TRAJECTORIES' CLASSIFICATION TO ENHANCE DECISION MAKING

Wided Oueslati and Jalel Akaichi
ISG - University of Tunis
41, Avenue de la Liberté
Cité Bouchoucha Le Bardo 2000
TUNISIA
widedoueslati@live.fr
jalel.akaichi@isg.rnu.tn

ABSTRACT

Decision support systems such as trajectory data warehouse are based on the fast and the best decisions that are becoming the means of success in many domains. However, the best and the fast decisions requires consistent and meaningful data, therefore the idea is to select only specific and interesting data from the whole trajectory data stored in the trajectory data warehouse. The classification or the clustering of trajectory data is the best solution in this case. The aim of this paper is to propose a new classification of trajectories of moving objects based on some criteria. The classification will appear in the conceptual modeling of the trajectory data warehouse and will play an important role in the trajectory data analysis process.

KEYWORDS

Mobile objects; Trajectory classification; Trajectory data; Trajectory data warehouse modeling.

1 INTRODUCTION

Making the best decisions in the right time, and acting efficiently require the fact that the decision maker must have on the one hand a set of sufficient, available, reliable, relevant, precise and recent data and on the other hand a set of subject-oriented, integrated, time-variant and non volatile data. This can be ensured in one hand by recent and incredible evolution of positioning technology, mobile devices, wireless networks, etc, and on the other hand by business intelligence techniques such as data warehouses.

In fact, these technologies make the collection of data and their updates feasible and in real time while moving. Moreover, these new technologies contribute to the creation of new jobs related to the collection, the management and the storage of data in order to have them exploited for analysis in depth, by new customers such as investors and new opportunities seekers.

In our inspiring example, Mobile Professionals are the Mobile Information Collectors (MIC). Those latter are in charge of the information collection. They are equipped of mobile devices, and transportation means equipped by sensors. They move through multiple trajectories and generate a large amount of data. Such data result from the trajectories of moving objects which are the MIC, that's why it is referred to as Trajectory Data (TD). Those latter are huge and they can be classified in order to reduce the analysis process. In fact, the classification has played an important role in the trajectory data analysis process that is why an efficient classification for trajectory data is required.

The trajectory classification will impact the trajectory data warehouse conceptual modeling. For this reason, this paper is organized as follows: In section 2, we will present different research works related to the conceptual modeling methodology, then we present research works related to trajectories' classification. In section 3, we will present the trajectory data warehouse conceptual modeling for our running example before the application of the classification algorithm. In section 4, we will propose a classification algorithm for trajectories then the new trajectory data warehouse model. In section 5, we will compare the analysis of trajectory data

before and after classification. In section 6, we will summarize the work and propose some perspectives to be done in the future.

2 STATE OF THE ART

We will present in this section different research works on conceptual modeling methodology and trajectories' classification and clustering. In the literature, we can find three categories of conceptual approaches; the top down approach, the bottom up approach and the middle out approach. The difference between those latter is situated in the starting point. In fact, each approach has its own starting point such as users' needs, data marts or both users' needs and data marts. Concerning the top down approach, this latter has to answer users' requirements without any exception. It is very expensive in term of time since it requires the whole conceptual modeling of the DW as well as its realization and it is difficult because it requires the knowledge in advance of dimensions and facts [1]. In this category, authors of [2] present a Multidimensional Aggregation Cube (MAC) method. This latter insures the construction of a multidimensional schema from the definition of decision makers' needs but the defined schema is partial because it describes only the hierarchies of dimensions. The goal of MAC is to supply an intuitive methodology of data modeling used in the multidimensional analysis. It models real world scenarios using concepts which are very similar to OLAP. In MAC, data are described as dimensional levels, drilling relationships, dimensions, cubes and attributes. Dimension levels are a set of dimension members. Those latter are the most detailed modeling concepts and they present real world instances' properties. Drilling relationships are used to present how one level element can be decomposed of other levels' elements. The dimension paths present a set of drilling relationships which are used to model a significant sequence of drill down operations. Dimensions are used to define a significant group of dimension paths. This grouping is essential to model semantic relationships. Cubes are the only concept which associates properties' values with real measures' values. They insist on the complex hierarchy structure defined by dimensions. The top down approach can be used in the Goal-driven methodology [3]. In fact, this latter focuses on the company's strategy by involving the executives of the company. For the bottom up approach, this latter consists on creating the schema step by step (data marts) until the obtaining of a real DW [1]. It is simple to be realized but it requires an important work in the data integration phase. Besides, there is always the risk of redundancy due to the fact that each table is created independently. Authors in [4] present a dimensional fact model. This latter relays on the construction of data marts firstly. This can insure the success in case of complex projects but it neglects the role of decision makers. Authors in [5] adopt the bottom up approach. In fact, they present a dimensional model development method from traditional Entity-Relationship models to insure the modeling of Data Warehouses and Data Marts. This method is based on three steps: the first step includes the classification of data models' entities into a set of categories. This leads to the production of a dimensional model from an Entity-Relationship model. We find the transactional entities that insure the storage of details concerning particular events in the company. We find also the component entities that are directly connected to transactional entities through the 1..* relationship. Those entities allow defining details of each transaction. Classification entities are connected to component entities through the 1..* relationship. Classification entities present the existing hierarchies in the model. The second step consists on identifying hierarchies that exist in the model. In fact, the hierarchy is an important concept in the dimensional modeling level. The third step consists on grouping hierarchies and aggregations together to form a dimensional model. At this level, we find two operators that are used to product models of dimensions. In fact, the first operator can transform the high level entities to low level entities. This can be done until the arrival at the bottom of the architecture. The aim is to have an only table at the end. For the second operator, it is applied on transactional data to create a new entity which contains summarized data. This approach is used as a base for the data

driven and user driven methodologies. In fact, as presented in [3], the data driven (supply driven) methodology starts by analyzing operational data sources to identify existent data. Users' intervention is limited to the choice of necessary data for the decision making process. This methodology is adopted when data sources are valid. For the user driven methodology, it starts by collecting users' needs. Those needs will be integrated in order to obtain one multidimensional schema. This approach is appreciated by users but it presents a big challenge. In fact, managers of projects must be able to take into account the different points of views. For the middle out approach, it is a hybrid method since it benefits from the two approaches cited above. Authors in [6] present an example of hybrid modeling method that is based on the top down and the bottom up approaches. The bottom up approach is based on three steps: the collection of needs, the specification and the formalization of those needs in the form of multidimensional constellation schema. The top down approach includes the data collection and the construction of a multidimensional schema that allows decision making. The approach is based on the description of decision makers needs. Those two approaches allow having two schemas, then from those latter only one schema will be derived and kept. The middle out approach is composed of four phases; the users' needs analysis, the confrontation/comparison, the resolution of conflicts and the implementation. Authors in [7] present another method which uses the middle out approach. This latter is based on three steps: the collection of users' requirements by the top down approach, the recovery of star schema by the bottom up approach and finally the integration phase. This latter connects the obtained star schema from the first step to the obtained star schema from the second step. The integration is realized thanks to a set of matrix. Users' requirements are collected by the Goal Question Metric (GQM) paradigm. This latter allow attributing metrics to identified goals. This facilitates the filtering and the deletion of not useful goals. Authors of [7] consider that the modeling of warehouses is a process based on goals, and then users' goals related to DW development will be present explicitly. Goals will be analyzed in order to reduce their number (authors take into account the similarity of goals). For the choice of star schema, authors use the Entity-Relationship model. This latter is exhaustively analyzed to find entities that will be transformed to facts and dimensions. The transformation process of Entity-Relationship model to a star schema is based on three steps. The first step is the construction of a connected graph that serves to synthesized data. The second step is to extract a snowflake schema from the graph. The third step is the integration phase. In fact, authors exploit the structure of the warehouse of the first phase and the set of possible schemas of the second phase, and then they apply a set of steps such as converting of schema to express them with the same terminology.

The clustering of moving objects' trajectories is bound in trajectory representation, trajectory similarity and the clustering algorithms. Those latter can be classified into four categories that are partitioning method (for example K-means [8]), hierarchical method (for example Birsh [9]), methods based on density (for example DBSCAN [10]) and methods based on grid (for example Sting [11]). The following table summarizes research works related to clustering.

Table 1. Classification of clustering algorithms.

Clustering category	principle	Algorithm example
Partitioning method	It consists on constructing several partitions then evaluate them according to some criteria	K-means [8] PAM (Partition Around Medoids)
Hierarchical method	It consists on creating a hierarchical decomposition of objects according to some criteria	BIRCH (Balanced I terative Reducing and Clustering using Hierarchy [9])
Method based on density	It is based on connectivity and density	DBSCAN [10] Optics [12]
Method based on grid	It is based on multilevel granularity structure	Sting [11]

For moving objects' trajectory clustering, many research works has used algorithms described above and adapted them to their activity domain. In fact, Gaffeny [13], [14] saw that the representation of trajectory based on vectors is inadequate in many cases and for this reason they proposed a clustering algorithm that is based on model. In this algorithm, a set of trajectories are represented by a probabilistic mixture regressions model and they proof that the EM algorithm can be used in case of trajectories' clustering. This approach groups the whole trajectories and non trajectory sections in a cluster.

Lee [15] proposed an algorithm called TRACLUS, this latter group similar trajectory sections in a cluster. They proved that the method of clustering based on density is the best in case of complex moving objects trajectories clustering.

There are several research works on moving objects' trajectories that are based on similarity criteria. Among those research works, we quote the works of Vlachas [16], Lin [17], Zeinalipour [18] and Pelekis [19].

Vlachas [16], proposed the use of non metric distance function that is based on the longest common subsequence (LCSS) with the conjunction of the SIGMOD Match function to match sequences of two given trajectories.

In [17], authors didn't take into account the spatial aspect in the computing of similarity and they proposed a new distance called One Way Distance (OWD). Zeinalipou and his team [18] introduced the notion of distributes spatio-temporal similarity based on the measure of LCSS Distance. They proposed two new algorithms with good performances. Pelekis [19] proposed a framework that consists on a set of distance operators based on the space, the temporal and other trajectories derived parameters such as the speed and the direction.

3 TRAJECTORY DATA WAREHOUSE CONCEPTUAL MODELING

There are three methods of conceptual modeling of DWs; the first one is the top down approach [20] that is based on the needs of the users, the second is the bottom-up approach [5] that begins with the operational data sources and finally the mixed approach [21] that combines the two previous approaches. We used the top down approach in our modeling phase because we were interested in users needs. In term of MDA (Model Driven Architecture) [22] our solution is situated in the CIM (Computation Independent Model) level because the models are not inevitably transformed into code. For the abstraction levels [23] (conceptual, logical and physical) our solution is established to cover the conceptual level.

As a multidimensional model, we choose the star schema since it is very popular because of its representation which is easy to understand. The main features of the star schema are the fact table and dimension tables. The fact table is the table at the center. It represents the variables to be analyzed. This latter contains records that are ready to explore, usually with ad hoc queries. Its primary key is a composite of all the columns except numeric values. The dimension tables represented axis of analysis of the variable to be analyzed. Each dimension table contains a primary key and other specific attributes. In our working example the fact table is the trajectory and the dimension tables are the mobile information collector, stop, move, pda, date, country, mean of transport, gps-data and point of interest. Those latter are described in the following table:

Table2. Trajectory data warehouse model description.

Table name	Description
Trajectory fact	Each MIC has a trajectory that is fixed by the responsible of the mission. Semantically, a trajectory is defined as an ordered set of stops and moves. Each trajectory has an identifier and a set of attributes such as t-begin-trajectory, t-end-trajectory and duration-trajectory.
Stop dimension	A stop is an important part of a trajectory. It considers that the object has not effectively moved. Each stop has a life cycle which is a simple time interval and a geometry which is a point and a kind. In fact, we considered that we have three types of stop: a planned stop in order to observe, to collect and to send trajectory data, a private stop in order to have a break and an unforeseen stop when there are some navigation events.
Move dimension	Each MIC moves through a given set of

	trajectories using a given mean of transport. Each move has a geometry which is a time varying point and a life cycle which is a simple time interval. This latter is represented by the "duration" attribute.
Mobile information collector dimension	MICs are involved in collecting information about various commercial and investment domains. A MIC is identified by an identifier and some attributes such as first-name, last-name…
Pda dimension	Each MIC has one PDA to send trajectory data and to communicate with the responsible of his mission. The PDA has an identifier and it is connected to a GPS.
Gps-data dimension	The GPS-data class can be connected to a given PDA and has an identifier and a set of attributes like the latitude and the longitude.
Point-of-interest dimension	There are several points of interest around each stop. Points of interest can be of type natural or artificial and have identifiers and other specific attributes such as name, type, location…
Mean-of-transport	While moving, the MIC uses a mean of transport. This latter has an identifier and some attributes like the minimum speed, the maximum speed…

The following figure represents the multidimensional modeling (the star schema) of our working example:
The classification is a used method to group similar things according to some criteria into groups or classes.

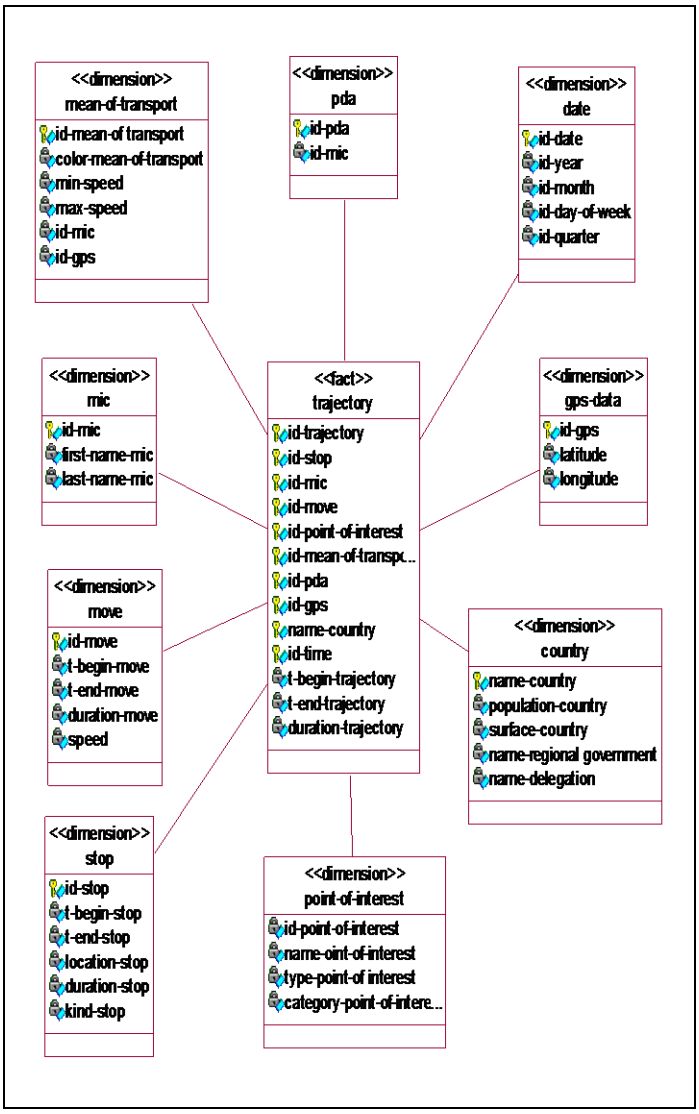

Figure1. The trajectory data warehouse star schema.

4 TRAJECTORY CLASSIFICATIONS

Our type of clustering allows to determinate trajectories that are characterized by the existence of natural or artificial projects around its stops.
The principle of our clustering is to check the type of projects existing in a given trajectory. If the type is natural then we put the trajectory in the natural-trajectory cluster, else, that is means that the type of projects is artificial, we put the considered trajectory in the artificial-trajectory cluster.
This clustering will help decision makers to have a good idea on the possibility to launch new investment in the adequate trajectory and will allow to analysts to reduce the time of response to query since he has to interrogate the natural

trajectory fact table or the artificial trajectory fact table and not the whole trajectory fact.
The following algorithm allows classifying each trajectory in the natural-trajectory class or the artificial-trajectory class. This depends on the type of points of interest existing in a given trajectory:

Algorithm of trajectories classification

Input
Pi: point of interest i
Trajectory i
Output
Artificial-trajectory class
Natural-trajectory class
FOR each Pi IN Trajectory i DO
Begin
IF (Proba(Pi.Type=Natural,Trajectory) > Proba(Pi.Type=Artificial,Trajectory)) THEN
PUT (Pi,Natural_trajectory)
Else if (Proba(Pi.Type=Natural,Trajectory) < Proba(Pi.Type=Artificial,Trajectory)) **THEN**
PUT (Pi,Artificial_trajectory class)
Else PUT (Pi,**Natural**_trajectory class)
PUT (Pi,Artificial_trajectory class)
End IF
End

The application of this algorithm will have an impact on the trajectory data warehouse conceptual model that is described in the previous section. In fact, the star schema will be transformed to a constellation schema since the fact table trajectory will be divided into the fact table natural-trajectory and the fact table artificial-trajectory as presented in the following schema:

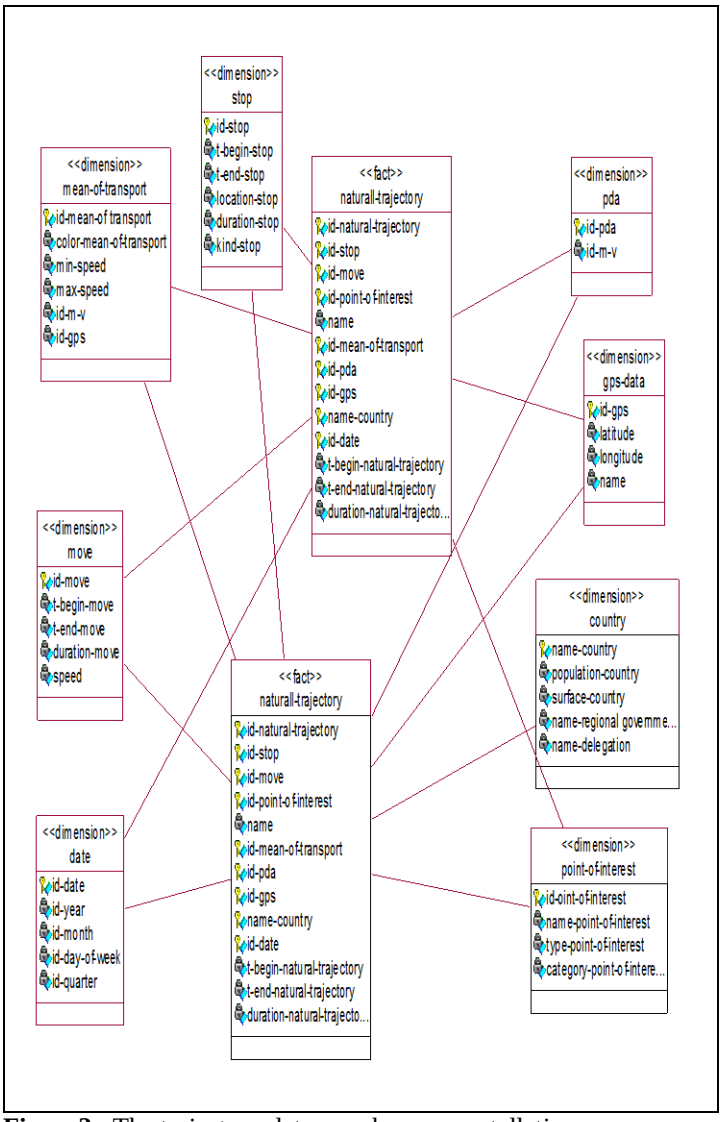

Figure2. The trajectory data warehouse constellation schema after classification.

5 QUERYING THE OLD TDW MODELVERSUS THE NEW TDW MODEL (after classification)

In this section, we will define some examples of queries about different existing points of interest and their localization in order to help investors in the choice of the suitable investment. Those queries will be applied on the TDW model before classification and after classification.

- Which trajectories that contain touristic projects?
  - (From old model)

*Select id-trajectory From trajectory T, point-of-interest P Where T.id-point-of-interest=P.id-point-of-interest and exists (select * from trajectory T' where (T'.id-trajectory=T.id-trajectory) and T'.id-point-of-interest In (select id-point-of-interest From point-of-interest Where id-tour-company Is Not Null and type.point-of-interest= touristic)))*

- (From model after classification)

Select id-artificial-trajectory From artificial-trajectory AT, point-of-interest P Where AT.id-point-of-interest=P.id-point-of-interest and exists (select * from trajectory AT' where (AT'.id-trajectory=AT.id-trajectory) and AT'.id-point-of-interest In (select id-point-of-interest From point-of-interest Where type.point-of-interest= touristic)))

- *How many agriculture projects are at sousse?*

- (From old model)

Select count (select id-point-of-interest) From point-of-interest P, trajectory T, country C, delegation D Where T.id-country=C.id-country and C.name-country="tunisia" and T.id-point-of-interest=P.id-point-of-interest and P.type=agriculture and C.id-delegation=D.id-delegation and C.name-delegation="sousse".

- (From model after classification)

Select count (select id-point-of-interest) From point-of-interest P, natural-trajectory NT, country C, delegation D Where T.id-country=C.id-country and C.name-country="tunisia" and NT.id-point-of-interest=P.id-point-of-interest and P.type=agriculture and C.id-delegation=D.id-delegation and C.name-delegation="sousse".

Those two queries need an entire route of the fact table trajectory, whereas in case of trajectory classification the analysts can save time by asking the natural-trajectory class or the artificial-trajectory class. In fact, the profits of trajectory classification are to decrease information and then to associate each trajectory to a specified class once typologies of trajectory classes have been defined. For example, when an investor is oriented or interested in natural projects, the decision maker has to check information in the natural-trajectory class instead of checking the whole trajectory fact table.

6 CONCLUSION and FUTURE WORKS

In this work, we presented concepts related to trajectory data resulting from a specific mobile object which is the mobile information collector. Such trajectory data resulting from MIC were gathered and then stored in a trajectory data warehouse. Contents of this latter were analyzed to make strategic decision about implanting new commercial activities and finding new opportunities.
A classification algorithm based on types of points of interest existing in given trajectories was presented in order to classify or to group trajectories that contains points of interest of type natural (such as sea, lake, mountain, desert)together and of type artificial (such as touristic projects, healthcare projects, industrial projects, educational projects) together. The mapping of each trajectory into the natural trajectory class or the artificial trajectory class will enable decision makers to obtain specific knowledge about points of interest existing in a specific trajectory with a least investment of resources. As future work, we propose to use data mining tools to analyze more deeply the trajectory data in order to perform the decision making

7 REFERENCES

1. E. wayane,. Four ways to build a data warehouse. 2003.
2. T. Karayannidis, N. Sellis, T. MAC: Conceptual data modeling for OLAP, Proc of the International Workshop on DMDW. 2001.
3. M, Golfarelli,. From user requirements to conceptual design in data warehouse design. Information science reference, 2009.
4. M. Rizzi, S,. WAND: A case tool for workload based design of a data mart. In Proc SEBD, Portoferraio, Italy, pp. 422-426, 2002.

5. D. Moody, M. Kortink,. From enterprise models to dimensional models: A methodology for data warehouse and data mart design. In Proc.2nd DMDW,Stockholm, Sweden, 2000.
6. J. Ghozzi, O. Teste,. Méthode de conception d'une base multidimensionnelle contrainte, 1iere journée francophone sur les entropots de données et l'analyse en ligne, toulouse, pp. 51-70, 2005.
7. Bonifati, A. Cattaneo, F. Designing data marts for data warehouses. ACM transactions on software engineering and methodology, pp. 452-483, 2001.
8. S.Lioyd, Least squares quantization in PCM, IEEE transactions on information theory, pp.129-137, 1982.
9. T.Zhang, R.Ramakrishnan. An efficient data clustering method for very large databases. In proc. ACM SIGMOD, Canada, pp.103-114, 1996.
10. E. Fayyad. From data mining to knowledge discovery in databases. AI magazine, pp 37-54,1996.
11. W.Wang, R.Muntz. A statical information grid approach to spatial data mining. In proc. Conference on very large databases, Greece, pp.186-195, 1997.
12. M. Ankerst, M.Breunig. OPTICS: Ordering Points to Identify the Clustering Structure. In proc. ACM SIGMOD, USA, pp.49-60, 1999.
13. S. Gaffney, P. Smith. Trajectory clustering with mixture of regression models. In proc. ACM SGKDD, Clifornia, pp.63-72, 1999
14. S. Gaffney, A. Robertson. Probabilistic clustering of extratropical cyclones using regression mixture models. Technical report UCI-ICS, university of California, 2006.
15. G. Lee, Y. Whang. Trajectory clustering: a partition and group framework. In proc. SIGMOD, china, 2007.
16. M. Vlachos, D. Gunopulos. Rotation invariant distance measures for trajectories. In proc. ACM SIGKDD, pp.707-712, 2004.
17. B. Lin, J. Su. Shapes based trajectory queries for moving objects. GIS, pp.21-30. 2005
18. Y. Zeinalipour, S. Song Lin, D. Gunopulos. Distributed spatio-temporal similarity search. CIKM,pp.14-23. 2006.
19. N. Pelekis, I. Kopanakis, G. Ntousti, I. Andrienko. Similarity search in trajectory databases. International symposium on temporal representation and reasoning, pp.124-140. 2007.
20. N. Tryfona, F. Busborg et J.G. Christiansen. "starER: A Conceptual Model for Data Warehouse Design". Dans: Proceedings of the ACM 2nd International Workshop on Data warehousing and OLAP (DOLAP'99), Kansas City, Missouri, USA, pp. 3–8. 1999.
21. C. Sapia, M. Blaschka, G. Höfling, B. Dinter. "Extending the E/R Model for the multidimensional paradigm". International Workshop on Data Warehousing and Data Mining (DWDM '98) in conjunction with the 17th Int. Conf. on Conceptual Modeling (ER '98), (Singapore). 1998.
22. J. Norberto, M. Juan Trujillo. "An MDA approach for the development of data warehouses ". Decision Support Systems, pp. 41–58. 2008.
23. N. Prat, J. Akoka and I. Comyn-Wattiaun. "A UML- based data warehouse design method". Decision Support Systems, pp. 1449–1473. 2006.