Data analysis using GIS and data mining.

F.-Y. Leu, T.-H. Wang

Data Analysis Using GIS and Data Mining

Fang-Yie Leu and Tai-Shiang Wang
Dept. of Computer Science and Information Engineering, Tunghai University, Taiwan
{leufy, g932810}@thu.edu.tw

Abstract

Recently, many commercial Geographical Information Systems (GISs) have been developed. Their functions are quickly growing up. Researchers and policymakers can input environmental data to a GIS system to gain spatial analysis result which can show up how data are geographically dispersed. Besides, the data mining and data warehouse technologies can automatically mine hidden knowledge and analyze/extract knowledge from raw data, respectively. If we can put them in use with GIS, the hidden meanings or rules embedded in the environmental data can be then more deeply and precisely uncovered. In this paper, we will discuss how to use the two data analytical tools, GIS and data mining, to analyze the data collected for the Situn district so that researchers can realize some facts that can not be superficially obtained from raw data.

Keywords: GIS, data mining, data analysis.

1. Introduction

Nowadays, a huge amount of geographic information has been produced and collected, especially from satellite remote measurement and map digitalization. A part of them have been transformed from traditional formats into digital so that they can be stored in a computer system. Geographical Information Systems (GISs) are widely used in modern time, particularly in designing and showing a city's road networks, underground pipes, power lines, and et al. Users can search roads or landmarks on a electronic map or in internet if the map provides a web version, to realize the locations they are interested in.

Besides, expert systems and machine learning are also well known intelligent techniques/models. Most of the researchers or decision makers rely on computers to analyze their data in deep which are always stored in computer databases or files. However, databases or files are passive facilities. We can query or manipulate them only. They never actively tell us the knowledge deeply embedded or hidden in them.

In the social or geographic domain, few applications deploy GIS and data mining at the same time. In this paper, we use them to analyze social and geographic phenomena, and then explain the phenomena according to the mining result.

The rest of this article is organized as follows. Section 2 shows the application domains that have been developed. Section 3 introduces the mining techniques. Section 4 describes GIS systems. Case study and examples are presented in section 5. Section 6 concludes this article.

2. Related work

To date, many application domains have employed data mining or GIS techniques, but not both, to promote their business.

In health care domain, Mitchell [1] described several prototypical uses of data mining, including an expert system able to predict women at high risk of requiring an emergency C-section. Merck-Medco Managed Care, a pharmaceutical insurance and prescription mail-order unit of Merck, used data mining to help uncover less expensive but equally effective drug treatments for certain types of diseases or patients [2].

In finance domain, Bank of America deployed data mining to detect which customers were using which Bank of America products so they could offer the right mix of products and services to better meet customer needs [2].

In sports domain, Brain James, assistant coach of the Toronto Raptors professional basketball teams, used Advanced Scout, a data mining/warehousing tool developed by IBM especially for NBA, to create favorable player matchups and help call the best plays [3].

Besides, many commercial products of GIS have been released, such as ArcGIS [4], TomTom Navigator [5], Google Map [6], Yahoo Map [7]. Some of the products are for single client use, and others for web-based service. For analysis purpose, the ArcGIS is much more mature than others since it can perform almost every

type of geographical analysis. or mobile or navigation purpose, Garmin and TomTom have released many products in this domain.

3. The "Mining" Techniques

Data mining is the process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data collected in a large database. Its purpose is to identify trends and patterns in data so that users can extract hidden predictive information from the database. It is a powerful technology with great potential to help researchers focus on the most important information in their raw data.

Machine learning is a complex process. Computers are sometimes good at learning concepts. A concept is a set of objects, symbols, or events grouped together due to sharing certain characteristics. Concepts can be well designed and structured for future retrieval and management. Common concept structures include trees, rules, networks, and mathematical equations.

3.1. Types of Learning

Many types of data mining techniques adopt induction-based learning [8], which is the process of forming concepts and definitions by observing concept examples and concept objects to be learned, as the core algorithms to mine knowledge. Learning can be classified into two types: supervised and unsupervised.

Supervised learning is a learning model that intercepts instances of concepts representing animals, plants, and the like, or labels given to individual instances, and then chooses what we believe to be the definite concept features. We can use supervised learning to build classification or prediction models from sets of data containing examples and non-examples of the concepts to be learned. Then the model (e.g., the decision tree.) is used to determine the classification or predict the outcomes of newly presented instances of unknown origin.

Unsupervised learning is a learning model that builds models from data without predefined classes. Data instances are grouped together based on specific features defined by the learning clustering system. Users have to interpret the meaning of the formed clusters with the help of evaluation techniques to determine whether the classification meets our requirements or not.

3.2. Data Mining and Data Query

Databases collect and store passive data in their predefined-format storages or data structures, from which users can retrieve the data and aggregate data. Data mining can mine the hidden rules or knowledge embedded in the raw data. Before deploying data mining as a problem-solving technique, we need to consider three questions.

(1). How to clearly define the problem? i.e., what we want to mine which gives us a mining direction.

(2). Does potential hidden meaningful data truly exist? If not, the mining process is in vain.

(3). Is the mining cost less than the profit gained from the mining process? If yes, we will lose much more during/after the process.

Without consideration of the three issues, a data mining is meaningless. There are four general types of knowledge that can help us determine whether data mining or data query is suitable for us.

(1). Data: sometimes data is also called shallow knowledge which can be easily stored in a database and manipulated by DBMS. Data query, for example, using SQL is enough. No data mining is required.

(2). Multidimensional data: Data of this type is often used to represent a multidimensional object in a multidimensional format. On-Line Analytical Processing (OLAP) [9] is an appropriate tool to manipulate this type of data.

(3). Hidden knowledge: patterns or regularities hidden in data that cannot be easily found using database query languages. Data mining algorithms are suitable for this type of knowledge.

(4). Deep knowledge: defined as the data that can only be found if we are given some hints or directions about what we are looking for. No current data mining tools and DBMSs are able to locate knowledge of this type.

Existing database query languages, such as SQL and QUEL, and OLAP are good enough to process data of the first two types [10]. Data mining leads us one step further to explore data of the third type. But no one dares to say that current mining techniques are sufficient to uncover all hidden knowledge. So, computer scientists have to work hard continuously.

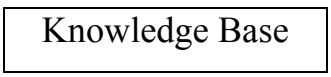

Fig.1 The framework of an expert system

3.3. Expert Systems

An expert system often comprises knowledge base and inference engine [11,12] as shown in Fig. 1. The former is the place to hold the knowledge of the system, whereas the latter is the mechanism that inferences new facts from exiting facts. From application viewpoint, an expert system is a computer program that gathers expertise from human experts to construct its knowledge base so as to emulate the problem-solving skills of human experts in specific problem domains. That means the program must solve problems using methods similar to those employed by the experts. Knowledge base is often implemented with rule-based approach. A rule, formatted by if x then y, can be created by data mining or extracted from human experts by knowledge engineers who are people trained to interact with experts to capture their knowledge, where x is the antecedent (or condition) and y is the action (or conclusion). To operate an expert system, inference engine tries to match known facts with "if" part (i.e., antecedent) of a rule to see whether the rule can be fired or not. If yes, the then part (action) of the rule is then executed. If not, inference engine continues to match other rules and facts.

4. Geographical Information System (GIS)

A GIS system (or GIS in short) is an application system for creating, storing, analyzing and managing spatial data and associated attributes [13]. In a more generic sense, a GIS is a software tool that enables users to create interactive queries, analyze spatial information,

edit data and display geographically-referenced information.

GIS is often used for scientific investigations, resource management, asset management, environmental impact assessment, city development planning, cartography, and route planning, for example, to identify a polluted area that need to be isolated from others.

4.1. Data Creation

Modern GIS technologies rely on digital information, for which there are a number of collection methods. The most common and popular one is digitization, where a hardcopy map or survey plan is transferred into a digital medium through the use of a digitization tool which is a computer-aided drafting (CAD) program with geo-referencing capabilities.

4.2. Data Representation

GIS represents real world objects (roads, wetlands, buildings) with digital data. Raster and vector are two common methods used to store data in a GIS for discrete objects and continuous fields. Raster images consist of rows and columns of cells where a cell stores a single value. The value recorded for each cell may be a discrete value, a continuous value, or a null value (if no data is available).

Vector uses geometries such as points, lines (series of point coordinates), or polygons (shapes bounded by lines), to represent objects. Examples include property boundaries for gardens represented as polygons and pond locations represented as points. Vector features can be made to respect spatial integrity constraints through the application of topology rules such as 'polygons must not overlap'. Vector data can also be used to represent continuously varying phenomena to show us the continuous change of objects, e.g., the annual development of last 20 years.

Raster datasets record a value for each point in the area covered which may consume more storage than representing data in a vector format that store data only as needed. Vector data can be displayed as vector graphics used on traditional maps, whereas raster data will appear as an image that may have a blocky appearance for object boundaries.

Additional non-spatial data can also be stored besides the spatial data, e.g., ages and genders collected through questionnaires or interview. In vector data, attributes of object are required. For example, a city inventory polygon may also have an identifier value and information about its population. In raster data, the cell value can be attribute information, or an identifier relating to records in another table.

4.3. Data Capture

Entering information into a GIS system consumes much of the time of its users/creators. There are a variety of methods used to enter data in a digital format into a GIS. Existing data printed on paper or film maps can be digitized or scanned to produce digital data. A digitizer produces vector data as an operator traces points, lines, and polygon boundaries from a map. Raster data produced by scanning a map could be further processed to generate vector data.

Positions from a Global Positioning System (GPS), a survey tool, can also be directly entered into a GIS.

Remotely sensed data also plays an important role in data collection. A sensing system consists of sensors attached to a collection mechanism. Sensors include cameras, digital scanners and so on, while collection mechanisms are often aircrafts or satellites.

The majority of digital data currently comes from photo interpretation of aerial photographs. After entering data into a GIS, it usually requires editing, removing errors, or further processing. For vector data it must be made "topologically correct" before it can be used for some advanced analysis. For example, in a city map, a polygon should be a closed area. Two adjacent lines of the object must connect together at an intersection. Otherwise, GIS will treat them as two disconnected line segments, i.e., errors such as undershoots and overshoots must also be removed or corrected. For scanned maps, blemishes on the source map need to be removed from the resulting raster. Otherwise two disconnected lines, for example, may become connected due to a dirtied spot located between the two lines and connecting the two lines.

4.4. Coordinate Systems

Two different maps might show data at different scales. Map information in a GIS must be modified or adjusted so that it can fit with information gathered from other maps. The modification or adjustment includes projection and coordinate conversions.

The earth is represented by various models, each of which may provide a different set of coordinates (e.g., latitude, longitude, elevation) for any given point on the earth's surface. As more measurements of the earth have been accumulated, the models of the earth have become more sophisticated and more accurate. In fact, there are models that apply to different areas of the earth to provide increased accuracy (e.g., North American Datum, 1983, NAD83, works well in North America, but not in Europe). Therefore, coordinate conversions are required.

A projection is the process of transferring information from a model of three-dimensional curved surface to a two-dimensional medium, e.g., a paper or a computer screen. Different projections are used for different types of maps because each projection particularly suits certain uses. For example, a projection that accurately represents the shapes of the oceans will distort their relative sizes.

Since much of the information in a GIS comes from existing maps, a GIS should benefit processing power of computer systems to accurately transform digital information, gathered from sources with different projections and/or different coordinate systems, to a common projection and coordinate system before we can correctly put the information of different sources together and then manipulate the integrated information precisely.

4.5. Current Systems

There are three common types of GIS hardware platforms: Single PC, Web-based (or Net-based) and mobile devices.

4.5.1 Single PC

We call this type of platforms resource-rich platforms since a PC as compared with a mobile device (e.g.,

pocket PC, smart-phone) often provides many more hardware and software resources. A GIS that operates in desktop or laptop has its own databases on which we can easily perform complex analysis or manipulation, such as overlapping, routing and 3D modeling. The major parameters that affect system performance include CPU capacity, memory capacity and so on.

4.5.2 Web-based

In a Web-based GIS system, the data is generally stored in network servers. The client side applications are just operational interfaces. Besides temporary results, they store nothing for the map currently manipulated. Platforms of this type are suitable for research teams or programmers in school in which most data are managed centrally.

Furthermore, interactive web GIS is most popular nowadays, such as the Google Maps. The Google Maps exposes an API, based on Asynchronous JavaScript and XML, enabling users to associate attributes with interactive maps.

4.5.3 Mobile Devices

GIS systems developed for running on mobile devices (such as cellphone, PDA) are rare. Their main applications focus on car navigation and disaster rescue. Due to limited device resources, venders often reduce down sizes of their digital geographic databases and confine their system analytical capabilities. So, most mobile systems are not able to analyze the geographic information as deeply as the system run on desktop.

5. Case Study

We had a research project concerning GIS and data mining, which is supported by Taichung City Government, Taiwan. More than 650 clients, whom were served by seven social service agencies for in-home services, made up the list of investigation for this project. These seven social service agencies have had contracting relations with the Taichung City Government in delivering in-home services to the elderly. A survey questionnaire was designed by our research team to be used as the main source for obtaining information regarding important variables of elderly needs and the satisfaction of clients towards the current service delivery system which carried out in-home services. GIS was used to enhance data storage and spatial analytical capacity, and to develop an in-home service information management system.

5.1. GIS Operations

Three main concepts of the project that use GIS to analyze social and in-home service resources are:

(1). Characteristics and satisfaction of clients. To understand the characteristics of elderly subjects who received in-home services, and to evaluate the satisfaction of the clients towards the current service delivery system for in-home services.

(2). How to use GIS to learn more about our services. To describe the use of GIS combining with other visualized statistical tools, such as correspondence analysis, and data mining in developing an in-home service management system to enhance our understanding of service satisfaction of the elderly and the issues of the elderly both for in-home services.

(3). How to use GIS to improve local government decisions. To explore the potential uses of information techniques for constructing decision support systems for local government who governs human services.

We analyzed the service satisfaction data and show them on digital maps. Thus we can easily understand that every recipient's satisfaction status. Furthermore, we used the "buffer zone" and "overlapping" functions to analyze the public facilities and in-home service centers' locations. Thus, we can learn which section is lacking of service center and/or public facilities. After that, the decision makers can refer to them to make the decisions more accurately and worthy.

5.2. Data Mining Application

We have analyzed the survey data gathered through questionnaires with a data mining tools. The following gives examples.

A. Completely free or partially pay the service fee against service satisfaction

(1). If (completely free) then the answer is "satisfied"

:rule accuracy 77.26%

:rule coverage 87.86%

The result represents that 77.26% of recipients, whose in-home services payment were totally paid by government, were satisfied with their in-home servants' services. Also, 87.86% of recipients who were satisfied with their in-home servants' services fitted this rule.

(2). If (Partially pay) then the answer is "satisfied"

:rule accuracy 64.71%

:rule coverage 11.61%

The result represents that 64.71% of recipients, whose in-home service payment were partially paid by government, were satisfied with their in-home servants' services. Also, 11.61% of recipients who were satisfied with their services fitted this rule.

We can conclude that most recipients enjoyed their in-home servants' services if the service payment was completely free or partially paid by government, no matter the services were truly what they wanted. That is, free lunch makes one feel happy and satisfied.

B. Participating home parties against service satisfaction

(1). If (the recipients have never taken part in home parties) then the answer is "satisfied"

:rule accuracy 76.05%

:rule coverage 62.01%

The result represents that 76.05% of recipients, who have never participated in home parties, were satisfied with their in-home servants' services. 62.01% of recipients who were satisfied with their in-home servants' services fitted this rule.

(2). If (the recipients have ever taken part in home parties) then the answer is "satisfied"

:rule accuracy 75.00%

:rule coverage 37.20%

The result represents that 75.00% of recipients, who have ever participated in home parties, were satisfied with their in-home servants' services. 37.20% of recipients who were satisfied with their services fitted this rule.

We can conclude that most recipients enjoyed their in-home servants' services no matter they have never or ever participated in home parties. The deep meaning is that most of the recipients feel lonely. They feel happy

and satisfied with the in-home services due to having the chance to talk with someone, even the one is their In-home servant.

6. Conclusion and Future Work

In the past, we have deployed GIS and data mining to analyze the data concerning social work, and got a series of results. In the future, we will apply these experience to analyze the data collected from Situn district regarding the development of this area during the past twenty or thirty years, and to uncover how the development of the Central Taiwan Science Park affects the development of Situn district. We expect to explore and learn what changes or advancement/regression were happened, and/or will happen.

In GIS, we expect to:

(1). Input, edit, store and manage the spatial data and attribute data collected from Situn district.

(2). Display data (maps, charts, and tables).

(3). Explore data (data query, geographic visualization).

(4). Analyze data (buffering, overlay, distance measurement, map manipulation, spatial interpolation, regions-based analysis, network analysis, etc.).

In Data Mining, we expect to:

(1). Code the questionnaires' result into databases.

(2). Use the supervised learning to mine the hidden knowledge embedded in the database.

(3). Display the mining result with GIS, and manually or automatically explain why they happen.

References

[1] T.M. Mitchell, "Does Machine Learning Really Work?" AI Magazine, vol.18, no.3, 1997, pp.11-20.

[2] V. McCarthy, "Strike It Rich, " Datamation, vol.43, no.2, 1997, pp.44-50.

[3] H. Baltazar, "NBA Coaches' Latest Weapon : Data Mining," PC Week, March 2000, pp.69-69.

[4] ESRI - The GIS Software Leader, http://www.esri.com/.

[5] Systèmes de navigation routière GPS portables de TomTom, http://www.tomtom.com/index.php.

[6] Google Maps, http://maps.google.com/.

[7] Yahoo! Maps, Driving Directions, and Traffic, http://maps.yahoo.com.

[9] H. Garcia-Holina, J.D. ullman and J. Widoma, Database System Implementation, Prentice Hall, 2000.

[8] R.J. Roiger and M.W. Geatz, Data Mining: A Tutorial-Based Primer, Addison Wesley, 2003.

[10] P. Adriaans and D. Zantinge, Data Mining, Addison Wesley, 1996.

[11] V.S. Moustakis, M. Lehto and G. Salvendy, "Survey of expert opinion: which machine learning method may be used for which task?" Special issue on machine learning of International Journal of HCI, 1996.

[12] M. Lavrac and S.K. Wrobel, Machine Learning: ECML-95, New York: Springer Verlag, 1995.

[13] Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/.