Modelling and Designing Spatial and Temporal Big Data for Analytics

Sinan Keskin[1(✉)] and Adnan Yazıcı[2(✉)]

[1] Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
keskin.sinan@metu.edu.tr
[2] School of Science and Technology, Nazarbayev University, Astana, Republic of Kazakhstan
adnan.yazici@nu.edu.k2

Abstract. The main purpose of this paper is to introduce a new approach with a new data model and architecture that supports spatial and temporal data analytics for meteorological big data applications. The architecture is designed with the recent advances in the field of spatial data warehousing (SDW) and spatial and temporal big data analytics. Measured meteorological data is stored in a big database (NoSQL database) and analyzed using Hadoop big data environment. SDW provides a structured approach for manipulating, analyzing and visualizing the huge volume of data. Therefore, the main focus of our study is to design a Spatial OLAP-based system to visualize the results of big data analytics for daily measured meteorological data by using the characteristic features of Spatial Online Analytical Processing (SOLAP), SDW, and the big data environment (Apache Hadoop). In this study we use daily collected real meteorological data from various stations distributed over the regions. Thus, we enable to do spatial and temporal data analytics by employing spatial data-mining tasks including spatial classification and prediction, spatial association rule mining, and spatial cluster analysis. Furthermore, a fuzzy logic extension for data analytics is injected to the big data environment.

Keywords: Meteorological big data analytics · DWH · SOLAP
Hadoop

1 Introduction

Data mining is the field of discovering novel and potentially useful information from large amounts of data [1]. Geospatial data mining is a sub-field of data mining that employs specialized techniques for dealing with geospatial data. There are two types of data mining tasks: (a) descriptive data mining tasks that describe the general properties of the existing data and (b) predictive data mining tasks that attempt to do predictions based on inference on available data. Predictive data mining tasks come up with a model from the available data set helpful in predicting unknown or future values of another data set of interest. Descriptive data mining tasks usually finds data describing patterns and comes up with new information or pattern from the available data set.

T. Czachórski et al. (Eds.): ISCIS 2018, CCIS 935, pp. 104–112, 2018.
https://doi.org/10.1007/978-3-030-00840-6_12

In this study, we design a spatial data warehousing (SDW) and do analytics on spatial for temporal big data. Daily measured meteorological data is stored in a NoSQL database and SDW that provides a structured approach for manipulating, analyzing and visualizing the huge volume of data. The Spatial OLAP-based system (SOLAP) is used to visualize the results of big data analytics using the big data environment (Apache Hadoop). The system learns the general features of the given data with descriptive data mining tasks and make predictions for the future with predictive data mining.

We propose a new model by designing SOLAP to visualize the trends for daily measured historical meteorological data. The daily meteorological data of different stations distributed over the regions is used as a case study. Also, recent theoretical and applied research in spatial data mining are studied. We also introduce fuzzy logic extension to our proposed system. Thus, fuzzy spatial temporal querying is supported by the system.

In this paper, the details of the study are described in the following sections, the related work on the topic is discussed in Sect. 2. The composite environment and the components of the architectural structure of the model (the proposed architecture) are briefly discussed in Sect. 3. A case study on the proposed architecture is given with examples in Sect. 4. In Sect. 5, conclusion and future works are given.

2 Related Works

Most of the meteorological data based prediction techniques and methods are based on statistical or widely used data mining techniques like clustering, classification, regression analysis, decision tree etc. Some of the related work is as follows:

Liang et al. [2] derived the sequence of ecological events using temporal association rule mining. Red tide phenomena occurred during 1991 and 1992 in Dapeng bay, South China Sea was taken as an example to validate T-Apriori algorithm which generated frequent itemsets and corresponding temporal association rules and K-means clustering analysis are used to map the quantitative association rule problem into the Boolean association rules.

Huang et al. [3] analyzed historic salinity-temperature data to make predictions about future variations in the ocean salinity and temperature relations in the water surrounding Taiwan. They use inter-dimensional association rules mining with fuzzy inference to discover salinity-temperature patterns with spatial-temporal relationships.

The other authors Kotsiantis et al. [4] proposed a hybrid data mining technique that can be used to predict the mean daily temperature values. Several experiments have been conducted with well-known regression algorithms using temperature data from the city of Patras in Greece. The methods used in their study needs be still validated by including temperature data with other meteorological parameters.

Kohail et al. [5] tried to extract useful knowledge from weather daily historical data collected locally at Gaza Strip city. The data include nine years period [1977–1985]. After data preprocessing, they apply basic algorithms of clustering, classification and association rules mining techniques. For each data mining technique, they present the extracted knowledge and describe its importance in meteorological field, which can be used to obtain useful prediction and support the decision making.

Sivaramakrishnan et al. [6] presented the method for prediction of daily rainfall. Meteorological data from 1961–2010 are used in their analysis. For the atmospheric parameters temperature, dew point, wind speed, visibility and precipitation (rainfall) are considered. They filter and discretize the raw data based on the best fit ranges and applied association mining on dataset using Apriori algorithm to find the hidden relationship between various atmospheric parameters to predict the rainfall.

3 Proposed Architecture

We designed a composite system to provide spatial and temporal data mining and analytics. This system consists of four layers structure. We can define the system from bottom to top as data sources, structured data, logic, and presentation layers. Multi-layer system architecture is represented in Fig. 1.

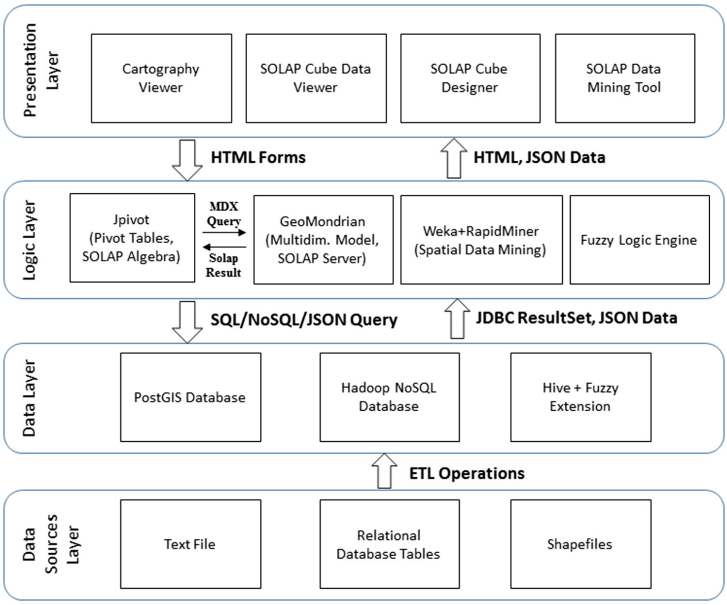

Fig. 1. Multi-layer composite system architecture.

At the bottom of the system we have text files, database tables and shape files that contain the pure data gathered from the meteorology service. Data in this layer is migrated to the structured data layer via Extract Transform and Load (ETL) operations.

Data layer is about semi-structured or structured data like relational database, Hive meta-store or Hadoop file system data nodes. Data in this layer is created by ETL operations. The upper logic layer requests data from data layer by using SQL, HiveQL or JSON request. Data layer returns the requested data via SQL tuples, JDBC result set

or JSON response. Data layer also provide fuzzy querying on Hive which supported User Defined Functions to contribute the system.

Logic layer contains integrated systems which provide spatial, non-spatial, temporal and fuzzy data mining tools and function sets. It also contains data analytics and geovisualization platforms that helps visually pattern detection. Another integrated part is reporting tools which provides common reports on data. SOLAP server is another main part of this layer that provides SOLAP data cube operations and MDX querying. Weka [7], RapidMiner [8] and ArcMap [9] tools are integrated for spatial data mining. Fuzzy logic engine is integrated with the system for fuzzy operations such as membership calculation, fuzzy clustering and fuzzy class identification.

At the top of the proposed architecture we have representation layer which provides using all the system categorized and simplified structure. For instance, we can demonstrate the data in map with cartography viewer. The composite system architecture mainly consists of three environments and is explained in following sub sections.

3.1 PostGIS Environment

One part of the system is PostGIS [10] that provides spatial objects for the PostgreSQL database, allowing storage and query of information about location and mapping. ETL operations are handled on the text-based data. Also, we design the spatial hierarchy of the inserted data in database such as stations belongs to cities and cities belongs to regions. In the real data we have only station data that contains latitude and longitude values, but we do not have city and regions information. The city and region information were collected, transformed into polygon form and inserted to database tables. After this operation spatial queries can be done on hierarchical data.

3.2 SOLAP Environment

Designed SOLAP cube contains two hierarchies, one of them is temporal hierarchy and the other one is spatial hierarchy. Temporal hierarchy consists of year-month-day values for each measurement record. Spatial hierarchy is about region-city-station values. In addition to hierarchies we have ten measurements in our spatial OLAP cube as shown in Fig. 2.

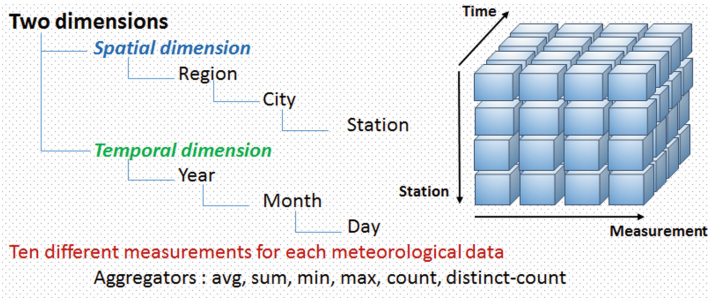

Fig. 2. Dimensions of the spatial OLAP cube.

After SOLAP cube is designed in Workbench [11] which provides SOLAP cube design, it is used as meta-data and feed the cube with data which was inserted in PostGIS database. The region-city-station spatial hierarchy and individual measurement data tables are used in cube definition. After providing SOLAP cube to PostGIS connection, MDX querying can be done and retrieved the SOLAP query result in Mondrian [12] system which is the main SOLAP server. System also provides geovisualization that show the query result in map after the integration of Spatialytics [13] which is a geovisualization tool. This tool also connected with the designed PostGIS database. Flow of the whole system including geovisualization is shown in Fig. 3.

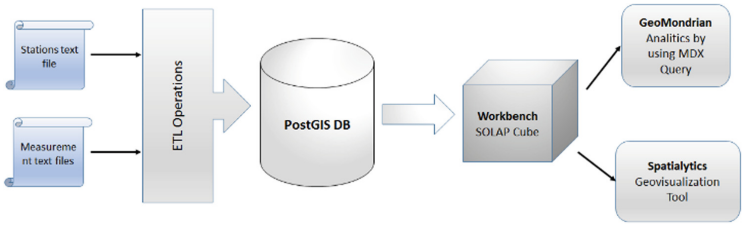

Fig. 3. SOLAP environment.

3.3 Big Data Environment

In big data environment, Apache Hadoop [14], which is an open-source software framework used for distributed storage and processing of dataset of big data is used. Another Apache product Hive [15] is selected to use above the Hadoop. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data and makes querying and analyzing easy.

The text base data is loaded to Hadoop HDFS by using Hive interface. Hadoop data nodes allocate the data in clusters. After the data is loaded to HDFS, we can do HQL queries on Hive for data analytics. After we load the data, apply analytics queries and fetch the result, we transform the result set into shape file. Created shape file can be load to geovisualization tool as map feature. Visual analyses can be done by importing features in to map. End to end data transformation is represented in Fig. 4.

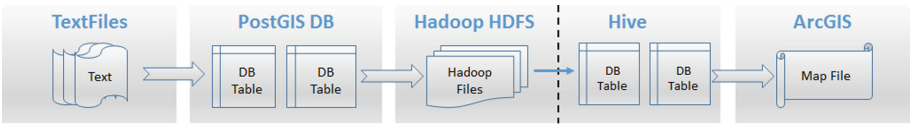

Fig. 4. End to end data transformation in big data.

Fuzzy Extension. One of the contribution of the study is fuzzy querying extension on big data environment. As mentioned that Hive supports UDFs which provides custom built function for querying. To achieve this, FuzzyValue UDF is implemented and

takes column value as parameter and return the fuzzy class name of the value. In this operation firstly, Fuzzy C-Means (FCM) [16, 17] is applied for clustering on whole data and determines the membership values. Each membership value according to their clusters is stored. For example, FCM is applied on temperature values and 3 clusters are determined such as cold, normal and hot. Temperature value 7.2 has membership like 0.1 cold, 0.8 normal and 0.1 hot. When we want to query hot days in a city, we can look up for the data which has hot membership is greater than 0.5. After we execute our fuzzy query, fuzzy spatial query or fuzzy spatial temporal query, we can fetch the result and transform them into shape file then load it to the map to do visual analytics. Fuzzy extension part of the study can be viewed as in Fig. 5.

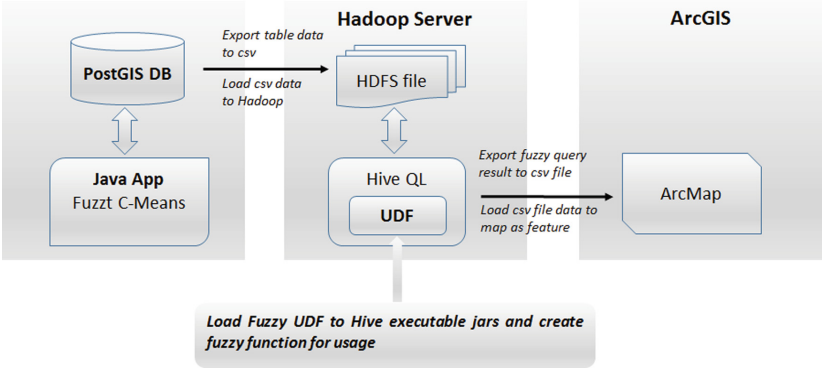

Fig. 5. Fuzzy extension part of study.

4 Case Study: Spatial Data Mining Tasks on Meteorological Big Data

In this section composed system is used with the following spatial data mining tasks that contains unsupervised classification, frequent item set mining, association rule learning and pattern discovery.

In this study, we have text files containing the results of 10 meteorological measurement types. The meteorological measurement types including the measurements of stations are as follows: daily vapor pressure, daily hours of sunshine, daily max speed and direction of the wind, daily average actual pressure, daily average cloudiness, daily average relative humidity, daily average speed of the wind, daily average temperature, daily total rainfall - manual and daily total rainfall - omgi.

These files contain daily measurements between 01.01.1970 to 01.01.2017. For each file there are records that are about station number, measurement type, and measurement date and measurement value data. Sample data in daily average speed of wind (m/s) is given in Table 1.

Table 1. Sample data of daily average wind speed file.

Station no	Station name	Year	Month	Day	Daily average speed of wind
8541	HASSA	1977	1	1	1.3
8541	HASSA	1977	1	2	1.1
8541	HASSA	1977	1	3	3.1
8541	HASSA	1977	1	4	3.4
8541	HASSA	1977	1	1	1.3

Firstly, we have crisp values of measurements and need to determine fuzzy membership values for each measurement. Before this operation we need to clarify uncertainty in the contents of measurement data by using interpretation motivated by the need to represent null or unknown values for certain data items, data entry mistakes, measurement errors in data, "don't care" values [18]. Then unsupervised classification based on FCM is applied over the measurement data to determine fuzzy classes and fuzzy membership values for each data. Applying FCM over 15 M data is extremely needs high computational resources in classical approaches. For this reason, adapted FCM on distributed environment is used to overcome resource limitation. Therefore, fuzzy classes as high, normal, low and membership values for temperature, rainfall and humidity is determined by using distributed FCM algorithm.

In the next step apriori is applied over each segmented data to determine the number of frequent item sets to support association rule learning. In executed apriori algorithm our item set is searching the values of high temperature, low rainfall and low humidity. At the end of execution, we have the support values for each winter partition.

Calculated support values are used for pattern discovery. Here we find meaningful pattern on these values. Using these patterns, it is possible to make prediction about the future meteorological events. Pattern discovery is studied on the data as Fig. 6.

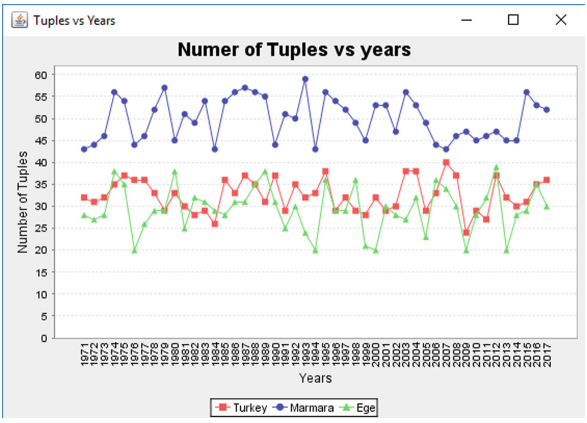

Fig. 6. The result of frequent item set between 1970 to 2017.

Geospatial data mining is employed on real data set, meteorological measurement data, of Turkey. 1161 different stations and 10 different measurement data types are used in the scope of this study. Each stations are chosen from different geographic regions of Turkey and used on text base data. Then the structured form of data set is built from that.

5 Conclusion and Future Works

The prosed approach introduces several common spatial data-mining tasks, including spatial classification, spatial association rule mining, spatial cluster analysis, frequent item set mining, pattern discovery and prediction. After using spatial data mining techniques to analyze the historical spatial meteorological data, make prediction about future meteorological measurement over Turkey climate by using geospatial predictive modelling is the primary target of this study. There are a number of possible future studies, such as developing a more efficient predictive model for meteorological data applications.

References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann Publishers, USA (2012)
2. Liang, Z., Xinming, T., Wenliang, J.: Temporal association rule mining based on t-Apriori algorithm and its typical application. In: International Symposium on Spatial-Temporal Modeling Analysis, vol. 5, issue 2 (2005)
3. Huang, Y.P., Kao, L.J., Sandnes, F.E.: Predicting ocean salinity and temperature variations using data mining and fuzzy inference. Int. J. Fuzzy Syst. **9**(3), 143–151 (2007)
4. Kotsiantis, S., Kostoulas, A., Lykoudis, S., Argiriou, A., Menagias, K.: A hybrid data mining technique for estimating mean daily temperature values. IJICT **1**(5), 54–59 (2007)
5. Kohail, S.N., El-Halees, A.M.: Implementation of data mining techniques for meteorological data analysis. Int. J. Inf. Commun. Technol. Res. (JICT) **1**(3) (2011)
6. Sivaramakrishnan, T.R., Meganathan, S.: Association rule mining and classifier approach for quantitative spot rainfall prediction. J. Theor. Appl. Inf. Technol. **34**(2), 173–177 (2011)
7. Weka is a collection of machine learning algorithms for data mining tasks. https://www.cs.waikato.ac.nz/ml/weka/
8. RapidMiner is a software platform for data science teams that unites data prep, machine learning, and predictive model deployment. https://rapidminer.com
9. ArcMap is the main component of Esri's ArcGIS suite of geospatial processing programs. http://desktop.arcgis.com/en/arcmap/
10. PostGIS is a spatial database extender for PostgreSQL object-relational database. It adds support for geographic objects allowing location queries to be run in SQL. https://postgis.net/
11. Mondrian Schema Workbench is a designer interface that creates and tests Mondrian OLAP cube schemas visually. https://mondrian.pentaho.com/documentation/workbench.php
12. GeoMondrian is an open source Spatial OnLine Analytical Processing (Spatial OLAP or SOLAP) server, a spatially-enabled version of Pentaho Analysis Services. http://www.spatialytics.org/blog/geomondrian-1-0-is-available-for-download/

13. Geovisualization tool for spatial data. http://www.spatialytics.org/
14. The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. http://hadoop.apache.org/
15. The Apache Hive data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. https://hive.apache.org/
16. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. **3,** 32–57 (1973)
17. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York (1981)
18. Gelenbe, E., Hebrail, G.: A probability model of uncertainty in data bases. In: ICDE, pp. 328–333 (1986)