A SCHEME AND IMPLEMENTATION FOR SPATIAL DATA MINING AND INFORMATION SHARING ON THE WEB

Chen Xiaogang Wu Shaohong Wang Yingjie
(Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences)
Fax: 010-64851844
Email: xc@unimelb.edu.au
wush@igsnrr.ac.cn
wangyingj@netease.com

Abstract

With the transition to web-based applications, spatial data mining is required to extend its scope of concept and functionalities, to cope with multi-data, to manage web database and locate cleaned data from distributed database for knowledge discovery. This paper presents an overall scheme and alternative approaches to spatial data mining and information sharing in web environment, with particular emphasis on i) relevant prior knowledge and conceptual data structuring, ii) meta-data-centered web data management and data search, iii) application-oriented data access and iv) integrated knowledge discovery and representation. Data standardization is described for fundamental information sharing; Meta-database is specifically designed for web data management and database locating; User levels and demands are categorized; web tools for knowledge representation and information sharing are presented.

1. Introduction

Spatial data mining and knowledge discovery in database are increasingly attracting attention in a wide range of geo-referenced data exploitation in depth (Koperski, 1995, MacEachren, 1999). Most efforts grow to be paid in data modeling, visualization and knowledge extraction from spatial database. Spatial data mining is normally considered as an approach or a process or applications of algorithms to identify spatial patterns or relationships in abstract spatial data, and to present them in an understandable way (Fayyad, 1996, Estivill-Castro, 1998). The intention is to explore seemingly unrelated data and present them in an organized way in order to reveal spatial phenomena and extract spatial knowledge, so that data can be fully utilized and interpreted into knowledge.

Challenges emerge with the rapid development of very large database with high dimensionality and complex relationships, Internet-based technology and expanding popularity of web applications. The data mining focus is being or has been transmitted from i) local database into network-based distributed database; ii) single data formats into multiple forms; iii) limited applications into world-wide domain; iv) high level of demands into manifold categories. Accordingly, spatial data mining not only limits to identifying hidden patterns or relationships, but extend to dealing with wide range of issues related to spatial data mining and knowledge discovery, such as multi-types of data treatment, data organization and management, metadata creation and web-site management. It relates to a certain kind of sciences, techniques as well as art. All these contribute to increasing the complexities and difficulties of data mining and knowledge discovery in spatial database. It calls for an urgent requisition to integrate techniques and art, deductive and inductive approaches, statistical and visualized modeling, iterative computer-based algorithm and human interactivity under the framework of strategy and methodology, with an attempt to discovery hidden knowledge and construct knowledge base for future prior input.

Basically, there are two alternative solutions for improving spatial data mining, one aims at improving computer-based algorithms to effectively handle multiple data sets; the other focuses on improving data organization and management incorporating human knowledge, with the goal of reducing search dimension and leading to quick final results. Due to the differences in software and hardware environment, levels of user demands, development emphasis and stages, one of them or combination of them would be given first priority. By addressing an overall strategy and methodology, this paper deals with spatial data mining and information sharing particularly on user-driven information publication and sharing, metadata-centered web data management and application-oriented data access, combined with knowledge representation by inference and geo-visualization. Effective data management is realized by data standardizing, data structuring, data cleaning and deriving for various applications. Data acquisition on the web mainly counts

on packed data at the categorized levels of users, and on the matadata-based data locating and extracting. Integrated web tools for data access, knowledge representation and information sharing guarantees the implementation, which has been carried out partially as a case study on the information sharing for social and economic development (Wu, 2001). Techniques include metadata management, web data locating, data characterizing and summarizing, geo-classification and web-based geo-visualization.

2. Mining Approach and Process

2.1 Task-driven approach

Data mining and information sharing on the web is challenging all sides from scientists and users. Scientists tend to solve problems by improving techniques. However, with the availability of vast amounts of data and the accessibility of web resources, it comes true to acquire multi-types of data on the web site. Web allows easy access to distributed data and acquisition in cyberspace. Scientists have noticed the trends that there has been a shift away from hypothesis-based deductive approaches, using small amounts of data, towards observation-based inductive approaches (Tang, 1992). Spatial data modeling, web database management and metadata creation all contribute to observation-based inductive approaches to mining data and sharing information and knowledge. It is obvious that conventional mining techniques are not sufficiently capable to meet most of needs, especially from multi-users. Choosing applicable approaches for mining data and sharing information is essential to successful attainment of knowledge discovery. Normally, there is a three-pronged approach: technology-driven (what we can do), perception-driven (what make sense), and task-driven (what users want) (Encarnacao, 1994). For the overall framework of data mining and information sharing by analyzing the aim of project and current limitation of techniques in terms of human interactivity and adaptive capability, task-driven or user-driven approach is assigned first priority on the basis of end-user demands, geo-applications, multi-dimensional data and relevant prior knowledge. Additionally, improving current techniques will facilitate the realization of users' objectives, and perceptual convention will make mined results meaningful and acceptable based on human perception (Table 1).

Table 1. Priority selection from a three-pronged approach

Series of Wants	First Priority	Second Priority	Third Priority	Comments
Users' needs	Task-driven	Perception-driven	Technology-driven	Task-driven first, Technology-driven second & Perception-driven last
Data Standardization	Task-driven	Technology-driven	Perception-driven	
Data Structuring	Technology-driven	Task-driven	Perception-driven	
Data Cleaning	Task-driven	Technology-driven	Perception-driven	

2.2 Mining and Sharing Process

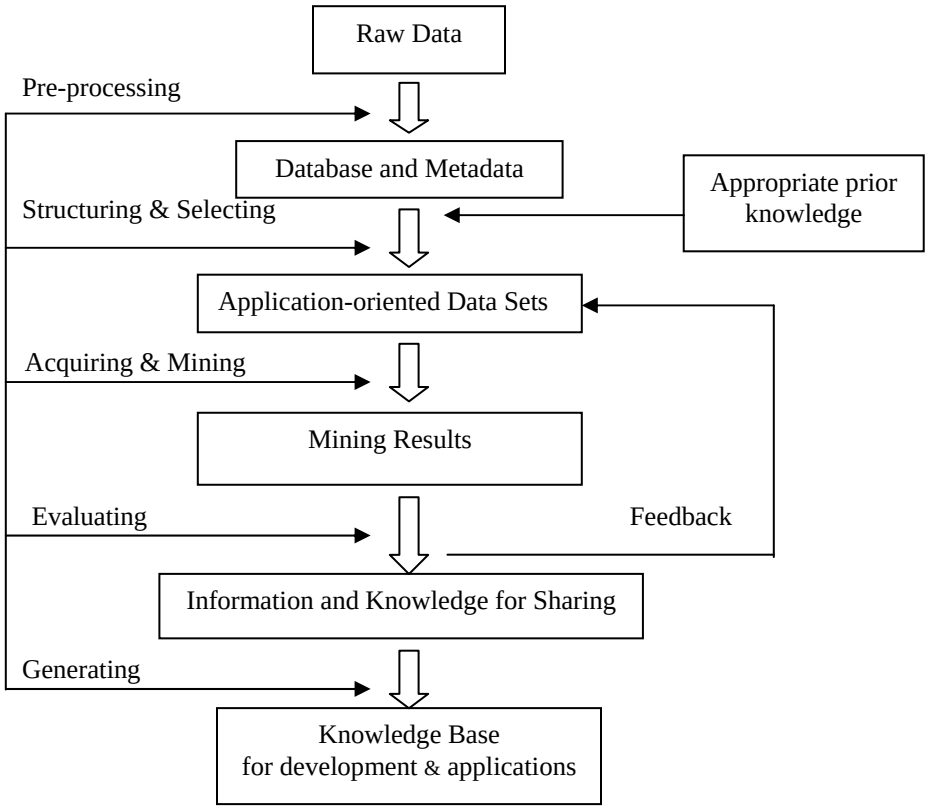
Instead of conventional hypothesis-based data mining with special emphasis on algorithms, observation-based spatial data mining on the web will focus on data organization and standardization, meta-data creation and navigation, database management.

Spatial data mining process is defined, following the task or user-driven approach to information and knowledge sharing (Fig.1).

2.3 Research Focus

As the transition of data mining focus seems to appear, multi-dimensional and dynamic forms of data, complex interrelations and auto-correlation between/among geo-phenomena, expanding popularity and diverse applications of web, manifold human demands and behavior should be fully considered under scrutiny. For on-going research on spatial data mining and information sharing, the preliminary research and implementation is emphasized on i) users' demands, perception and behavior, ii) data standardization, organization and management, iii) meta-data design and creation, iv) database and web site navigation, v) data type analysis and modeling, vi) spatial patterns identification and interpretation and vii) knowledge representation and sharing.

Figure 1. Data mining process for information sharing



3. Design Strategies

Following the task-driven mining approach, satisfying users' demands at different levels is prior, based on users survey and observation. The contents and details, formats of information sharing and forms of knowledge representation, style of data access and convenience of interactivity are fully brought into design.

3.1 User-centered scheme

The purpose of data sharing is to present high quality of various data to potential users with multiple representations through the web and allows users to exploit the applications of data with most possibility. Consequently, data mining and information sharing is largely driven by levels of user demands. Presently, most users are specifically not interested in raw data, instead the results of data exploration in a given context. As a result, information and knowledge sharing is targeting mainly to general and application-oriented users. Typically, users are categorized as:

1) Popular users, who are mainly provided with general introduction on themes or project background, major contents contained, data source, statistics, documents and other common information. Information is presented in majority of characters, graph or tabular data.

2) In addition to above information, application users in related disciplines request thematic information supplied on the web for their specific applications, such as presenting visualized maps for addressing important environmental and social issues, attaching graphs to show thematic patterns or trends, citing analytical reports in forms of maps, tabular data and relevant detailed document. These users may sometimes have special demands and request approaches to specific problems in a certain area, which requires flexible and integrated functions for problem solving.

3) Professional users, who may not be satisfied with the above information and require further information for their interactive exploration of data, may extract data from on-line database, or visualize multi-variable data dynamically, and model application-oriented spatial relationships. It is designed in the

information sharing system to incorporate metadata at two levels and data dictionary with varieties of details, which could be used for database guide and consultation in advance of data extraction.

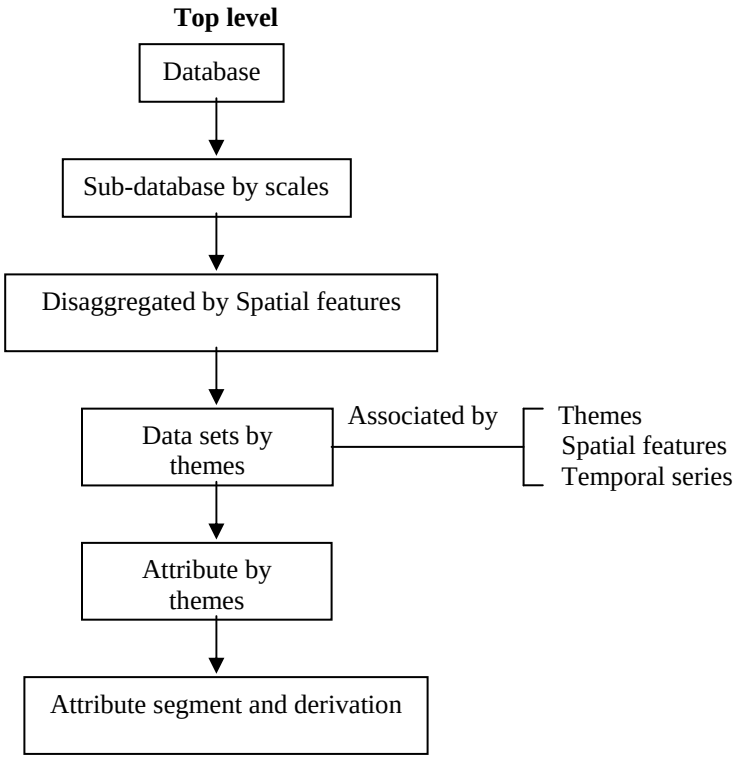
3.2 Data organization

Organizing data in flexible and convenient forms could greatly improve the efficiency and effectiveness of data access, dataset generation and data representation in a variety of forms.

1) Data structuring

A nation-wide sharing database of territorial physical and economic resources is designed to meet the demands of different levels of users, particularly various levels of government for decision making, educational cycles for teaching, scientific domains for research, mass media or users of interests. Data on environment, land cover and land resources, soil, fragile eco-belts, population, socio-economy, life quality and sustainable development, which are related to widely-concerned issues, are selected for the first stage of implementation. Moreover, it is required that data be structured in hierarchical and associative ways to facilitate users access and adapt to web environment. Hence, data are modeled by scales, attribute components, spatial features (by region, province, county and city proper or by other application units), temporal series (by year and month), element layers and so on (Fig.2).

Figure.2 Conceptual data structuring by hierarchy and associations



2) Data set deriving

To derive theme-related data sets by scales from primary database mainly serves for data mining and knowledge representation towards further applications driven by levels of users in web environment. It involves extraction or aggregation of attributes or indicators, spatial divisions, temporal components and basic geo-features from database. Depended on hierarchically and associatively structured data, it is readily to group data into a variety of sets. Users may acquire these available data by interacting with web interface for straight applications or to tailor them for specific uses.

3.3 Multi-forms of data sharing

Text, tabular data, graphs, maps, images, on-line geo-visualized maps, spatial/attribute query results and statistical models for applications are typical forms of information for sharing. Data in various formats could be accessed to derived database through web database server.

3.4 Guide for information sharing

Information guide is preferable for data sharing on the web. Viewing data on the web is mainly guided by themes in associations as theme, hierarchy and spatial relations, etc., on which web page connection is based. It is designed that themes are organized in combination of hierarchical and linear structure to define the common path of data access. In addition, users might lose their way while viewing data, guiding page could be returned directly from any current web pages. Returning path to homepage or database web pages is specifically designed in case of way loss, and theme-associated path between web pages is constructed for concurrent study on relevant subjects.

4. Data Standardization

For spatial information sharing, classifying geo-referenced elements and other application-oriented disciplines or activities, under nation-wide standards, to aggregate data according to a certain standard scheme is essential in terms of sensible data semantics and knowledge construction. Followings are examples of data standardization scheme applied in the project.

4.1 Classification and code for administrative divisions

It was formulated nation-widely in 1980's, and later improved and updated by years while administrative divisions have been changed at the level of province or city/county. Normally, a county is subject to dividing into more units or more counties merging into one unit. Occasionally, a county may be separated and each part affiliating to more administrative units.

The classification and code for administrative divisions are standardized at the levels of province, county and city proper, representing the fundamental units for information sharing at nation-wide scale. The standard mainly serves for thematic applications, at point-based and area-based data aggregation.

4.2 Classification and code for national economic activities

It was drafted by State Bureau of Statistics and formulated nation-widely in 1980's and later improved and updated according to the economic and social development with reference to international standards. The economic database refers to the standard for economic activities classification and coding. Forthcoming industrial and agricultural database, and future database for tertiary industry have been or will be designed upon the standard.

4.3 Classification and code for varieties of disciplines

The detailed classification and code for China Soil Database is based on the standard classification within the discipline and approved by relevant authority. World Soil Classification Scheme issued by FAO is listed for comparison study by professional users. The classification and code for land use, land resources and so on, abide by rules of corresponding disciplines.

4.4 Classification and code for geo-referenced elements

The general classification and code for geo-referenced elements was drafted by State Bureau of Surveying and Mapping, and used for designing and establishing National Topographic database. Referred to the standard in principle, other derived geo-referenced database extends it for specific applications, which are also standardized in a certain domain.

Besides, the standard projection for small-scale maps is adopted according to the national convention. The import and export format is set for spatial data conversion.

5. Methods, Techniques and Implementation

5.1 Metadata guiding

Metadata is the data about data for users' on-line consultation to check the availability and applicability of data supplied on the web in advance of various applications. Metadata is standardized for data description. In the light of data exploration, metadata serves for reducing available and applicable data

search dimension on the web, locating data with much efficiency and providing information semantics. There are a few selective options for locating database through metadata:

1) Full title of database: Database could be located by selecting full title in pull-down selection box, where total database titles are listed. Fresh users are suggested for this option.
2) Short title of database: Frequent users may type in short form of title for each database to locate data required.
3) Abstract of database: This option could guide new users to the most proper database they want.
4) Key words: Relevant database will be prompt up as long as key words are related to them for users to consult cross-cutting issues or topics.
5) Semantic search: Semantic search could bring intelligent structure to the meaningful contents of web pages by extending current prevailing web structure, and assists in locating data semantically. It might be developed during next stage.

In addition, more detailed information about data property in each database, such as field name, format, length, description, measure and so on, could be consulted in data dictionary as well.

5.2 Geo-visualized Mapping

The most important form of representation is visualized map. Visualization mapping defines an abstract visualization technique by establishing a set of bindings between the manipulated data and visualization primitives, such as geo-referenced elements and parameters, symbols, composition rules, perception convention, etc. It offers a means to combine and view several channels (attributes) of data concurrently (Mark Gahegan, 1999). Based on users' demands, a series of visualized maps with a variety of data attributes, representing spatial distribution, patterns and trends of objects, are produced beforehand and loaded on the web, and relevant numeric data for mapping are attached for detailed reference. These maps mainly serve for application users in their disciplines or levels of government for general information and decision-making.

5.3 Statistical modeling

Comparison, summarizing, classification and clustering are designed in the first generation of prototype, aiming at finding out knowledge hidden in statistical data distribution, patterns and aggregation. User interface is designed for selecting mining methods. Frequently selected ones are comparison and summarizing, which is comparatively easy to realize on the web in the first stage of mining, considering the rate of data access, selection, conversion and representation. Classification and clustering are emphasized in spatial visualization while data are classified and represented.

5.4 On-line visualization and spatial query

Professional users focus on mining data in a variety of space and in multi-forms of representation. It is obviously that static data and visualized maps could not support such exploration. Instead, on-line data visualization technique is aided dynamically by web mapping software developed using VB and MapObject, which has functions of geo-coding, interactive data access, map visualization, map zoom in/out, roam, pan, and query by location and attribute. Users can interactively select single or multi-data indicators, and ways of representation, and self-define data classification and symbol for representation. Data aggregation by various administrative divisions can also be decided in advance of visualization for the demand of mining data and generating spatial distribution, patterns and trends at different spatial levels.

5.5 Knowledge discovery

By integrating inductive learning methods with deductive database technologies in the context of knowledge discovery from database, spatially statistical models are applied in the area of users' interests. GDP is a popular indicator used in China for revealing economic growth. Consequently, GDP level (L), growth (G) rate and stability (S) are selected as variables for assessment. Each category of statistical indicator data by province from 1990 to 1998 is averaged. L, G and S are combined to deduct some categories on economic level, growth and stability, and then inducted into several spatial and attribute patterns by classifiers. Besides, population projection by sex at provincial level is explored in the year of 2005, 2010 and 2020. Outcome is represented in tabular data as well as visualized maps. On-line spatial and attribute data processing and analysis in web environment is under study during the first stage of knowledge representation and sharing.

6. Summary and Envision

We have designed, developed and partly implemented a spatial data mining and information sharing system on the web, which has following features or functions:

6.1 Users-centered and task-driven strategy and applications

In order to share information on the web, we propose an overall scheme and methodology for data mining and information sharing driven by user demands for valid and useful knowledge. Basically, three levels of users are categorized, i.e. popular users, application users and professional users. We expect it could be positively suggested for scheme planning.

6.2 Well-organized and structured database and derived application-oriented data sets

Data organization is followed to structure data by spatial units, temporal series, attribute indices and categories, hierarchical levels, element layers and so on. Derived database are created by considering the spatial units, temporal series, attribute class, hierarchical levels of original data for data transmission on the web when users are guided to select data by any categories.

6.3 Metadata-guided database location and data extraction

Considering the limitation of web environment constrained by software and hardware or other network-related factors, we merely provide easy-to-use and quick-to-run tools for users to acquire data via metadata-guided page by inputting optional selection or extract data from database on request. As criteria for existing database are met, candidate databases are prompted for users.

6.4 Combination of statistical and geo-visualized approaches for knowledge discovery

Statistical models are used to describe data patterns and display data in tables or graphics. Geo-visualization is to identify the geo-patterns, distribution and trends of spatial data over the nation-wide scale by province or county. Spatial and attribute data can be readily accessed dynamically in request-return feedback mechanism. By integrating inductive learning methods with deductive database technologies from database, spatially statistical models with prior knowledge from each discipline and geo-visualizing mapping are combined for data mining and knowledge discovery in the area of users' interests.

There should be two major prospecting tasks envisaged for future research into a more knowledgeable prototype:

i) Tracking user interests by web data mining

It aims at studying users' behavior in detail by recording times and duration of their selection regarding to types of web pages, database, data sets, data formats, representation of knowledge, etc., to improve the mechanism of user-centered and task-driven applications. Also, web users' survey could be done in conjunction with automatic web recording into a service database.

ii) Semantic search

By using XML and RDF to express both data and rules for reasoning, semantic search could bring intelligent structure to the meaningful contents of web pages by extending current prevailing web structure. It indicates the helpful assistance in making inference for locating required web site, data sets and other relevant information.

Main References

Encarnacao, J., Foley, J., Bryson, S., Feiner, S. and Gershon, N. (1994), Research Issues in Perception and User Interfaces, *IEEE Computer Graphics and Applications*, pp. 67-69.

Estivill-Castro, V. and Murray, A. (1998), Mining spatial data via clustering, Proceedings of 8th International Symposium on Spatial Data Handling, edited by Poiker, T and Chrisman, N

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), From data mining to knowledge discovery: An overview. In: *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: AAAI/MIT Press, pp.1-34

Gahegan, M.N. (1996), Visualization strategies for exploratory spatial analysis. Proceedings, *Third International Conference on GIS and Enrironmental Modeling*, Santa Fe, NM: NCGIA.

Koperski, K. and Han J. (1995) Discovery of spatial association rules in geographic information database. *In Advances in spatial databases, proceedings of 4th Symposium, SSD'95*. Springer-Verlag, Berlin, pp. 47-66

MacEachren, A., Wachowicz, M., Edsall, R. and Haug, D. (1999), Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualization with knowledge discovery in database methods, International Journal of Geographical Information Sciences, Vol.13, No.4, 311-334

Tang, Q. (1992), A personal visualization system for visual analysis of area-based spatial data, Proceedings, GIS/LIS'92. Vol. 2, *American Society for Photogrammetry and Remote Sensing*, Bethesda, Maryland, USA, pp.767-776

WU Shaohong, CHEN Xiaogang, et.al., 2001, Study on Land Resources Information Sharing, Resources Science, Vol. 23, No. 1. 49-53, (Chinese).