POLISH ASSOCIATION FOR SPATIAL INFORMATION
ANNALS OF GEOMATICS 2008 ❍ VOLUME VI ❍ NUMBER 1

ANALYTICAL TOOLS FOR BUSINESS INTELLIGENCE IN SPATIAL DATABASES

ANALITYCZNE NARZĘDZIA BUSINESS INTELLIGENCE W BAZACH DANYCH PRZESTRZENNYCH

Jędrzej Gąsiorowski

Institute of Geodesy and Cartography, Warsaw, Poland

Keywords: spatial databases, business intelligence, data warehouses, OLAP, data mining
Słowa kluczowe: bazy danych przestrzennych, business intelligence, hurtownie danych, OLAP, eksploracja danych

Introduction

For a dozen years, Information Technology, including new ideas and more and more powerful hardware devices, has developed very rapidly. Many technologies, which ten years ago were described as „future technologies", are very common today. In particular, the fastest development we can see in these technologies which can be used in business and commerce, is surely that caused by the huge assets assigned to science by commercial corporations. This has led to the creation of the new domain of Information Technology, *Business Intelligence* (BI), which is a part of the wider domain, *Decision Support Systems* (DSS). D.J. Power describes BI as a set of concepts and methods to improve business decision making by using fact-based support systems (Power, 2007). "Fact-based" is the keyword here. It means that BI uses existing data and relationships, existing, but not necessarily visible, to return results. In other words, BI converts data into knowledge.

On this basis, all BI tools and technologies are concerned with processing and analysing existing data. To process and analyse data efficiently, the data have to be organised and stored in databases. Where location is the key subject, a spatial database is required. Even though BI was developed to support commercial decision making, it can be very helpful for scientific research and geographic and environmental research, where spatial databases are used.

The goal of this article is to present two tools of Business Intelligence – *OLAP* and *data mining* – and to discuss possibilities of using these tools in spatial databases.

Database vs data warehouse

Unfortunately, the architecture of a regular database doesn't allow for efficient use of Business Intelligence analytical tools, because it is designed for recording data. Although it allows for simple or more advanced querying and reporting, it is optimized for as fast recording as possible. Typically, these are is known as *OLTP systems* (On-line transaction processing). For example, when a customer wants to perform some operation in a bank, the biggest priority is to record this operation in the bank's database as fast as possible, using, for example, the SQL operations `INSERT`, `DELETE` or `UPDATE`.

The main goal of Business Intelligence is to analyse and report data, which can benefit from an architecture that is designed for fast response to queries and will enable effectively performing advanced analyses and reports, such as are performed in SQL using the `SELECT` operation. Such a database is called a *data warehouse*. R. Kimball introduces several requirements for the data warehouse:

- it must make an information easy accessible and it's content must be understandable;
- it must store the information consistently and the data that it contains must be credible;
- it must be adaptive and resilient to change;
- it must protect information assets;
- it must serve for improved decision making;
- the business community must accept it if it is to be deemed successful.

From these requirements, four components must be taken into consideration while thinking about the data warehouse: operational source systems, data staging area, data presentation area and data access tools (Kimball, Ross, 2002). Figure 1 illustrates structure of the data warehouse, using these four components.

Operational source systems comprise single databases that record every operation online. They are employed in OLTP systems and can only be queried in a simple and predictive way. Operational source systems do not store any historical data. Data from operational source systems is extracted, read, understood and copied, into the next component, the data staging area. Here data is being transformed and prepared for presentation or analysis. Kimball describes this component as being for the data warehouse what a kitchen for a restaurant. Processed data is then loaded onto the presentation area. In this component, data is stored and organised for query, analysis or report generation. The presentation area is specified as a set of data marts. A data mart is a data repository, which is oriented for a specific, single subject and its task is fast performing processes connected with this subject and being easily accessible to users. Data from presentation area can be reached by using the access tools, the final component. Access tools are applications that allow query, advanced analysis and reporting of the data warehouse contents. Access tools are, for example, OLAP, data mining or just simple queries. Such a data warehouse structure enables advanced analysis and reporting of the data it contains with high speed and efficiency.

A slightly different point of view on data warehousing, differing mostly in details, was presented by W. H. Inmon (2005), who together with R. Kimball is the biggest authority in this domain, but these differences are not pertinent to this paper.

Data Warehouse has one feature, which is very important for analytical tools like OLAP and data mining, that its data structure is built using a multidimensional paradigm (Bedard et al., 2001). This simplifies navigation within the database, especially through such functions

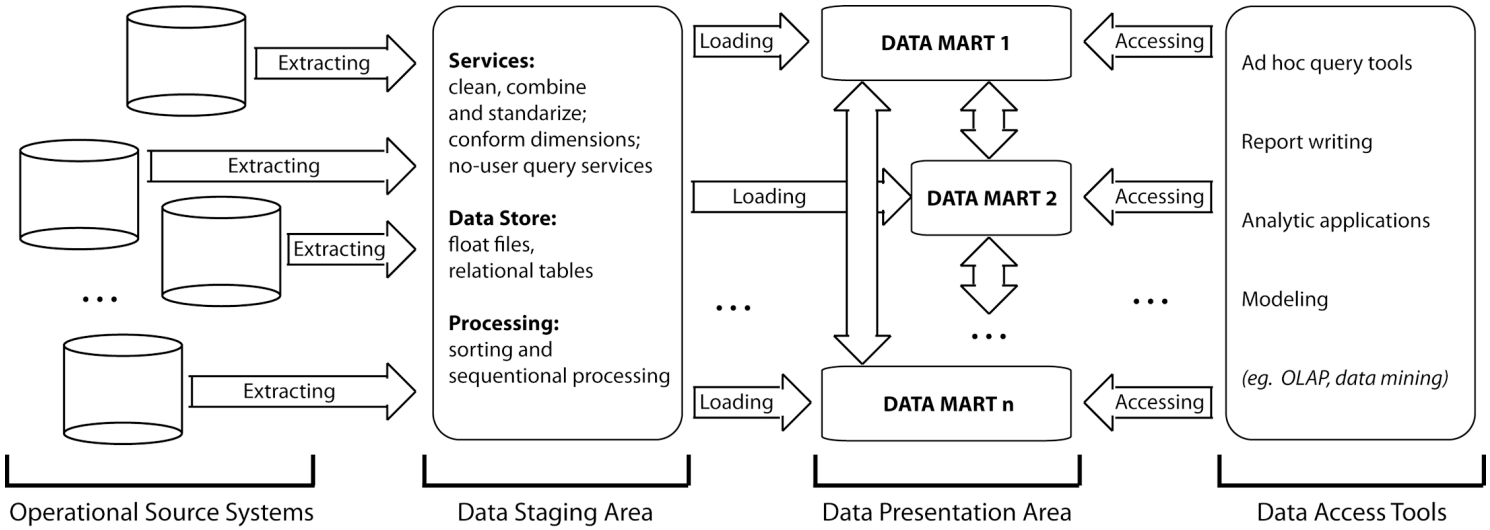

Figure 1. Data warehouse structure (Kimball, Ross, 2002)

like drill-up, drill-down and drill-across. These functions are responsible for navigation between information hierarchy levels: for example in a folder system, the drill-down function enables movement from the parent folder to the specified file in a subfolder. This multidimensional approach will be discussed later.

On-line Analytical Processing

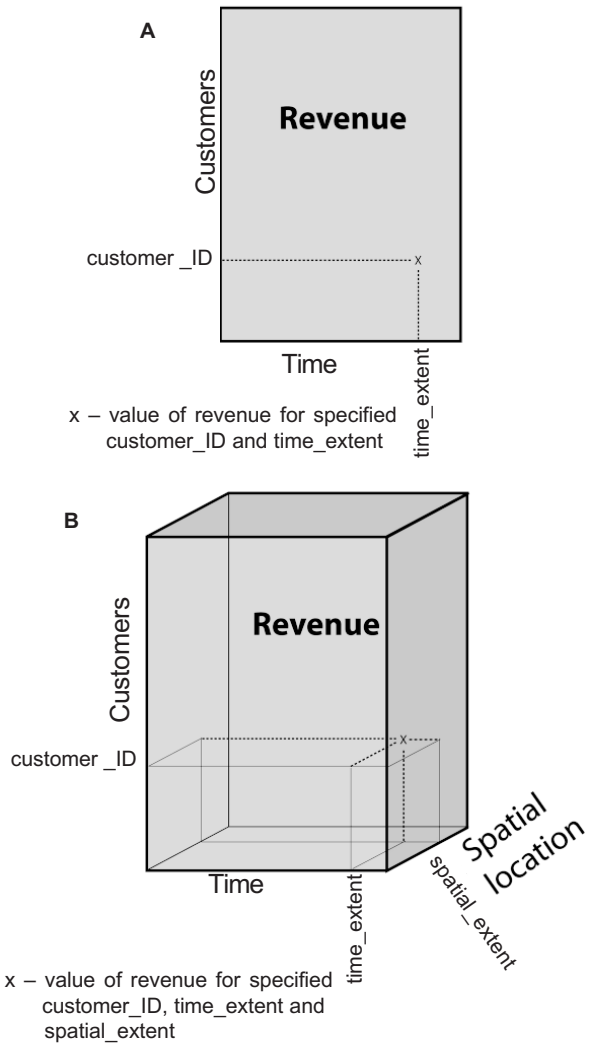

Figure 2. Illustration of ad hoc query (A) and OLAP (B) results (Wu, 2000)

One of the simplest forms of data analysis is an ad hoc query. In this case, the user will get an answer exactly to the question asked. J. Wu (2000) describes the essence of ad hoc query and OLAP in a very simple way. The revenue of some a company will be considered. Figure 2A shows the result of an ad hoc query concerning revenue depending on customers and time. Examples ofad hoc query questions that answers are: *How much revenue was registered every month of this year by each customer?* or *Which customer generated the highest/lowest revenue this year?* Many other queries can be performed here, but all of them will have common characteristics: they show a numeric answer depending on two non-numeric criteria and they deliver only that information which was requested. In a spatial database, spatial location might be queried instead of time, eg. *How much revenue was generated by each customer inside the ranges of 50 km from 10 biggest cities in Poland?* Unfortunately, in this case it is unable to set time extents.

This problem can be solved by using On-line Analytical Processing (OLAP). On the basis of the above pattern, the result of OLAP can be presented as shown in Figure 2B. The revenue of a company is shown here depending on three dimensions: customers, time and spatial location. So OLAP enables to answer the question: *How much revenue*

was registered every month of this year by each customer inside the ranges of 50 km from 10 biggest cities in Poland? This question joins OLAP and simple GIS analysis, buffering. Of course, more than three dimensions can be considered, for example the kinds of. product that company sells in addition to location, revenue and time. In such cases, the visualisation is different and mapping multiple logical dimensions onto two physical dimensions, the computer screen, is necessary (Thomsen, 2002).

This multidimensional approach in the data warehouse is based on *dimensions* and *measures* (Rivest et al., 2001). Dimensions represent descriptive, non-numeric data, like *what?*, *where?*, *why?*, *who?*, etc. Measures represent numeric data, like *how much?*, *how many?* and are attributes of *facts*, in the background of dimensions. So OLAP shows measures depending on dimensions. In Figure 2B, the fact is the revenue of a company: customers, time and spatial location are dimensions, while all values of revenue are measures. In addition, dimensions are hierarchically organised in *levels*, from the most general to the most detailed. The dimension «time» can be organised using levels such as: year › month › day. Levels are used to aggregate dimensions to limit the number of cells in the visualisation. This is shown on Figure 2 B (X symbolizes a single cell). Such a multidimensional structure for anlysis is known as the *OLAP cube*.

Put simply, OLAP shows one numeric aspect depending on many non-numeric aspects in complex views. Moreover, it can highlight some patterns and relationships, only within the data, that was requested and it depends on the user to identify them (Wu, 2000).

Data mining

Data mining is the most advanced BI tool. Generally, data mining means a process of discovering some pattern and relationships in data by using various analytical tools. M. Kantardzic (2003) introduces two primary goals for data mining: prediction and description. Prediction uses known variables or fields in the data to predict unknown or future values: it produces some model of the system described by the given dataset. Description uncovers patterns that describe the data to enable interpretation by humans: it produces some new, non-trivial information based on the given dataset. So generally, data mining means getting hidden knowledge from data. As an example, the following question will be considered: *What are the characteristics of customers in Warsaw who buy product A?* One of the results of descriptive data mining processing in this query may be such statements: *70% of customers in Warsaw who buy product A, also buy product B at the same time* or *All customers in Warsaw who buy product A pay with a credit card*. An example query of predictive data mining may be following statement: *How many of product A will be bought by customers in Warsaw each month of the next year*. These examples shows trends and predictions which often can be discovered without such a powerful tool like data mining and which an advanced user can identify just by using OLAP. But, above all, it can uncover such patterns and make predictions which are totally unknown, unexpected and surprising.

Data mining applications use a variety of analytical techniques to achieve the the goals of prediction and description. These can include neural networks, regression models, rule induction, decision trees, nearest neighbour classification, statistics and more. Each of these techniques demands more or less human involvement: while neural networks demands little

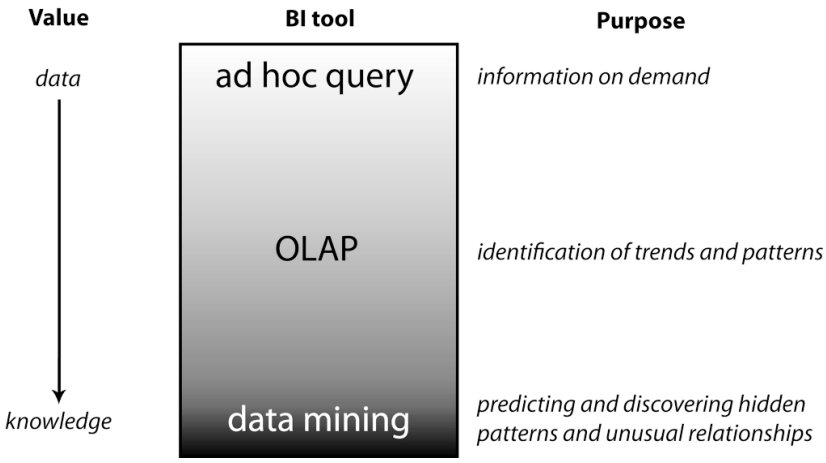

Figure 3. Summary of BI tools goals and abilities (Wu, 2000)

human manipulation, statistics needs very much. But to deploy all of them, a user has to know either how they work or what is the core of knowledge he wants to discover from the data.

As a summary of this general description of the BI tools, Figure 3 shows the main goals and abilities of ad hoc query, OLAP and data mining and the path of information from data (ad hoc query) to complex knowledge (data mining).

BI tools in spatial databases

The main goal of this article is to show how these tools can be used in databases with a spatial reference and how to join them with Geographic Information Systems (GIS) functionality.

The concept of Spatial On-line Analytical Processing (SOLAP), OLAP in spatial databases, was introduced by S. Rivest, Y. Bedard and P. Marchand (Rivest et al., 2001). Spatial data can be considered in two different ways: non-geometric and geometric. The first approach treats spatial data only as descriptive values (eg. Country: Poland, city: Warsaw, etc.). In this case, the dimension connected with spatial location in OLAP is treated as any other dimension, because its nature is the same. This approach will not be developed in this article. The most interesting and challenging is the second approach, in which the dimensions connected with spatial location include geometric shapes, spatially referenced, and allow its dimension members to be visualized and queried graphically (Rivest et al., 2001). There are two possibilities here. In the first, spatial data for geometric shapes is only queried and is not included in visualisation. The example mentioned in OLAP description is applicable here: *How much revenue was registered every month of this year by each customer inside the ranges of 50 km from 10 biggest cities in Poland?* In this case some GIS software makes a pre-analysis which extracts customers inside the buffers mentioned in the question and this result is sent with some attributes to the proper OLAP engine and the cube is generated. The final result for further user analysing can be visualised as shown on Figure 4.

In this case, two dimensions (time and location) are mapped onto one physical dimension (horizontal axis). Users may identify in this examples some patterns, which are highlighted by such a visualisation. Of course operations drill-up, drill-down, etc. can be performed here to adjust a view for user needs, for example aggregation months into quarters (function drill-up).

The second possibility is more interesting and concerns visualising spatial data in the form of geometric shapes as an OLAP result. In this case, the solution described below is proposed. A three-dimensional cube will be considered. Two of the physical dimensions will be reserved for any spatial dimensions involved in OLAP and will be considered as spatially referenced layers, like in a common GIS software. These two dimensions correspond to geographical coordinates. The remaining physical dimension will be reserved for any non-spatial dimensions, that will be mapped onto it. This is shown in Figure 5, with several example dimensions.

The biggest problem here is that the view of the three-dimensional cube with spatial shapes will not be clearly visible on a two-dimensional monitor screen. A reasonable solution would be to move every non-spatial dimension into the form of list, from which the requested values will be chosen and spatial dimensions will be displayed, depending on these chosen values. Such a solution allows users fast browsing of the requested displays and identification of trends and patterns in the data. Moreover, this approach allows the operations drill-up, drill-down and drill-across to be easily performed either for non-spatial dimensions, for example changing time level from year to month, or spatial dimensions, for example changing land cover layer from generalized to more detailed. Of course this case needs special dedicated software which will join the GIS software functionality, georeferencing, symbology creating, spatial analyses, etc., with the OLAP engine to create OLAP cubes.

Spatial data mining concepts and techniques were presented by M. Ester, H.-P. Kriegel and J. Sander (Ester et. al., 1999). Following their concept, the major difference between

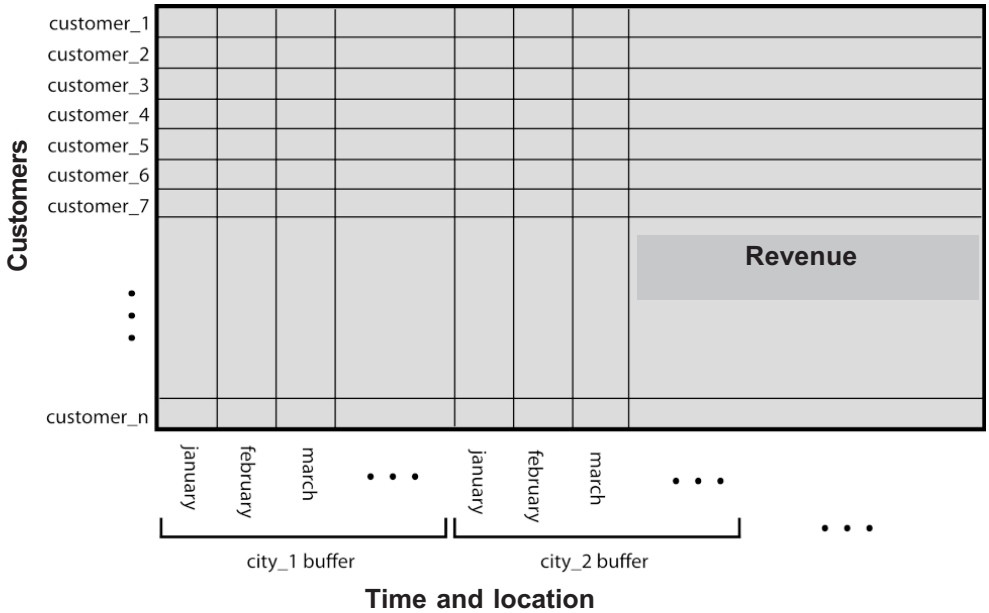

Figure 4. Spatial data OLAP result in non-spatial visualisation

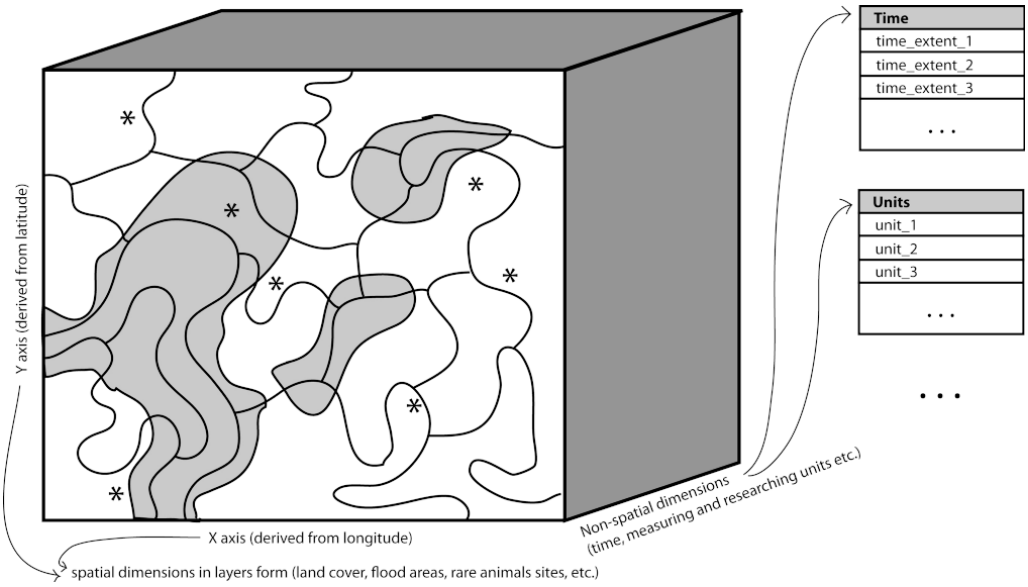

Figure 5. Spatial data OLAP result in spatial visualisation

discovering knowledge in relational databases and in spatial databases is that the attributes of the neighbours of some spatially referenced object may have an influence on that object. A spatial relationship between two objects can be considered using three kinds of relations: *topological*, *distance* and *direction*. The topological relation describes the mutual relationship between two objects which is constant while making topological transformations such as scaling, rotation, etc. Examples of such relations include: object A *contains* object B, A *overlaps* B or A *is equal* to B. The distance relation concerns the distance between two objects, while the direction relation describes the direction from object A to object B in the specified reference system. These three relations can be combined in the *complex neighbourhood relation* (Ester et al., 1999). Spatial data mining can be treated as the combination of processing and analysing attributes of spatial and non-spatial data with analysing and processing the neighbourhood relations of spatial data. To perform spatial data mining, in addition to classic data mining techniques mentioned earlier, techniques dedicated for spatial data are needed. These techniques are described by M. Ester, H.-P. Kriegel and J. Sander and include spatial clustering, spatial characterization and spatial trend detection (Ester et al., 1999).

As an example of applied spatial data mining, the following spatial data will be used: land cover, factory locations, with a pollution emission attribute, locations for lynx, and wind strength and direction for a specified area. This is shown in Figure 6.

It shows how the relations described above may participate in knowledge discovery using data mining. The following statement can be the result of analysis for this example: *Lynx sites are located inside forest areas at minimum 30 km west and minimum 100 km east from factories with high pollution emission and at minimum 15 km west and minimum 30 km east from factories with low pollution emission*. Of course data mining is dedicated for analysing really large and complex datasets and it is very difficult to show its application for uncovering hidden and unusual patterns in an example. Nevertheless, this simple example is

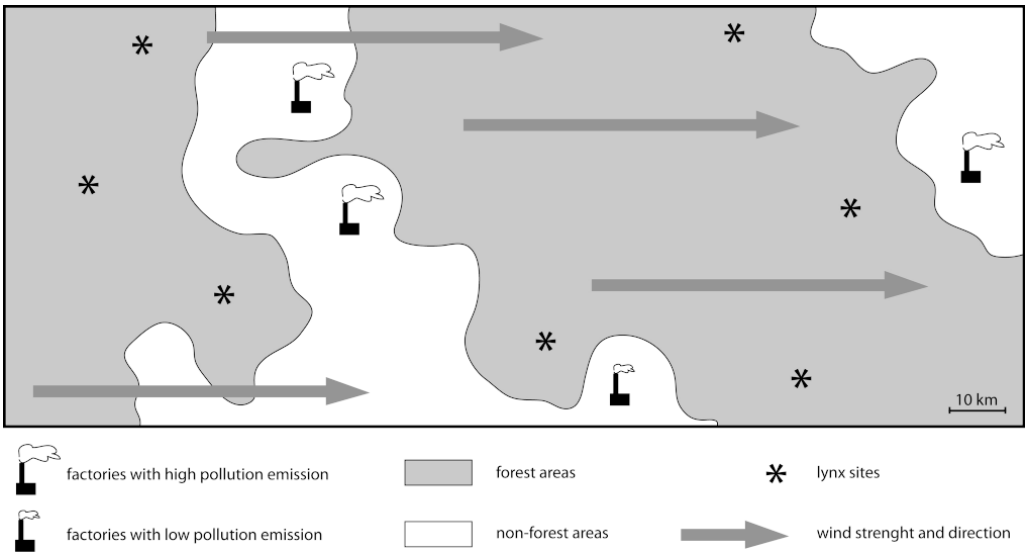

Figure 6. Presentation of spatial data mining application

provided to show how data mining can be used with spatial databases; it does not present, for example, any historical data, which is very important in data mining. But it does show the possibilities of using data mining in geographical and environmental research, as well as in commercial applications.

Conclusion

Nowadays BI tools for spatial data are at the development level. Pilot studies in universities and other research units are ongoing and in the recent past, the first attempts of creating software for spatial OLAP and spatial data mining have been undertaken, but up to now they are not commonly used by spatial data users. Nevertheless, these tools are so useful, have such powerful possibilities and, importantly, have been successfully used in the commercial field for a dozen years, that it is very probable that will become important techniques for analysing spatial data in the near future.

References

Bedard Y., Merrett T., Han J., 2001: Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery. [In:] Miller HJ, Han J. (eds.) Geographic Data Mining and Knowledge Discovery. Taylor and Francis, London, pp 53-73.

Ester M., Kriegel H-P., Sander J., 1999: Knowledge Discovery in Spatial Databases. [In:] Lecture Notes in Computer Science, vol 1701, pp 61-74.

Inmon W.H., 2005: Building the Data Warehouse. John Wiley & Sons, New York.

Kantardzic M., 2003: Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, New York.

Kimball R., Ross M., 2002: The Data Warehouse Toolkit. John Wiley & Sons, New York
Power D.J., 2007: A Brief History of Decision Support Systems. DSSResources.COM, World Wide Web, http://DSSResources.COM/history/dsshistory.html, version 4.0.
Rivest S., Bedard Y., Marchand P., 2001: Towards better support for spatial decision-making: Defining the characteristics of Spatial On-Line Analytical Processing (SOLAP). [In:] *Geomatica, the journal of the Canadian Institute of Geomatics* 15: 539-555.
Thomsen E., 2002: OLAP Solutions. Building Multidimensional Information Systems. John Wiley & Sons, New York.
Wu J., 2000: What is data mining? DM Review Online, World Wide Web, http://www.dmreview.com/news/2582-1.html

Abstract

The term Business Intelligence (BI) stands for technologies and applications that support decision making in commercial business. It is based on data analysis in a specific kind of database, termed the data warehouse. The architecture of data warehouses is optimized for searching, analysing and reporting of data. Nowadays, some spatial databases, especially in the commercial area, are so large and complex that simple analysis and reporting are not able to show all relationships and connections between data.

The article focuses on two BI tools: On-line Analytical Processing (OLAP) and data mining, and on the potential for using these tools in spatial databases. OLAP allows the creation of multidimensional views of data reports in the form of multidimensional cubes. In the spatial database, such views can be useful to show complex reports, including information about spatial location, time and other dimensions. Data mining is based on analytical searching of some regular relationship and pattern in databases, which are hidden and not visible while using simple analysis. The aim of data mining for spatial databases, can be to predict the influence of a geographic object on (a) neighbour object(s), including their attributes.

Streszczenie

Termin Business Intelligence (BI) oznacza technologie i aplikacje, które wspomagają podejmowanie decyzji w sferze biznesowej. Opiera się na analizie danych w specyficznym rodzaju baz danych, które określane są mianem hurtowni danych. Architektura hurtowni danych zoptymalizowana jest pod kątem przeszukiwania, analizy i raportowania zawartych w niej danych. Obecnie niektóre bazy danych przestrzennych, zwłaszcza w zastosowaniu komercyjnym, są tak obszerne i złożone, że prosta analiza i raportowanie nie są w stanie pokazać wszystkich związków pomiędzy danymi.

Niniejszy artykuł skupia się na dwóch narzędziach BI: przetwarzaniem analitycznym on-line (OLAP, ang. On-line Analytical Processing) i eksploracją danych (ang. data mining) oraz możliwościami zastosowania tych narzędzi w bazach danych przestrzennych. OLAP pozawala na tworzenie wielowymiarowych widoków raportowych w formie wielowymiarowych kostek. W bazie danych przestrzennych widoki takie mogą być użyteczne do pokazania złożonego raportu, zawierającego informację o położeniu przestrzennym, czasie, czy dowolnym innym wymiarze. Eksploracja danych jest techniką opartą na analitycznym wyszukiwaniu w bazach danych stałych związków i wzorów, które są ukryte i niewidoczne podczas stosowania prostych analiz. Celem eksploracji danych w przypadku baz danych przestrzennych może być przewidywanie wpływu obiektu geograficznego na obiekt(y) sąsiedni, przy uwzględnieniu jego atrybutów.

mgr inż. Jędrzej Gąsiorowski
Jedrzej.Gasiorowski@igik.edu.pl
phone: +48 22 329 19 91