S. Sathappan* et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.4, Issue No.1, December - January 2016, 2621 – 2623.

A Novel Application of Spatial Data Mining in Air Pollution

S.SATHAPPAN
Research Scholar
Computer Science and Engineering
Sathyabama University, Chennai

Dr. D.C. TOMAR
Research Guide, Professor
IT Department
Jerusalem College of Engineering, Chennai

Abstract: **Spatial Data Mining is an exploratory process aimed at discovering hidden patterns from spatial data. The extracted knowledge can be used to perform efficient spatial prediction. It allows taking advantage of the growing availability of geographically referenced data and their potential richness. Spatial data mining techniques can be applied to various fields' namely health care, metrological data, traffic analysis, customer intelligence, transport management, urban planning and utilities industry. This paper proposes to support spatial data mining techniques of air pollution. A web based system is proposed to investigate the effect of meteorological and air pollutant elements on air pollution. It majorly consists of the collection, transformation and query and mining elements. The collecting element helps providing access to collection of data. The transformers convert the data into the required format and the query and mining element provide an interface to the user, for querying and mining requests and provide the results.**

Key Words : **Spatial Data Mining, Air Pollution**

I. INTRODUCTION

Indian metropolitan cities like Delhi, Mumbai, Kolkata, etc. have high emission of air pollutants, which is degrading the ambient air quality day by day. The degradation of air quality is a major environmental problem that affects many urban and industrial sites and the surrounding regions worldwide. Air pollution can reach levels, where it significantly influences human health, diminishes crop yield, and destroys infrastructure and patrimony. The inventorying of sources of emission, local meteorology and air quality gives the status of air quality. Most of the data is available, spatial data mining techniques can be applied to support public queries and spatial data mining. Spatial data mining is aimed to discover spatial and non-spatial relationships and the air pollution situations. The air pollutants concentrations and other metrological data are made available by various agencies. The results after mining the spatial data can help decision making, planning and forecasting. A predictive model may then be prepared to enact the pollution control regulation to achieve ambient standards/ goals. Spatial Data mining techniques can used to generate a better, efficient and cost effective approach for estimating the ambient air quality. The results can assist in making of strategies to control the air pollution. Town planners can plan to develop residential areas in areas with good air quality. Environmental engineers can concentrate on improving the areas with bad air quality and the public can estimate and forecast the air quality and meteorological condition by providing some parameters.

II. AIR POLLUTION IN MUMBAI

Mumbai is located on the western seacoast of India at 180 53' N to 190 16' N latitude and 720E to 72059' e longitude. It occupies and area of 437SQKm. The assessment of Mumbai's environment suggests that air quality and noise are issues that would require attention. MCGM monitors air quality for criteria air pollutants namely Sulphur Di-Oxide (SO2), Nitrogen Di-Oxide (NO2), Ammonia (NH3), Suspended Particulate Matter (SPM) and Respirable Suspended Particulate Matter (RSPM) regularly. Maharashtra Pollution Control Board (MPCB)'s ambient air quality monitoring network comprises of 16 receptor oriented monitoring sites spread all over Mumbai. Every day sites are monitored continuously throughout the year for four air pollutants. Air quality levels are evaluated for the for its compliance with ambient air quality standards for SO2, NO2, NH3, SPM, covering all monitoring sites under residential zone. Studies report a steep increase of RSPM concentration from 180 ug/m3 to 270 ug/m3 (annual average). The increasing concentration of SO2, NO2and SPM is really causing worry. Its concentration is beyond acceptable limits prescribed by National Air Quality guideline. Based on the monitoring of ambient air quality, fifty-one (51) non-attainment cities have been identified in the country in which the prescribed Respirable Particulate Matter (RSPM) levels, specified under the National Ambient Air Quality Standards (NAAQS), are not met. The Central Pollution Control Board, Delhi has proposed the Indian Air Quality Index (IND - AQI) in simple and lucid terms. A segmented linear function is used relating the actual air pollution concentrations (of each pollutant) to a normalized number. IND-AQI is primarily a health related index with the following descriptor words: "Good (0 - 100)", "Moderate (101 - 200)", "Poor (201 - 300)", "Very poor (301 - 400)", "Severe (401 - 500)".

The factors that are related to the air pollution are wind, rainfall, temperature, fog and air pressure. The sources are mainly re-suspension from road caused by vehicles, emissions from diesel and gasoline vehicle, domestic wood and refuse burning. Multiple setting up of reliable and continuous air quality monitoring and predicting system, building up of reliable dispersions model random data, application of

solutions thrown up by the model which may be hard and harsh, are some of the challenges .[7]

III. SPATIAL DATA MINING

Spatial data mining is the application of traditional data mining techniques to spatial data. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography. Spatial data mining methods are applied to extract interesting and regular knowledge from large spatial databases. A challenging problem for spatial data mining is to consider both spatial and aspatial aspects together, for example, detecting multivariate spatial clusters [1] deriving classification rules that involve spatial relations ,or discovering spatial association rules [5].

Clustering and classification is the task of grouping the objects of a database into meaningful subclasses (that is, clusters) so that the members of a cluster are as similar as possible whereas the members of different clusters differ as much as possible from each other. Both classification and clustering are aimed towards the division of series into differentiated groups in accordance to a similarity or distance measure. Different types of spatial clustering algorithms such as CLARANS, (clustering Large Applications based upon RANdomized Search) algorithm have been developed.. The main advantage of using clustering is that interesting structures or clusters can be found directly from the data without using any prior knowledge. Clustering algorithms can be roughly classified into hierarchical methods and non-hierarchical methods. Non-hierarchical method can also be divided into four categories; partitioning methods, density-based methods, grid-based methods, and model-based methods [6]. The task of classification is to assign an object to a class from a given set of classes based on the attribute values of the object. In spatial classification the attribute values of neighboring objects may also be relevant for the membership of objects and therefore have to be considered as well. [2][4]. For classification the number of groups and the characteristics of each group are known beforehand, whereas in clustering these are determined in the clustering process.

Spatial data mining can be categorized based on the kinds of rules to be discovered in spatial databases. A spatial characteristic rule is a general description of a set of spatial related data. For example the description of the general weather patterns in a set of geographic regions is a spatial characteristic rule. A spatial association rule is a rule which describes the implication of one or a set of features by another set of features in spatial databases. [1][3][4]. The spatial association rules are of the form $X \rightarrow Y$ (c%), where X and Y are sets of spatial or non-spatial association rule and c% is the confidence level.

In mining the metrological and the air pollution data we propose to apply spatial data mining methods that assume the existence of background knowledge represented on the form of concept hierarchy. The information thus becomes more and more general and remaining consistent with the other lower levels.[2][8]

IV. A FRAMEWORK FOR AIR POLLUTION MINING

Information that is available is in an unorganized and distributed form. The availability type and the reliability of information services are constantly changing and updating. The information is difficult to collect, filter evaluate and use in problem solving and decision making. This poses a problem of locating the information sources, accessing filtering and integrating information in support of spatial data mining is very difficult.

We propose to design a client-server based framework with an aim of supporting the spatial data mining of air pollution data. A solution to investigate the effect of meteorological and air pollution elements on air pollution using spatial data mining technologies is provided. It uses three types of agents consisting of the interface agent, collaboration agent and the information agent.

Interface agent

Interface agent interacts with the user accept the queries and give out the results. This comprises of the query and the mining agent. The spatial mining agent can use classification-based mining methods as the algorithm to mine the data and give out the results. For air pollution framework the three levels of hierarchy can be suggested, the first level consisting of the various values of IND-AQI, the second level consisting of values “good”, “moderate”, “poor”, “very poor” and “severe”. The third level has two values “acceptable “and “not acceptable”. The mining agent would carry out the spatial data mining process. The mining process takes the required data form the collaborator, carries out the spatial data mining and sends out the results to the user or the GIS database.

Collaboration Agent

The collaboration agent help the user perform the task by formulating the problem, plan solving and carrying out the plans through querying and exchanging information with the other software agents namely the transformers. The raw data that is available in different formats and frequencies are configured to make the data in a consistent format.

Information Agent

The information agent provides intelligent access to a heterogeneous collection of information sources i.e the collectors. These are designed to support the querying and spatial data mining of air pollution data. The meteorological and air pollution data are available from the MPCB and the Indian Metrological Department. The is collected and updated frequently form the data sources in different formats and are stored in the GIS database

The basic framework for the air pollution mining is as shown in the figure.

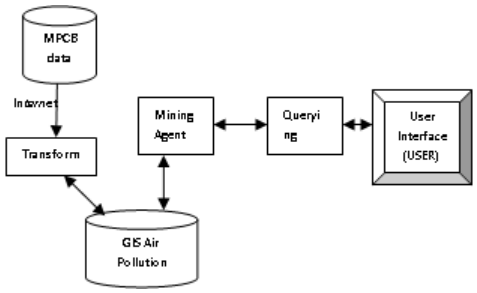

Figure 1: Mining Air Pollution data

The framework is divided into two layers. The front end accepts the user query request and gives out a detailed query report on the air pollutant dat. These results are as a result of mining or evaluating the on demand queries. The back end is responsible for the collection of the air pollution and metrological data transform the data as per the requirement and prepare it to be mined or queried for the front end. The events can majorly be to two categories namely the regular events and the events on demand.

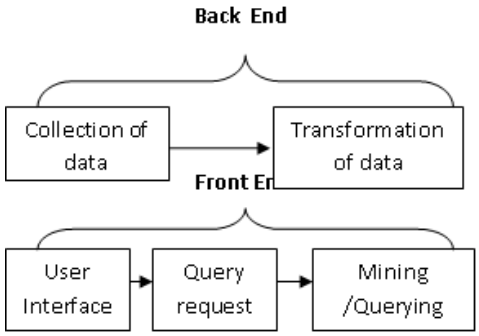

Figure 2: Tasks involved in Back and Front End

The scheduled events include the accessing resources from the Internet and collect the metrological and the air pollution data on a regular basis. The raw data the HTML format data collected are further sent to the transformers who transform the raw data and store it. This data is passed on to the mining and querying element that carries out the spatial data mining on the data. The results obtained after the mining are passed to the GIS databases.

The events on demand are the ad-hoc queries by the users. The user requests for a query which is evaluated by the query element and give out the results from the database. If the query element is not able to find the requested results due to lack of data it passes the query to the mining element. The mining element then coordinates with the transformers to extract, convert and analyze the data. The results then are returned to the query element to be delivered to the end user.

V. CONCLUSIONS

Spatial data mining the air pollution data acquired from the various departments and the web can be fruitfully used so as help the end user to understand and know the air quality at various locations. The system can provide a means for querying, data extraction and transformation. This system can be further enhanced to provide more features and support various other spatial data mining algorithms.

VI. REFERENCES

[1]. Cherkassky, V . and Mulier , F . (1998). Learning from data: concepts, theory, and methods . New York, NY : John Wiley & Sons .

[2]. E.M. Knorr, RT Ng, " Finding Aggregate Proximity Relationships and Commonalities in Saptial Data Mining", IEEE 1996

[3]. Jawaiei Han Michelle Kamber," Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[4]. Koperski, K. and Han, J. (1995) "Discovery of spatial association rules in geographic information databases," In 4th SSD, Portland, p. 47-66. Springer.

[5]. Koperski, K.; Adhikary, J. and Han, J. 1996. "Spatial Data Mining: Progress and Challenges." In Proceedings Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada.

[6]. Ng RT , Han J (1994) Efficient and Effective Clustering Methods for Spatial Data Mining. In Proceedings of the 20th VLDB Conference Chile.

[7]. PR Dohen Acyyer, M Wang, An Open Agent Architecture. AAAI Spring Symposium, pp1-8, March 1998

[8]. Vincent Ng, Stephen CF chan, Sandra Au, "Web agents for Spatial Mining on Air Pollution".

