International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(1), pp.545-550
DOI: http://dx.doi.org/10.21172/1.81.071
e-ISSN:2278-621X

A REVIEW ON SPATIAL DATA AND SPATIAL HADOOP

Kirandeep Kaur[1], Sukhjit Singh Sehra[2] and Priyanka Arora[3]

Abstract- Spatial data mining is the process to mine data from spatial data sets like OpenStreetMap or any kind of geographic data. With the emergence of time, Geographic Information get more value from the researchers. This is used to mine useful data and features based on different locations. In related work, existing techniques are based on Dijkstras algorithm and Euclidean distance for spatial data mining. But this is not effective way for spatial data. This data is manipulated by mobile phones, road networks and remote sensing data. In this paper we compare all these techniques to process large data sets of OpenStreetMap.

Index Terms—Spatial Data Mining, SpatialHadoop, OpenStreetMap Data.

I. INTRODUCTION

This paper provides an view of different technologies in spatial data mining. The research is based on how to handle spatial data using these techniques.

II. SPATIAL DATA MINING

Spatial data is freely available richer in production and geographic information. Spatial Data Mining (SDM) is the way to mine new information from large data sets like OpenStreetMap. This technique is based on GPS (Global Positioning System) services and LBS (Location Based Service). Geographic context make aware of Spatial Data Analysis and Geo-Sensor Data. Colocation pattern mining is an important concept in SDM. It is based on spatial proximity include set of spatial features. Spatial data is used as main network of Geographic Information Science under the category of UGC (User Generated Content). Clustering tendency can be dignified by joining points and lines to generate network based model.

A. Spatial Data and Big Data

Mainly Data consist of two types Spatial and non-Spatial. The data is available for mining processes like crowdsourced is known as spatial data. Spatial Data Mining is used by Earth observing system, Census Bureau, dept. of Commerce etc. SDM technique is designed to work with organized relationships to discover patterns from large set of data. The data which is correlated to business or market is called Big Data. Big data is a term for mixture of data when its volume, variability and velocity create challenge in technologies or tools like relational databases. The main difference between classical data mining and spatial data mining is classical and statistical approach.

III. SPATIAL DATA MINING TOOLS

Spatial Data Mining is based on databases which can be presented in many ways. SDM is divided into two types

[1] *MTech Student, GNDEC*
[2] *Assistant Professor, GNDEC*
[3] *Assistant Professor, GNDEC*

Descriptive and Predictive. They have further categorized into:

1) **Clustering and Detection:** Clustering is process of grouping objects. It is used in machine learning, pattern recognition, image processing, computer graphics and analysis. It can be based on density or distribution and uses k-means clustering. This is effective way for pattern recognition and data analysis. It can be done in four ways:
   - Hierarchical Method
   - Partitioning Method
   - Density Based
   - Grid Based
2) **Classification:** It is basically use to identifying the set of objects or categorized them in specific order. Classification algorithms are used to build the classifiers. It involves data cleaning to remove noise and treatment of missing values.
3) **Co-location and Association:** Association rule mining and co-location pattern mining are proposed to find interesting and useful relations or patterns from large data sets. This method provides high speed to work with large data sets.
4) **Trend Detection:** In spatial data mining trend detection is the identification of objects and events which are expected to be a pattern. This technique is applicable in fraud detection, event detection in sensor networks and Ecosystem disturbances.

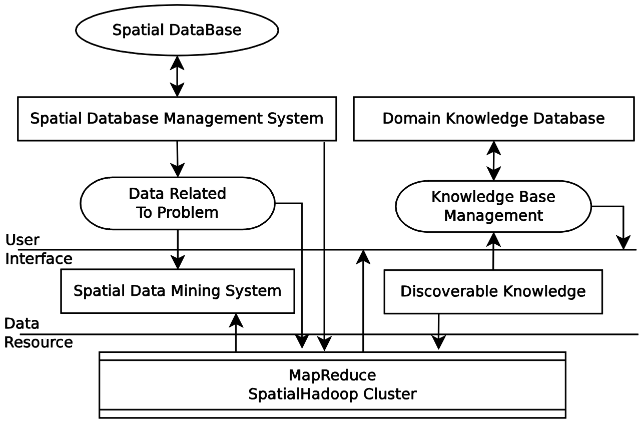
Spatial DataBase
Spatial Database Management System
Domain Knowledge Database
Data Related To Problem
Knowledge Base Management
User Interface
Spatial Data Mining System
Discoverable Knowledge
Data Resource
MapReduce SpatialHadoop Cluster

Fig. 1. Spatial Data Mining Architecture

IV. Spatial Data Mining Tools

Spatial Data Mining can be used with number of tools. Spatial Data Analysis and Modeling (SDAM) software system can execute programs developed in different environments like C, C++, MATLAB with a integrated control and Graphical User Interface (GUI). This technique is liable to execute on a local machines and remotely. Following are the tools used for Spatial Data Mining:

A. Radoop

In 2011 Prekopcsak defined RapidMiner, a tool for analyzing large data sets. This is designed for integration of hadoop. Rapidminer is highly efficient to use in hadoop functionalities. Using Radoop data is stored in Hive's distributed file system. The operations work with every row in dataset and provide results.

B. Hadoop-GIS

This is developed in 2013, a open source platform. It is specially designed for Geographic Information System. It can be used for spatial and non-spatial queries [1]. It also incorporates Hive to boost declarative spatial inquiries. Hadoop-GIS work on RESQUE (Real Time Spatial Query Engine). Hive consists three steps for query processing

- Query translation
- Logical plans
- Physical plans.

C. SpatialHadoop

SpatialHadoop is also an open source platform it helps to implement spatial operations. It has better performance for spatial data as compared to hadoop. Cluster of SpatialHadoop is similar to Google File System (GFS). It provides high throughput and best suitable for large applications. SpatialHadoop is built in Hadoop for spatial constructs inside the core of Hadoop. It acts as powerful backend for system applications like MNTG: is based on real road network and traffic, TAREEG: is used to extract crowdsourced data from OpenStreetMap and SHAHED: is a tool for analyzing and processing remote sensing data.

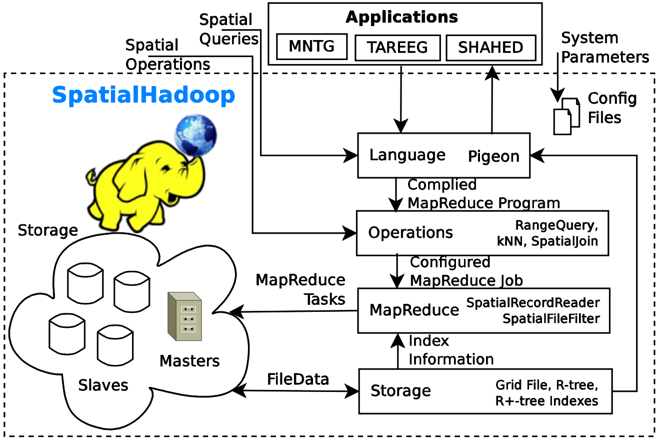

Fig. 2. Architecture of SpatialHadoop

SpatialHadoop is based on a computational prototype called MapReduce follows map and reduce functions. Computation operations can be divided into fragments and its distribution among the cluster nodes is done in map phase. A reduce phase merges all given values related to the same key. SpatialHadoop has four main components.

1) Spatial hadoop does not have its own language. Pigeon extension is added to Pig language in SpatialHadoop. This language is easy for non-technical users and those who are familiar to use PostGIS tools. Pig provides support to process large amount data sets. Pig is SQL language suitable for data analysis. It has provision for user defined functions and optimizations.
2) Second component is Storage layer includes three spatial indexes, Grid File, R-tree and R++tree. These Indexes are organized in two-layers, one is global index and other is multiple local indexes.
3) Third MapReduce layer has two segments SpatialFileSplitter and SpatialRecordReader. MapReduce runs on large cluster of machines and provides powerful interface for parallelization with high speed of performance.
4) Operation layer is the last and fourth one which encapsulates the spatial operations using SpatialHadoop. It has three standard operations- range query, kNN and spatial join.

V. RELATED WORK

Aji et al. [1] described that HadoopGIS is integrated with Hive, which is a data warehousing system for spatial queries on hadoop. Flexibility is supported by this system in map reduce mode and also provides spatial join with mapreduce. This is able to work on geospatial data and location based services. In this work, two data sets are taken one from OpenStreetMap (OSM) and other is Pathology Imaging (PI). This has been used for visualization of 2D and 3D images. Main drawback is the lack of slow data loading, limited support of spatial queries and high cost.

Eldawy et al. [3] defined that HadoopViz used to visualizing spatial data on heat maps. It is an efficient tool for extensible interface provides high image resolution in gigapixel and smoothing functionality for distributed environment. It also provides plotting and partitioning for abstract functions. This has been tested on many networks like Roads Network, Scatter Plot, Frequency Heat Map, Satellite Data and Vectorized Map.

Eldawy and Mokbel [5] included MapReduce systems like Hadoop-GIS, ESRI Tools and SpatialHadoop; Parallel DB systems such as Parallel Secondo; MD-HBase and GeoMesa; and Resilient Distributed Datasets (RDD) like SpatialSpark and GeoTrellis; Skyline operations. It shows that how each feature of extraction is implemented in different systems to make it easy to choose the most suitable approach based on the system architecture. In case of big non-spatial data distributed systems are emerging, they will require more work to extend those systems to support spatial data.

Eldawy [7] described that Spatial data is increasing in large amount day by day. Spatialhadoop and Hadoop become common tool for analyze big spatial data. There is high level language like Pig Latin which support spatialhadoop. Basically these types of languages are designed for non-spatial data and have no support for spatial data types and its functions. So Pigeon is a spatial extension added to Pig for spatial functionality and spatial operations [9]. Pigeon can be implemented with User Defined Functions (UDFs) and this is compatible with all versions of Pig. It also uses spatial operations like FILTER, JOIN and GROUPBY. Open Geospatial Consortium is also compatible with Pigeon like PostGIS. Pigeon is processed to find interesting patterns from given spatial data with lesser code.

TABLE - COMPARISON OF SPATIAL DATA MINING TECHNIQUES

Spatial Data Mining Techniques			
Author	Technology	Data Set	Conclusion
Yu [2]	C++ with database	OpenStreetMap Data	This can provide better performance by changing parameters of algorithm.
Eldawy et al. [3]	Apache Hadoop	OpenStreetMap Data	This method provides high efficiency and it can be improved by changing parameters. It has large use in Multilevel Visualization and heat maps.
Zielstra et al. [4]	HadoopGIS with Hive, Spatial Queries in C++	Pathology Imaging, OpenStreetMap	Pathology Imaging can be used for 3D analytical for detection of tissues.
Eldawy and Mokbel [5]	Real Time Application SHAHED,TAREEG and EarthDB	OpenStreetMap	The system can further worked by visualization component
Daneshyar [6]	Hadoop & MapRedduce	Data from Amazon Web Services	The efficiency can be produced by using spatialhadoop in mapreduce mode with cloud computing.
Eldawy [7]	SpatialHadoop and Hadoop, Pigeon Extension to Pig Language	OpenStreetMap Data	This provides high efficiency for mapreduce function to mine data.
Eldawy et al. [8]	SpatialHadoop	OpenStreetMap Data	Provide extraction of spatial features from crowdsourced data.

Daneshyar [6] explained that Big Data is a large set of data which cannot be handled by traditional systems. This can be intuitively managed by MapReduce framework with large cluster of machines in cloud computing. Big data can be of three types in structured, semi-structured and unstructured form. The architecture of Hadoop is used to process unstructured and semi-structured data with help of mapreduce functions to locate data items and respond the spatial inquiries. It has more chances and challenges in Big Data for business purposes.

Yu [2] explained that how road transports increase the road networks at every place. This needs to capture co-location patterns with help of network distance. Example like in mobile applications, clients usually consider services to the co-location patterns of urban facilities. In this manner these patterns need to be measured along road network with distance due to driven vehicles have not access on predefined roads to move on under some conditions. The computed patterns of facilities which set the limit of movements into account can help service providers to provide more attractive location- sensitive recommendations [10].

Eldawy and Mokbel [11] explained that Pig is member of Hadoop ecosystem act as a procedural version of SQL.

Pig is very simple and high level language. But it doesn't require Java professionals for execution. This is compatible to user and flexible in its syntax and variable allocation. It adds extensibility for the framework using user-defined functions. It can process for unstructured data to process data in parallel form. Pig latin has very less code as compare to other map reduce programs. It can easily process large data sets like OSM. This language is similar to the scripting languages like Perl. In spatial hadoop it uses pigeon extension for implementation of data processing.

Sengstock et al. [12] illustrated that Co-location pattern mining focuses on different algorithms to mine patterns like measure thresholds, e.g,[13]. The techniques used to identify patterns by different properties and parameters of measures such as index. A pattern is superset considered for the frequent pattern. This includes the identification of co-locations having rare events and lead to rules with high confidence.

Eldawy et al. [8] has explained different services for mobile PDAs to use with locating devices like GPS. These services are helpful to locate geographic area to find nearest hotels, parking slots etc. Service providers based on locations find requested services that located in spatial proximity. This improves the effectiveness of location based systems where a user can request a service to its nearby location [14]. Using co-location patterns based on location services may also enable the prefetching to deliver the service fast. In ecology, scientists are curious to discover regular co-occurrences within spatial features as drought, substantial increase or drop in vegetation and extremely high precipitation.

VI. PATTERN DISTRIBUTION IN SDM

Spatial pattern is a set of spatial features. They can be defined depending upon spatial neighborhood which can be checked by interval space. These are useful to upgrade facilities in urban or non-urban areas [2]. This is an important concept using spatial association rule mining, also provides spatial join for co-location patterns.

Spatial patterns can be measured in small or large eras [12]. It is essentially a distribution which characterizes spatial features. This is rule based association data mining defining their spatial characteristics on OpenStreetMap data. This is based on support and entropy using the high resolution grid.

VII. APPLICATIONS OF SPATIAL DATA MINING

Spatial Data Mining is same as data mining with objective to extract special information from large type of data sets in geographic area, remote sensing area, traffic data and mobile generated data.

1) **Trend Detection in GIS** It contains one or more non-spatial attributes to start object from neighborhood paths to the movement and then perform a regression analysis on the respective of attributes for the objects with their neighborhood path to define the precision of change.
2) **Spatial Clustering** Spatial Clustering identifies the location parameters with a number of objects correlated with geographical regions. Location of objects can be assigned by the given methods obtaining an address that specifies the ZIP code, city, state and country.

VIII. CONCLUSION

In this paper we have compared different types of spatial data mining techniques. Techniques employed earlier are changing in the era of spatial and geospatial data mining is rising every day, as the need to extract frequent spatial patterns for OpenStreetMap is today's demand. This can be concluded in supervised and unsupervised manner. Clustering and Co- location pattern mining performs better with SpatialHadoop. Also, SpatialHadoop with Pigeon provides higher efficiency to analyze the current facts for prediction of future events.

REFERENCES

[1] A. Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, and J. Saltz, "Hadoop GIS: a high performance spatial data warehousing system over mapreduce," *Proceedings of the VLDB Endowment*, vol. 6, no. 11, pp. 1009–1020, 2013.

[2] W. Yu, "Spatial co-location pattern mining for location- based services in road networks," *Expert Systems with Applications*, vol. 46, pp. 324–335, Mar. 2016.

[3] A. Eldawy, M. Mokbel, and C. Jonathan, "HadoopViz: A MapReduce Framework for Extensible Visualization of Big Spatial Data," in *IEEE Intl. Conf. on Data Engineering (ICDE),* 2016.

[4] D. Zielstra, H. Hochmair, P. Neis, and F. Tonini, "Areal Delineation of Home Regions from Contribution and Editing Patterns in OpenStreetMap," *ISPRS International Journal of Geo-Information*, vol. 3, no. 4, pp. 1211–1233, Nov. 2014.

[5] A. Eldawy and M. F. Mokbel, "The era of big spatial data," in *Data Engineering Workshops (ICDEW), 2015 31st IEEE International Conference on*. IEEE, 2015, pp. 42–49.

[6] S. Daneshyar, "Large-Scale Data Processing Using MapReduce in Cloud Computing Environment," *Interna- tional Journal on Web Service Computing*, vol. 3, no. 4, pp. 1–13, Dec. 2012.

[7] A. Eldawy, "SpatialHadoop: towards flexible and scalable spatial processing using mapreduce." ACM Press, 2014, pp. 46–50.

[8] A. Eldawy, L. Alarabi, and M. F. Mokbel, "Spatial partitioning techniques in SpatialHadoop," *Proceedings of the VLDB Endowment*, vol. 8,

no. 12, pp. 1602–1605, 2015.

[9] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, “Pig latin: a not-so-foreign language for data processing,” in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 1099–1110.

[10] Y. Zhang, X. Li, A. Wang, T. Bao, and S. Tian, “Density and diversity of OpenStreetMap road networks in China,” *Journal of Urban Management*, vol. 4, no. 2, pp. 135– 146, Dec. 2015.

[11] A. Eldawy and M. F. Mokbel, “Pigeon: A spatial MapRe- duce language,” in *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*. IEEE, 2014, pp. 1242–1245.

[12] C. Sengstock, M. Gertz, and T. Van Canh, “Spatial interestingness measures for co-location pattern mining,” in *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 821–826.

[13] Y. Huang, S. Shekhar, and H. Xiong, “Discovering colocation patterns from spatial data sets: a general approach,” vol. 16, no. 12, pp. 1472–1485.

[14] L. Alarabi, A. Eldawy, R. Alghamdi, and M. F. Mokbel, “TAREEG: a MapReduce-based web service for extract- ing spatial data from OpenStreetMap.” ACM Press, pp. 897–900.