Chapter 6
The Integral of Spatial Data Mining in the Era of Big Data:
Algorithms and Applications

Gebeyehu Belay Gebremeskel
Chongqing University, China

Chai Yi
Chongqing University, China

Zhongshi He
Chongqing University, China

ABSTRACT

Data Mining (DM) is a rapidly expanding field in many disciplines, and it is greatly inspiring to analyze massive data types, which includes geospatial, image and other forms of data sets. Such the fast growths of data characterized as high volume, velocity, variety, variability, value and others that collected and generated from various sources that are too complex and big to capturing, storing, and analyzing and challenging to traditional tools. The SDM is, therefore, the process of searching and discovering valuable information and knowledge in large volumes of spatial data, which draws basic principles from concepts in databases, machine learning, statistics, pattern recognition and 'soft' computing. Using DM techniques enables a more efficient use of the data warehouse. It is thus becoming an emerging research field in Geosciences because of the increasing amount of data, which lead to new promising applications. The integral SDM in which we focused in this chapter is the inference to geospatial and GIS data.

INTRODUCTION

Data Mining (DM) is a rapidly expanding field in many disciplines. It plays a significant role in human activities and has become an essential component in various areas and issues that employed to Knowledge Discovery (KD) process to analyzing large-scale data from different sources and perspectives (Martin et al., 2001; Krzysztof et al., 1996). It is greatly inspiring to analyze massive data types, which

DOI: 10.4018/978-1-5225-2031-3.ch006

includes geospatial, astronomic, climate, image and other forms of data sets. The fast growths of data are characterized as high volume, velocity, variety, variability, value and others, which can be collected and generated from various sources. The data includes, satellite, remote sensing, Geographic Positioning System (GPS), areal images, photographs, log files, social media, machines, video, textual, which are too complex and big to capturing, storing and analyzing, and challenging by traditional tools (Diansheng & Jeremy, 2009). These sources have strained the capabilities of classical relational Database Management Systems (DBMS) and spawned a host of new technologies, approaches, and platforms called "Big Data." Therefore, the potential values of geospatial data using DM techniques are great and are clearly established by a growing number of studies (Deepali, 2013; Ranga et al., 2012).

In the last two decades, the integrity of DM and Geographic Information Systems (GIS) was limited, and the actual spatial data analysis techniques suffer from the huge amount of complex data to process (Anselin, 1998). Indeed, earth observation data (acquired from optical, radar and hyperspectral sensors installed in terrestrial, airborne or space-borne platforms) is often heterogeneous, multi-scale, incomplete, and composed of diverse objects. However, the existing data analytics were traditional and doing very basic spatial analysis functionality, which confined to analysis that involves descriptive statistical displays, such as histograms and/or pie charts. Moreover, the complete DM process is a combination of many sub-processes, which includes data extraction and cleaning, feature selection, algorithm's design and other analytics of the spatial data (Zaragozi, et al., 2012). In many geospatial research works, the non-spatial data did not well addressed and synthesis by any of the various data analysis techniques. Based on these and other facts, we proposed an Integral Spatial DM (ISDM) to discuss and introduced a novel way of handling and analyzing geospatial and non-spatial data that allows flexibility to describe elements together to optimize spatially based decision-making process (Carlos et al., 2013; Xing et al., 2013).

In this chapter, we discussed the integral DM techniques and algorithms and its significance in GIS and spatial data analysis. The key intervention, and success of Spatial DM (SDM) are a clear business need that data analytics aligns with the Geo-referencing environment and IT strategies, which improve and optimize a spatial decision-making culture, a strong data infrastructure, the right analytical tools, and people skill in the use of spatial data analysis. It is thus becoming an emerging research field in Geosciences because of the increasing amount of data, which lead to being new promising applications (Gennady et al., 2006). Therefore, the need of DM in the domain of GIS is manifold advantages as to advance and automate spatial data search for hidden patterns in complex databases, offers great potential benefits of applying GIS-based decision-making. Recently, the task of integrating these two technologies has become critical, especially as various public and private sector organizations possessing large databases with thematic and geographically referenced data begin to realize the huge potential to gain valuable information and knowledge implicit. For example, to analyze Geo-referenced statistical data, searching for explanations of disease clusters, assessing the impact of changing land-use patterns on climate change and doing customer segmentation based on spatial location are mentioned among its applications (Luís et al., 2005).

This chapter research work is vitally essential to DM and GIS specialists, scientists and researchers by giving novel ideas and techniques of DM for GIS domain in which we focused the inference of geographic positioning or coordinating data that can include astronomical, satellite and spacecraft, and other forms of referenced image data. The DM algorithms are employed when analyzing spatial and related types of data on the use of spatial warehouses, data cubes, OLAP, and clustering methods. Therefore, the field of GIS is a well-identified domain of DM, which is complex due to the nature of spatial data types, relationships and correlation between features that an effectiveness of many algorithms depends