Spatial Data Mining Approaches for GIS – A Brief Review

Mousi Perumal, Bhuvaneswari Velumani, Ananthi Sadhasivam, and Kalpana Ramaswamy

geodatabases representing spatial features, with respect to latitude and longitudinal positions. Geodatabases are increasing day by day generating huge volume of data from satellite images providing details related to orbit and from other sources for representing natural resources like water bodies, forest covers, soil quality monitoring etc. Recently GIS is used in analysis of traffic monitoring, tourist monitoring, health management, and bio-diversity conservation. Inferring information from geodatabases has gained importance using computational algorithms. The objective of this survey is to provide with a brief overview of GIS data formats data representation models, data sources, data mining algorithmic approaches, SDM tools, issues and challenges. Based on analysis of various literatures this paper outlines the issues and challenges of GIS data and architecture is proposed to meet the challenges of GIS data and viewed GIS as a Bigdata problem.

Keywords: GIS, SDM, Geodatabases, Bigdata, Spatial and non-spatial data, Topology.

1 Introduction

Geographic Information System (GIS) has emerged as a new discipline due to the development of communication technologies. GIS is applied in various domains to infer information with respect to location. Enormous amount of data is generated in the form of image, fat files from sources like satellite imaginary sensors and other devices. Understanding of information stored in these large databases requires computational analysis and modeling techniques. Spatial data mining has emerged as a new area of research for analysis of data with respect to spatial relations. SDM techniques are widely used in GIS for inferring association among spatial attributes, clustering, and classifying information with respect to spatial attributes.

The objective of this paper is to provide with a brief summary of GIS data models data sets, data sources to provide better understanding of GIS for analyzing data analysis using data mining techniques. This paper is organized as follows the section 2 provides with a detailed over view of GIS data sources, data representations. Section 3 provides with description of SDM tasks applied in various domains of GIS data. This paper also presents with the overview of SDM tools for GIS. Section 4 describes the issues and challenges with respect to GIS data set are discussed and architecture is proposed for the same finally drawn by conclusion in section 5.

2 Overview of GIS

The development of information and communication technologies in GIS Domain has generated huge volume of data representing spatial information of water bodies, forest reserves, urbanization, etc., GIS databases stores spatial and non spatial data received from heterogeneous components connected, with each other such as sensors, laptop, mobile etc. Analysis of data deposited in GIS has gained importance in domains related to knowledge management and data mining. Recent widespread use of spatial databases has lead to the studies of Spatial Data Mining (SDM), Spatial Knowledge Discovery (SKD), and the development of SDM techniques. GIS can be viewed as collection of components such as Data, Software, Hardware, Procedures and methods used by people for analysis and decision making with respect to location. Fig 1, represents the components of GIS. The focus of this section is to provide with an overview of GIS data source, data formats, trends and Data Mining applications in GIS.

People
Data
Software
GIS
Procedures
Hardware

Fig. 1. Components of GIS

Spatial Data mining techniques combined with widely used in various studies to mine interesting facts associated in domains Transport, Tourism, Soil quality monitoring, water resource monitoring, and deforestation [7], [9], [13], [17], [20-22], [24-25], [27-29], [31], [40-42], [44-45], [48-49], [54], [60], [62-63], [73], [75], [78], [82], [87].

2.1 Data Sources

The Geodatabase is used as a "container" used to hold a collection of datasets for representing GIS features. The features of GIS systems are stored in form of tables and raster images. The various dataset of GIS available are listed in Table 1 with its description.

Table 1. GIS Data Sources

Data source	Site	Description	O/P
Yahoo! BOSS	http://www.yahooapis.com	BOSS (**Build your Own Search Service**). Provides a facility of Place finder & Place Spotter to make location aware.	*P
CityGrid	http://developer.citygridmedia.com http://docs.citygridmedia.com/display/citygridv2/Getting+Started	Incorporates local content into web and mobile applications.	*O
Geocoder.us	http://geocoder.us	Provides latitude & longitude of any **US** address , Geocoding for incomplete address ,Bulk Geocoding and Calculates distances	O
GeoNames	http://www.geonames.org	It contains geographical names, populated places and alternate names. All categorized into one out of nine feature classes.	O
US Census	http://www.census.gov http://www.census.gov/geo/www/tiger	Offers several file types for mapping geographic data based on data found in our **MAF/TIGER** database. **MAF**-**M**aster **A**ddress **F**ile. **TIGER** -- **T**opologically **I**ntegrated **G**eographic **E**ncoding.	O
Zillow	http://www.zillow.com	Zillow is a home and real estate marketplace.	O
Natural Earth	http://www.naturalearthdata.com	Natural Earth is a public domain map dataset available at 1:10m, 1:50m, and 1:110 million scales.	O
OpenStreet Map	http://www.openstreetmap.org	OpenStreetMap is a free worldwide map, created by many people.	O
MaxMind	http://www.maxmind.com	GeoIP - IP Intelligence databases and web services minFraud-transaction fraud detection database	O
ArcGis, MapGis.	http://www.esri.com/software/arcgis	Helps to organize and analyze geographic data	O
OpenEarthquake Data	http://earthquake.usgs.gov/earthquakes/map	Gives the databases related to earthquake	O

*P - Licensed Data Source, *O – Open Source Data

2.2 Data Representation in GIS

GIS Systems collects data from various heterogeneous data sources from wide range of communicating devices. The data from the communicating devices is available in different representation and file formats. IN GIS the data representation from these devices is classified into two main categories as Raster and Vector Data types. Fig 2 represents the visual representation of GIS data type. Raster is a two dimensional data type, which stores the value of pixel colors of raster images in a cell and the attribute values are continuous in nature. Raster data type is used to represents information

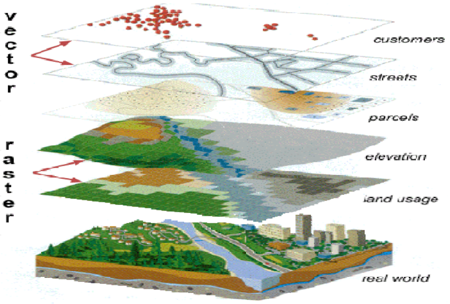


Fig. 2. Visual representation of GIS data type (ref. 91)

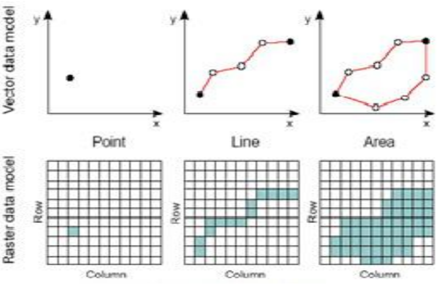


Fig. 3. Data models of GIS

from sources such as air photos, scanned maps, elevation layers; remote sensing data. Vector data types are used to represent discrete features in GIS and have a layered architecture representing point, line, and polygon. Vector data types are used to represents information from sources such as roads, rivers, cities, lakes, park boundaries with a layered hierarchy. The data model of GIS is given in Fig 3.

Table 2. File Formats in GIS

Raster Data		Vector Data	
Description	**File format**	**Description**	**File format**
Arc/Info ASCII Grid, Binary Grid,	prj, adf	ESRI Generate Line, shapefile	arc, shp
ADRG/ARC Digitilized Raster Graphics	gen,thf, gen, jpeg, tif	MicroStation Design Files	Dgn
Magellan BLX Topo	blx, xlb	Digital Line Graphs	Dlg
Bathymetry Attributed Grid	bag	Autodesk Drawing, eXchange Files	dwg,dxf
Microsoft Windows Device Independent Bitmap	bmp	ARC/INFO interchange file	e00
BSB Nautical Chart Format	kap	Geography Markup Language	Gml
VTP Binary Terrain Format	bt	ISOK	kf85
Spot DIMAP (metadata.dim)	dim	MapInfo Interchange Format	mif,mid
First Generation , New Labelled USGS DOQ	doq	Spatial Data Transfer System	Sdts
Military Elevation Data (.dt0, .dt1, .dt2)	dt0, dt1, dt2	Scalable Vector Graphics	Svg
Arc/Info Export E00 GRID	adf ,shx	Topologically Integrated Geographic Encoding and Referencing Files	Tiger
ECRG Table Of Contents (TOC.xml)	xml	Vector Product Format	Vpf
ERDAS Compressed Wavelets (.ecw)	ecw	Idrisi32 ASCII vector export format	Vxp
Eir – erdas imagine raw	bl,raw	Microsoft Windows Metafile	Wmf

2.3 Challenges

Analysis of GIS databases representing spatial information is complex due to the varied formats, representation and data sources. The various challenges involved mining spatial data from Geodatabases are listed below:

- Need help of domain experts to relate and understand Spatial and Non-Spatial Data.
- Selection and representation of data for mining from Geodatabases due to wide range of file formats.
- Understanding of information represented in image files (raster data).
- Selection and Transformation of spatial attributes from Non spatial attributes.

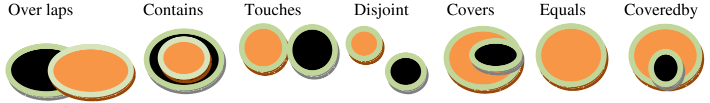
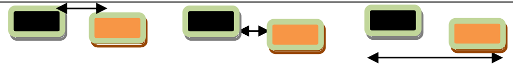
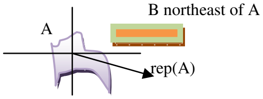
3 Spatial Data Mining for GIS

Spatial Data Mining or knowledge discovery in spatial databases refers to the extraction of implicit knowledge or other patterns that are not explicitly stored in spatial database [47][34][72][56]. The word spatial refers to the data associated with the geographic location of the earth. A large amount of spatial data has been collected in various applications, ranging from remote sensing to GIS, computer cartography, environmental assessment and planning. The collected data is huge in such a way that it's far of human knowledge to analyze it, new and efficient methods are needed to discover knowledge from large spatial databases [38]. Spatial data mining is the analysis of geometric or statistical characteristics and relationships of spatial data. The advance in spatial data has enabled efficient querying of large spatial databases.

Over the last few years, spatial data mining has been often used in many applications, like Marine Ecology [89].remote sensing[29], space exploration[79], traffic analysis, climatic change [17]NASA Earth Observing System (EOS), Census Bureau, National Inst. of Justice, National Inst. of Health etc. there is a need of evaluating the structural and topological consistency among multiple representations of complex regions with broad boundaries, mining the frequent trajectory patterns in a spatial–temporal database [4], extracting the spatial association rules from a remotely sensed database [3], generating polygon data from heterogeneous spatial information [71], and analyzing the change of land use [2]. Extraction of spatial rules is one of the main targets of spatial data mining [72], [83] and has been used in many real time applications. [14] Proposed a model to that selects the locations of land-use by using the decision rules generated by nearest neighbours.

Mining of Spatial data set is found to be complex as the spatial data is not represented explicitly in geodatabases [52]. Analysis of Spatial data has become important for analysing information with respect to location. So, analysis of spatial data requires mapping of spatial attributes with non spatial attributes for effective decision making. Spatial data is also known as geospatial data contains information about a physical object that can be represented by numerical values in a geographic coordinate system. Spatial data are multidimensional and auto correlated. Spatial data includes location, shape, size and orientation. [18]. Non spatial data is also called as attribute or characteristic data which is independent of all the geometric considerations. Non spatial data includes height, mass and age, etc. The spatial attributes are classified in three major relations as Distance relation, Direction relation and Topological relation. Topological relation is always non spatial data, so it requires spatial mapping to convert non spatial to spatial data.[57],[47]. Table 3 provides with an overview of spatial relations for modelling attributes.

Table 3. Classification of Spatial Relations

Topological relation	 Over laps Contains Touches Disjoint Covers Equals Coveredby
Distance Relation	
Direction Relation	

3.1 Spatial Data Mining Tasks

Brief overview of Spatial Data Mining tasks are discussed in the section below. Mining of Spatial Data using data mining techniques such as association, classification, clustering, and trend detection generates interesting facts associated in various domains. Spatial Data Mining tasks are generally an extension of data mining tasks in which spatial data and criteria are combined [50],[51],[71] to form various tasks to find class identification, to find association and co-location of Spatial and Non-Spatial data, make the clustering rules to detect the outliers and to detect the deviations of trends.

3.1.1 Spatial Classification

The spatial object has been classified by using its attributes. Each classified object is assigned a class. Spatial classification is the process of finding a set of rules to determine the class of spatial object[11]Spatial classification methods extend the general-purpose classification methods to consider not only attributes of the object to be classified but also the attributes of neighboring objects and their spatial relations. The spatial classification techniques such as Decision trees (C4.5), Artificial Neural Networks (ANN), remote sensing, Spatial Autoregressive Regression to find the group the spatial objects together. The classification problem is applied in the area of transporting for dividing spatial locations based on the area.

3.1.2 Spatial Association Rule

Association Rule Mining is the process of finding frequent patterns, associations, correlations among sets of items or objects in transaction databases, relational databases, and other information repositories Frequent pattern is described as set of items, sequence etc., that occurs frequently in a database the main motivation is to finding regularities in data[23] ,[61][67],[68].An association among different sets of spatial entities that associate one or more spatial objects with other spatial objects. The association rule is used to find the frequency of items occurring together in transactional databases. Spatial Association relation is based on the topological relation. Association Rule is an expression of the form X ==> Y. A spatial Association Rule describes a set of features describing another set of features in spatial databases. Spatial association is a rule A->B where A, B are set of predicates, the predicates can be Spatial or Non Spatial but needs at least one Spatial predicate [3],[18],[47].

Spatial Association is used to find positive and negative Association Rule Mining which extracts multilevel interesting patterns in Spatial or Non-Spatial predicates using topological relation[2],[90]. A multilevel Association Rule has been generated to find association between the data in a large database [14] and suggested a method by applying different minimum confidence thresholds for mining associations at different levels of abstraction.

3.1.3 Spatial Clustering

Clustering is a process of grouping the database items into clusters. All the members of the cluster have similar features. Spatial Clustering task is an automatic or unsupervised classification that yields a partition of a given dataset depending on a similarity function. Spatial clustering is based on the distance and direction relation. Spatial clustering techniques used ranges from partitioning method, hierarchical method, and density-based method to grid-based method. Similarity can be expressed in terms of a distance function.

3.1.4 Trend Detection

A spatial trend is a regular change of one or more non-spatial attributes when spatially moving away from a start object. Spatial trend detection is a technique for finding patterns of the attribute changes with respect to the neighbourhood of some spatial object. One of the trend detection techniques is kriging to predict the location from outside the sample. Table 4 describes spatial data mining tasks and its techniques used in various domains such as Transportation, Tourism management, Environmental and agriculture from various literatures

Table 4. Spatial Data Mining Technique

SDM Domains	Techniques/Methods	Usage	References
Transport	Add-on Environmental Modelling System (TRAEMS)	Add-on module to existing transport plan, provides rapid information based on traffic related outcomes	[20],[24],[40]
	GIS-based DSS (Classification Technique)	evaluates urban transportation policies, provides estimates of road traffic to traffic policies	[12],[48],[77]
	ArcView GIS 3.3 and ArcInfo (Clustering technique)	Represents geometry of streets and establish the connection between the GIS street data and the roadway links.	[41],[84]
	Transportation object-oriented modelling (TOOM)	Gathers data from the GPS trace, matches GIS and mathematical algorithms to propose a new schedule.	[24],[74],[78],[62],[85]
Tourism Management	Association technique	Used to integrate the ICT technologies with tourism	[21]
	Web GIS	Provides a new generation interface and expands the ways in which travel information can be accessed.	[31],[74]
	Tourism GIS	Used to provide users with a quick and convenient travel information query method	[7],[25],[26]
	Unified GIS database on the cycle infrastructure (UDCI)	Provide information about the overall cycle trail network on tourism	[53],[73],[78]
	spatial tourism interaction model or gravity model	Used to provide about green tourism potential, human resources, and the shortest distances among villages.	[74],[9],[28],[6],[21],[45]
Ecological model	Indicates the water catchments by integrating geographic area.	Maintains the Bio-Diversity	[18],[22],[75]
Argiculture	GIS and Saptial Data mining	Used to access the soil quality, water resource management	[69],[82],[86],[87]
land allocation (MOLA)	Location prediction	Provides the selection of land use models, illegal land fills	[8],[15],[16]
Environmental Decision Support System	Site selection, Location prediction, clustering	Provides environmental support of assessment of various relates feature with environmental degradation etc.,	[18],[22],[25],[44],[50],[59],[63],[65],[76],[89]
Health Care	Accessing of resource	Provides with an decision making support for various health related domains	[8],[14],[17]

3.2 Spatial Data Mining Tools

Spatial Data Mining tools are used by researchers to mine spatial relations among spatial datasets in various application domains. There exists many numbers of tools as opens source and propriety software. It is found that the tools are input specific in terms of file formats. So common file formats is not available for any specific tools which is a challenge for choosing the correct tool. Table 5 describes about various Spatial Data Mining Tools.

Table 5. Spatial Data Mining tools

Spatial Data Mining Tools	O/ P	Developer	Language used	Main Features and Data Source uses
DBMiner (DBlearn)	O	Data mining research group, Simon Fraser University, Canada	Data Mining Query Language	Supports data mining functions including (https://www.dbminer.com)
GeoMiner	O	Data mining research group, Simon Fraser University, Canada	Geo-Mining Query Language	Supports the data mining tasks (http://www.downloadcollection.com)
GeoDA(ESDA, STARS)	O	Dr. Luc Anselin	Python	Supports spatial autocorrelation statistics, spatial regression (http://geodacenter.asu.edu)
Weka-GDPM	O	University of Waikato, N Z	Java	Supports several standard data mining tasks (http://weka.software.informer.com)
R language (sp, rgdal, rgeos)	O	Ross Ihaka and Robert Gentleman at the University of Auckland,NZ	C, FORTRAN, Python	Used to analyze statistical and graphical techniques, (http://www.r-project.org/)
Descrates	O	Jankowski and Andrienko	Python	Used to visualize and analyse source data and results of the classification on the map.(https://www.descartes.com)
ArcGIS (ArcView, ArcInfo, ArcEditor)	p	Environmental Systems Research Institute (ESRI)	Python, Web API, .NET	Supports spatial analysis and modeling features including overlay, surface, proximity, suitability, and network analysis, as well as interpolation analysis and other geo statistical modeling techniques. (http://www.esri.com/software/arcgis)

4 Issues and Challenges

The issues and challenges in applying Spatial Data Mining in GIS can be viewed in terms of Integration of data and Mining huge volume of data. Architecture is proposed to address the issues of data integration and volume of data based on the analysis of data of GIS from literature is given in Fig 4.Data warehousing technology is used as a tool for data integration and stores summarized data. Currently a Bigdata approach has gained attention for mining data parallel using architectures like Hadoop and Mapreduce. In this paper we propose an architecture which can have a Bigdata platform modeled for representing a data warehouse. The other challenge in integration of data is semantic representation of information organized from various data sources. An ontology layer is proposed for semantic representation of data [13]. The data mining techniques can be applied above these layers by modifying the algorithmic approaches suitable for BigData architecture.

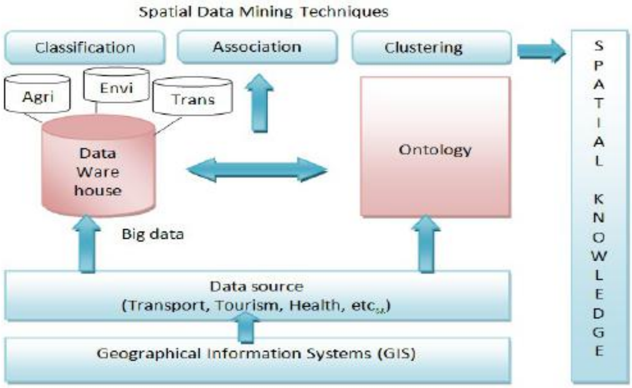

Spatial Data Mining Techniques
Classification
Association
Clustering
Agri
Envi
Trans
Data Ware house
Ontology
Big data
Data source (Transport, Tourism, Health, etc.,)
Geographical Information Systems (GIS)
SPATIAL KNOWLEDGE

Fig. 4. A Big Data Approach – Integration of GIS Data

5 Conclusion

This paper provides with a detailed survey on spatial data bases and its characteristics. A detailed analysis and description of GIS Data sources, data formats and data representation is presented from various literatures. The classical data mining algorithm applied for various applications in GIS are also discussed. On analysis it is identified that semantic integration of GIS datasets is necessary for analyzing spatial attributes with respect to non spatial attributes in all domains. The other major challenge of GIS databases can be viewed as volume and data formats in this work we have proposed an architecture for data integration using data ware house approach and ontology in GIS. The other major issue in GIS is huge volume of data generation for which we have proposed to view the problem as Bigdata and represented the same in our proposed architecture design. In future the proposed architecture would be implemented and tested for any specific domain. SDM tools would also be presented with a brief overview discussing the merits and limitations

References