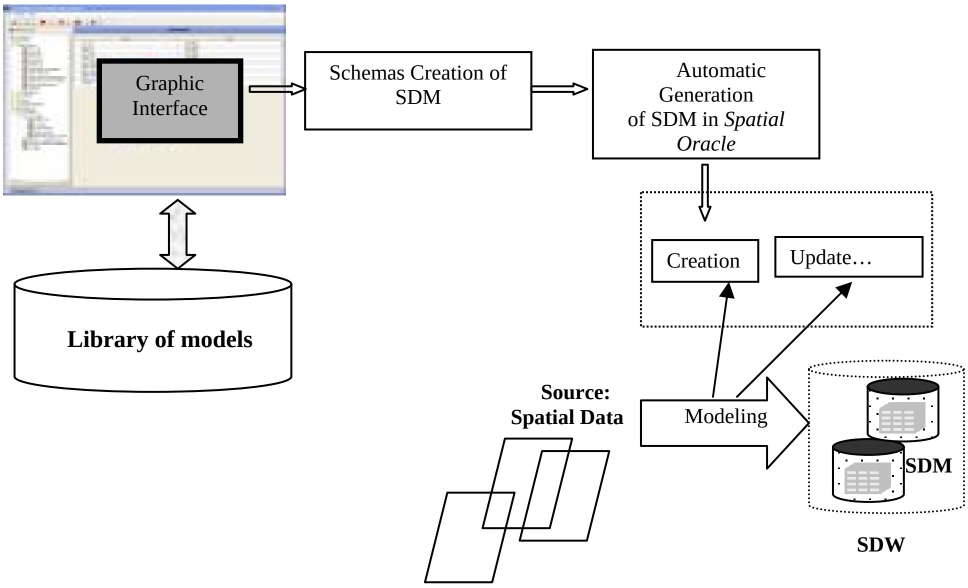

Figure 4. Architecture of CASME tool

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.

7 Conclusions and future works

Motivated by the increasing of spatial information volume, we present in this paper the spatial information, the spatial decision support, the related works and their extensions to spatial field. In contrast to existing approaches and prototypes we present architecture system specific to spatial information: our approach for modeling and implementing Spatial Data Mart (SDM) is exposed. This approach is based on a meta model, which is being supported by a Computer Aided Spatial Mart Engineering (CASME) environment. This environment generates the implementation of the modeled instance. We are now improving the exploration process and the knowledge extraction in CASME by using multi-agents system in order to take into account the user profile and style.

References

1. Abello A., Samos J., and F. Saltor. Understanding Analysis Dimensions in a Multidimensional Object-Oriented Model, In 3rd International Workshop on Design and Management of Data Warehouses (DMDW). SwissLife, 2001.
2. Agrawal R., Gupta A., Sarawagi A., "*Modeling Multidimensional Databases*", ICDE'97 pages 232-243. IEEE Press, 1997.
3. Baazaoui H., Faiz S., Ben Ghezala H., “Exploration Techniques of the Spatial Data Warehouses: Overview and Application to Incendiary Domain”, Acts of International Conference On Computer Systems and Applications (AICSSA 2001), Beirut, Liban, 2001.
4. Bedard Y.: “Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery” Authors Yvan Bedard, Centre for Research in Geomatics, Laval University, Quebec City, Canada, 1999.
5. Bedard, Y., 2002, Geospatial Data Warehousing, Datamart and SOLAP for Geographic Knowledge Discovery, Université de Muenster, Germany, 2002.
6. Ester M., “The database approch to spatial data mining”, In Proc. Journées du data mining spatial et analyse du risque, Paris, France, 2000.
7. Faiz S., “*Systèmes d’Informations Géographiques : Information Qualité et Data Mining*”, book in french, Editions CLE, Tunis, 1999.
8. Golfarelli M., Maio D., and Rizzi S., “The Dimensional Fact Model: a Conceptual Model for Data Warehouses”, Int. Journal of Cooperative Information Systems, 1998.

9. Han J., Koperski K., Stefanovic N., « GeoMiner : A system prototype for spatial Data Mining », Proc. 1997 ACM-SIGMOD Conf. on Management and Data (SIGMOD'97), Arizona, 1997.
10. Kimball R., The Data Warehouse Toolkit: Practical techniques for building dimensional data warehouses. John Wiley. 1996.
11. Koperski K.: "A progressive refinement approach to spatial data mining", Doctor of Philosophy of computing Science, University of Simon Fraser, British Columbia, Canada, 1999.
12. Lu W., Han J.: "Discovery of general knowledge in large spatial databases", *Far East Workshop on GIS,* Singapore, 1993.
13. Muller P-A., *"Modélisation objet avec UM*L", *e*d. Eyrolles, ISBN 2212091222, 2000.
14. Pedersen T. B., Jensen C. S., "Multidimensional data modeling for complex data", In Proc. of 15th Int. Conf. on Data Engineering (ICDE), IEEE Computer Society, 1999.
15. Rivest,S., Bédard, Y. & Marchand P., 2001, Towards better support for spatial decision-making: Defining the characteris Spatial On-Line Analytical Processing (SOLAP), Geomatica: The journal of the Canadian Institute of Geomatics, 2001
16. Trujillo J. C., M. Palomar, and J. Gomez. Applying Object-Oriented Conceptual Modeling Techniques to the Design of Multidimensional Databases and OLAP applications, In Proc. of 1st Int. Conf. on Web-Age Information Management (WAIM), Springer, 2000.
17. Tryfona N., Busborg F., Borch Christiansen J.G., *"starER: A Conceptual Model for Data Warehouse Desig*n", (DOLAP' 99), Kansas city, Missouri, USA, 1999.
18. Vassiliadis P., Sellis T., A Survey of Logical Models for OLAP Databases. SIGMOD Record, December 1999.
19. Zeitouni K, "Introduction aux bases de données spatiales et au data mining spatial", In Proc. Journées du data mining spatial et analyse du risque, Paris, France, 2000.