BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Tomul LIX (LXIII), Fasc. 3-4, 2013
Secţia
HIDROTEHNICĂ

KNOWLEDGE DISCOVERY BASED ON SPATIAL DATABASES

BY

MIRELA DANUBIANU*

"Ştefan cel Mare" University, Suceava
Faculty of Electrical Engineering

Received: October 11, 2013
Accepted for publication: November 22, 2013

Abstract. In the last years we faced a storing process of huge quantities of data, in databases, data warehouses, geographic information systems or other forms. It is obvious that, currently, these data quantities increase rapidly.

In the given context, the data mining (DM) and knowledge discovery in databases (KDD) became research fields of high interest. In essence, *these imply the discovery of interesting knowledge, unknown previously and implied from large databases* [FPSU96], and are located at the intersection between several research fields, including automatic learning, database systems, statistics, recognition of shapes and the theory of information.

Keywords: spatial; databases; information system.

1. Introduction

Many organizations have built data warehouses able to store hundreds of data terabytes concerning the exploitation of natural resources; the databases in the astronomy area have the size of terabytes; it is estimated that the Earth monitoring systems can send up to 50 gigabytes per hour. These issues represented a challenge for the traditional data analysis methods able to extract information and knowledge.

*Corresponding author: *e-mail*: mdanub@eed.usv.ro

In addition to the various studies on knowledge discovery in relational databases, or in transaction-related databases, the spatial data mining, which refers to the *extraction of implied knowledge, of spatial relations or of other patterns which are not expressly memorized in the spatial databases,* was contemplated by several research studies (Ester *et al*., 1995; Knorr & Ng, 1996; Koperski & Han, 1995; Yhang *et al*., 1996).

The spatial data has a series of distinctive characteristics from the relational databases. We mainly refer to the fact that these contain topological or distance information, use multidimensional spatial indexation structures, are accessed through specific access methods and often require geometrical computation and knowledge representation spatial techniques. Consequently, the spatial data mining requires the integration of data mining technologies with those specific for spatial databases. One of the sensitive issues of spatial data mining consists of the efficiency of used techniques, due to the work with important volumes of complex data and with spatial access methods.

The spatial data mining can be used in order to display the spatial databases, to understand this type of data, to discover the spatial links and the relation between spatial data and the non-spatial data, to reorganize the spatial databases, optimize the spatial interrogations etc. The results of such research works find their application in geographical information systems, in mining the databases of image type, in navigation and in any other fields operating with spatial data.

This paper intends to be an overview of spatial data mining, used techniques, with the pros and cons thereof, enforcement manners and related challenges.

2. Use of Relational Database Exploration Techniques

One of the first attempts of spatial data analysis was conducted in 1993 by Major and Mangano, who used a commercial instrument in order to mine a database related to a tropical storm, in order to forecast whether this would reach the US. The data describing the hurricane was decomposed into observations at various points, and such observations were memorized in a relational database. Attributes of the following types were used: position of the hurricane, speed, circulation direction, gradient from the coast etc. Because the description of the hurricane at various points was memorized of several tuples, certain data was interdependent. Such interdependencies generates a number of shortcomings because the used algorithm implied the independence of data. A GIS system was used as support of the selection for the best rules.

The conclusions of this study were that it is required to extend the data mining traditional techniques towards the mining of spatial data for an appropriate analysis of complex spatial objects and phenomena.

3. Process of Knowledge Discovery Based on Spatial Databases

The experience of the last years showed that the knowledge discovery in huge databases implies more than the mere application of sophisticated data

mining algorithms to a preset set of data.

One of the major problems of the research related to knowledge discovery in data refers to the KDD understanding as a "***non-trivial process of identifying the valid, novel, potentially useful and ultimately intelligible patterns from data.***" (Fayyad *et al*., 1996).

From this point of view, the concept of pattern is understood in a very general manner. A pattern is what a data mining algorithm can extract or generate from data, such as a model which classifies a group of data or a multitude of association rules based on a decision tree or a neuronal network.

The knowledge discovery from spatial data can be modeled through the architecture (Matheus *et al*. , 1993), set out in Fig. 1.

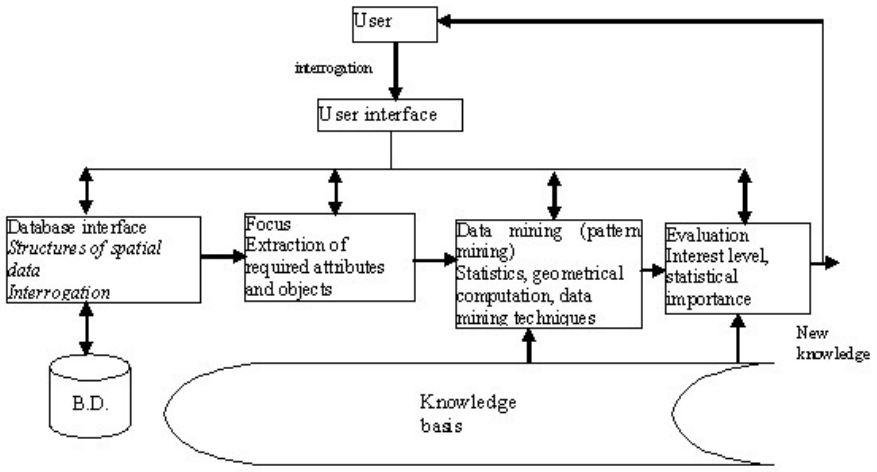

Fig. 1 – Model of knowledge discovery from spatial data.

The experience up to date lead to the conclusion that the process of knowledge discovery is not reduced to a "simple touch of a button", but on the contrary, is complex, iterative and highly interactive. Human presence is felt on each stage thereof through the user (or analyst) who decides the contents of the next stage, if the current stage is to be resumed or if one should revert to a previous stage.

Briefly, the knowledge discovery takes place as follows: the background knowledge, such as the hierarchy of spatial and non-spatial concepts or information related to the database is memorized in the knowledge base. The data is extracted from the storage media by means of the database interface which enables, inter alia, to optimize the interrogations. Indexes can be used for spatial data, for an efficient processing thereof. *The focusing component* decides which data parts are useful for the recognition of patterns. For instance, one can decide that only certain attributes are relevant for the knowledge discovery task or may extract objects the use of which guarantees

the best results. The rules and patterns are discovered during the data mining techniques stage performed by the module bearing the same name. This can use statistical procedures, automatic learning methods and classification and grouping techniques in conjunction with geometrical computation algorithms, in order to discover rules or links. The importance and significance of the discovered patterns is determined by means of the evaluation module which can remove the already existing or redundant knowledge. The discovered knowledge is ultimately provided to the user for review purposes.

4. Exploration methods and techniques of the used data

The knowledge discovered in spatial databases may have different forms, namely: rules characteristic for the description of spatial data, discriminative rules for differentiating a class of spatial data from other classes, association rules which associate one or more characteristics to a distinct set of characteristics, deviation and evolution rules for describing the time changes or rules which describe special structures. All these can be set out in various forms and may be used in order to describe the spatial objects.

Because data mining requires many challenging issues for research, the direct application of methods and techniques developed by related fields, such as automatic learning, statistics or databases cannot solve the problems. This requires the conduct of dedicated studies in order to identify new integrated methods or techniques for an effective mining. As a result, the data mining became a distinct research field.

As component of the knowledge discovery from data, data mining implies the iterative application of certain specific methods, in particular of specific algorithms.

Two types of targets can be identified for this stage, namely: *verification*, in which case the system is used to verify the presumptions of the user, and the *discovery,* when the system autonomously finds new patterns. Furthermore, discovery can be divided into: *forecast* if the system finds patterns in order to forecast the future behavior for certain entities, or *description* when the system finds patterns in order to provide these to a user, in an intelligible form.

The objectives of forecast and description, respectively, are achieved through the following primary data mining methods:

a) **classification**: finding a function which includes a data item in one of several preset classes.

b) **regression**: is used in order to forecast the value of a continuous variable based on the values of other variables, implying a linear or non-linear dependency model. Logical regression is used in order to forecast the value of a binary variable. It is a classification tool used in order to determine the value of a variable, such as, for instance, “is an individual is a buyer or not”, and is also

used in order to determine continuous variables, such as: "the likelihood of an individual to go shopping".

c) **clustering**: identifies a finite set of categories or clusters in order to describe data. Closely related is the method of estimating the probability density, which consists of estimation techniques from data of the probability function associated for all variables/fields of a database.

d) **summary:** identifies a compact description for a subset of data.

e) **dependencies modeling:** identifies a model which describes the significant dependencies between variables.

f) **identification of changes and deviation:** discovers the most significant data changes in the time frame between two consecutive measurements.

Based on the data mining fundament, the data mining can be classified into mining based on generalization, mining based on patterns, mining based on mathematical and statistical theories, integrated access etc.

5. Exploration Based on Generalization

The data and items in the databases often contain detailed information at the level of received concepts. In this case it would be advisable to make a summary of the data sets in order to present them through higher level concepts. For instance, can be summarized the detailed data o temperatures and precipitations of a region in order to provide the general climate pattern. This implies a data mining based on generalization, in which case the first stage performs a generalization of a significant volume of relevant data from a low level concept to a relatively high one, and afterwards extracts the knowledge from the generalized data. This implies the existence of fundamental knowledge in the form of a hierarchy of concepts, expressly provided by the experts in the field, which can be automatically generated through data analysis.

In case of spatial databases, two categories of concept hierarchies can be defined: *non-spatial and spatial*. Once the concept hierarchies is imposed, the information becomes more and more general, but still consistent at lower concept level. In case of spatial data, an example of hierarchy in the generalization process could be the following: the regions representing regional-administrative units can be united into counties, which can be united into stated etc.

The attribute-oriented induction is an efficient data generalization technique (Fayyad *et al*., 1996). First of all is considered a data mining interrogation expressed in a SQL similar language, (for instance DMSQL) which collects the relevant set of data in a database. Afterwards the generalization takes place by increasing the generalization hierarchy and by summarizing the connections between the spatial and non-spatial data at higher level concepts. In case of non-spatial data, this takes place by: increasing the concept hierarchy when the values of attributes in a tuple are replaced with generalized values, by eliminating the attributes when it is impossible to

continue the generalization and there are too many distinct values for an attribute and by coupling the identical tuples. The induction continues until all attributes are generalized to the intended level. During the process of uniting the identical tuples the number of coupled tuples is memorized as a meter. In addition, the aggregated values of certain quantitative attributes may also be memorized in order to enable the quantitative presentation of the acquired knowledge.

Generalized data can be expressed in the form of generalized formulas or data cubes which can be subject to other operations in order to be turned into various forms of knowledge. For instance, the "drill-down" and "roll-up" operations enable the display of data at various generalization levels; the generalized relations can be transposed into summary tables, maps or charts for viewing and presentation purposes, characteristic rules or discriminating rules may be extracted etc.

The attribute-oriented induction was also extended on spatial data. Lu *et al*. (1993), sets out two generalization algorithms: in case of spatial data dominance and in case of non-spatial data dominance.

The generalization algorithm in case of spatial data dominance enables the description of space regions by means of high level attributes. First of all the regions are united in compliance with the special hierarchies and this generates a map with a low number of areas. Afterwards, a non-spatial description of each area is performed, by means of the attribute-oriented induction technique. The reply of an interrogation is the description of all regions through the disjunction of few attributes which characterize each of the generalized regions.

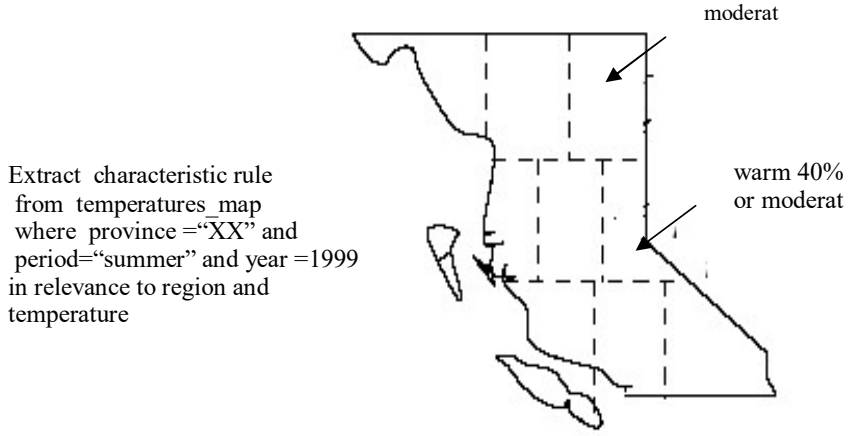

Fig. 2 – Interrogation and obtained result by applying the generalizing algorithm with spatial dominance.

The generalization algorithm for non-spatial dominant data generate maps which consist of a low number of regions with the same hierarchical level of non-spatial description. This algorithm begins with the non-spatial attribute-

oriented induction and the generalization thereof to higher conceptual levels. After that, the neighboring areas with the same values for the generalized attributes are united. Approximation can be used in order to ignore the small regions with different non-spatial descriptions. The next figure shows an example of the mentioned applied algorithms. It shows the interrogation in the generalization algorithm with spatial dominance, as well as the obtained result.

5.1. Mining of Spatial Associations

The above mentioned method enables the identification of characteristic rules which describe spatial objects based on the non-spatial attributes thereof. However, under certain circumstances, it is intended to discover the spatial association rules, which are those rules associating one or more spatial objects with other spatial objects.

The concept of association rules was introduced by Agrawal *et al*. (1993) in order to mine large transaction databases, and by Koperski and Han (1995) was extended to spatial databases.

A spatial association rule is noted:

$$X \rightarrow Y\ (c\%),$$

where: X and Y are sets of spatial and non-spatial attribute sets and c% is the confidence degree of the rule.

An example of such rule is as follows:

is (x, school) $\wedge$ *close_to (x, sport_center)* $\rightarrow$ *close_to (x, park) (80%)*

This rule mentions that 80% of the schools located close to a sport center are also close to a park. There are different types of spatial attributes which can generate spatial association rules, among which: topological relations, such as *intersection, overlaying, disjunction* etc., or distance information, such as *close_to, far_from* etc.

The large databases may contain many associations between objects, but most of them either apply to an extremely low number of objects, either have a low confidence degree.

For instance, a user may not be interested in a relation which associates 5% of the houses with a certain school, but may be interested in a rule applied to at least 50% of the houses. There are two thresholds, the *minimum support and the minimum confidence degree* which are used in order to control the filtering of associations describing a small percent of objects or low confidence rules. These thresholds may be different at each level of non-spatial descriptions of objects, because the use of the same threshold might not find the interesting associations at low hierarchic level because the number of objects which meet the same attribute may be low.

The mining process is initiated through an interrogation which describes a class of objects S using other classes of objects relevant for the proposed task and a multitude of relevant connections. For instance, a user

might want to describe the parks of a town by representing the connections between the parks and other objects, such as: railways, restaurants, zoos, streets etc. Instead, the user may determine that is interested only in the objects of an area with one kilometer width from the park edges. In order to mitigate the cost of spatial computations, the algorithm uses various approximations and applies the most detailed spatial computation, which is more expensive, only to those patterns with an appropriate support at the considered approximation level.

5.2. Data Clustering

Spatial clustering, which associates similar spatial objects in classes, is an important data mining technique, which can be used in order to identify the similar resource mining areas, in merging the regions with similar climate characteristics etc.

Can be used as a freestanding instrument in order to better understand data distribution, to notice the characteristics of each cluster and to easily focus on a certain set of clusters for further analyses, or may be a preparation step for the classification or characterizing algorithms which operate on the identified clusters.

Due to the fact that the cluster analysis was an active data mining research field, efficient clustering methods were developed in the last years. These can be classified into: partitioning methods (Kaufman & Rousseeuw, 1990; Ng & Han, 1994; Bradley *et al*., 1998), hierarchic methods (Kaufman & Rousseeuw, 1990; Yhang *et al*., 1996; Guha *et al*., 1998; Karypis *et al*., 1999), density-based methods (Easter *et al*., 1996; Ankerst *et al*., 1999; Hinneburg & Keim, 1998), grid-based methods (Wang *et al*., 1997; Sheikholeslami, *et al*., 1998; Agrawal *et al*., 1998) and model-based methods (Shavlik & Dietterich, 1990; Kohonen, 1982).

5.3. Use of Approximates and Aggregations

The clustering methods can provide answers for questions such as "which is the location of clusters in spatial databases". However, there is also another aspect of the problem, namely "why are the clusters located there". This can be rephrased as follows: which are the characteristics of clusters, based on the features on contained objects. The problem refers to the measuring manner of the aggregation proximity because an allegation such as "90% of the cluster houses might have characteristic X" is more interesting and contains more information than the allegation "one house definitely has X characteristic". The aggregation proximity is the proximity measure of the cluster set of objects to a certain characteristic as opposed to the distance between the cluster borders and the limits of a characteristic.

Knorr and Ng (1996), sets out a method which enables the fast identification of a characteristic close to the cluster.

5.4. Mining of Raster Databases

The mining of raster databases, such as the image type ones, is a particular case of spatial data mining. The studies in this field include POSS II (Second Palomar Observatory Sky Survey) which uses the decision trees in order to classify galaxies, or other cosmic objects starting from approximately 3 terabytes of cosmic images and Magellan study, which analyzes approximately 30000 high resolution radar images of the surface of Venus planet, in order to identify the potential volcanoes.

5.5. Future Challenges

In the years of transition to an information society it is useless to prove the necessity of a research field such as the one of data mining from large volumes of geographic data.

Nonetheless, what should be mentioned is the fact that there still are research issues which require the extension of studies. Some of these issues refer to: identifying certain additional methods such as classification, mining based on patterns or similarities, specially dedicated to geographic data; analysis of interactive data mining opportunities by using visual reactions, and analysis of possibilities to design a language dedicated to spatial data mining; the mining of data originating from distributed sources (Internet/Intranet) and memorized in various formats. A real challenge consists of combining the spatial data mining methods with the advanced spatial databases (object-oriented spatial databases or spatial-temporal databases) and with the technology of expert systems in order to generate the so-called smart GIS systems.

6. Conclusions

In this paper we reviewed the methods used for data mining, understood as a stage of the knowledge discovery process from spatial databases. This is a current topic and also an attractive one which aims at providing the geographic information systems with the power to discover knowledge and adjust the new GIS applications to the future requirements.

REFERENCES

Agrawal R., Imielinski T., Swami A., *Mining Association Rules Between Sets of Items in Large Databases.* Proc. of the ACM SIGMOD Internat. Conf. on Manag. of Data, 1993, 9.

Agrawal R., Gehrke J., Gunopulos D., Raghavan P., *Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications*. Proc 1998 ACM-SIGMOD Internat. Conf. Manag. of Data (SIGMOD'98), Seattle, 1998, 94-105.

Ankerst M., Breunig M., Kriegel H.P., Sanders J., *OPTICS: Ordering Points to Identify the Clustering Structure*. Proc. 1998 ACM- SIGMOD Internat. Conf. Manag. of Data (SIG-MOD'99), Philadelphia, 1999, 49-60.

Bradley R., Fayyad U., Reina C., *Scaling Clustering Algorithm for Large Databases*. Proc. 1998 Internat. Conf. Knowledge Discovery and Data Mining (KDD'98), New York, 1998, 9-15.

Easter M., Kriegel H.P., Sander J., Xu X., *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases*. Proc. 1996 Internat. Conf. Knowledge Discovery and Data Mining (KDD'96), Portland, 1996, 226-231.

Ester M., Kriegel H.P., Xu X., *Knowledge Discovery in Large Spatial Databases: Focusing Techniques for Efficient Class Identification.* Adv. in Spatial Databases, SSD'95, Springer-Verlag , Berlin, 1995, 67-82.

Fayyad U., Piatetsky-Shapiro G., Smyth P., *The KDD Process for Extracting Useful Knowledge from Volumes of Data*. Comm. of the ACM, **39**, *11*, 27-34 (1996).

Fayyad U., Piatetsky-Shapiro G., Smith P., Uthurusamy R., *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, Menlo Park, CA, 1996.

Guha S., Rastogi R., Shim K., *Cune: An Efficient Clustering Algorithm for Large Databases*. Proc. 1998 ACM-SIGMOD Internat. Conf. Manag. of Data (SIGMOD'98), Seattle, 1998, 73-84.

Hinneburg A., Keim A., *An Efficient Approach to Clustering in Large Multimedia Databases with Noise*. Proc 1998 Internat. Conf. Knowledge Discovery and Data Mining (KDD'98), New York, 1998, 58-65.

Karypis G., Han J., Kumar V., *CHAMELEON: A Hierarhical Clustering Algorithm Using Dynamic Modeling*. COMPUTER, **32**, 68-75 (1999).

Kaufman L., Rousseeuw P.J., *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Soons, 1990.

Knorr E., Ng R.T., *Finding Aggregate Proximity Relationship and Commonalities in Spatial Data Mining*. IEEE Trans. Knowledge and Data Engng., 1996, 884-897.

Kohonen T., *Self Organized .Formation of Topologically Correct Feature Maps*. Biolog. Cybern., **43**, 59-69 (1982).

Koperski K., Han J., *Discovery of Spatial Association Rules in Geographic Information Databases*. SDD'95 Springer-Verlag, Berlin, 1995, 47-66.

Lu W., Han J., Choi B.C., *Discovery of General Knowledge in Large Spatial Databases*. Proc. of Far East Workshop on Geographic Information Systems, Singapore, 1993, 275-289.

Major J., Mangano J., *Selecting Among Rules Induced from a Hurricane Database*. Proc. Knowledge Discovery in Databases Workshop AAAI Press, 1993.

Matheus C.J., Chan P.K., Piatetsky-Shapiro G., *Systems for Knowledge Discovery in Databases*. IEEE Trans. Knowledge and Data Engng., **5**, *5*, 903-913 (1993).

Ng R., Han J., *Efficient and Effective Clustering Method for Spatial Data Mining*. Proc. Internat. Conf. Very Large Data Bases (VLDB94), Santiago Chile, 1994, 144-155.

Shavlik J.W., Dietterich T.G., *Reading in Machine Learning*. Morgan Kaufmann, 1990.

Sheikholeslami G., Chatterjee S., Yhang A., *WaveCluster:A Multi-Resolution Clustering Approach for Very Large Spatial Databases*. Proc. Internat. Conf. Very Large Data Bases (VLDB98), New-York, 1998, 428-439.

Wang W., Yang J., Munty R., *STING: A Statistical Information Grid Approach to Spatial Data Mining*. Proc. Intenat. Conf. Very Large Data Bases (VLDB'97), Athenes, 1997, 186-195.

Yhang T., Ramakrishnan R., Livny M., *BIRCH: an Efficient Data Clustering Method for Very Large Databases*. Proc. ACM-SIGMOD Internat. Conf. Manag. of Data, Montreal, 1996, 103-114.

DESCOPERIREA DE CUNOŞTINŢE DIN BAZE DE DATE SPAŢIALE

(Rezumat)

În perioada ultimilor ani am asistat la un proces de stocare a unor cantităţi uriaşe de date, în baze de date, depozite de date, sisteme informaţionale geografice sau sub alte forme. Este evident faptul că, în prezent, aceste cantităţi de date cresc cu o viteză uluitoare.

În contextul dat, explorarea datelor (DM – data mining) şi descoperirea cunoştinţelor din date (KDD – knowledge discovery in databases) s-au impus ca domenii de cercetare de mare interes. In esenţă, *acestea implică descoperirea de cunoştinţe interesante, anterior necunoscute şi implicite din baze mari de date*, şi se află la confluenţa mai multor domenii de cercetare, incluzând învăţarea automată, sisteme de baze de date, statistică, recunoaşterea formelor şi teoria informaţiei.