SMART: A Service-Oriented Statistical Analysis Framework on Spatio-Temporal Big Data (Short Paper)

Jie Zhou[1,3](✉), Weilong Ding[1,2], Zhuofeng Zhao[1,3], and Han Li[1,2]

[1] Data Engineering Institute,
North China University of Technology, Beijing 100144, China
136139154@qq.com
[2] Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, Beijing 100144, China
[3] Beijing Urban Governance Research Center, Beijing 100144, China

Abstract. Spatio-temporal data is one of the most important assets in the context of smart cities. Spatio-temporal big data comes from a variety of sensor devices, implies the state of urban operation, insight into the development trend. Due to the multidimensional characteristics and diverse analysis needs of spatial-temporal data, data analysis based on spatial-temporal data must take into account the large capacity, diversity and frequent changes of data. This makes spatial and temporal data analysis more difficult. In order to simplify the analysis of spatio-temporal data, a service-oriented intelligent framework is proposed. Firstly, the concept of spatio-temporal data service is introduced into the framework, and several common spatio-temporal data service models are defined. Then, a configurable scripting language was proposed to define the analytic application. We also developed a prototype tool to implement spatio-temporal data services on Hadoop. In order to prove the applicability of our method, we demonstrate the effectiveness of our work through a practical application-based study.

Keywords: Spatio-temporal data · Service composition · Configurable

1 Introduction

In nowadays, various sensors are adopted in modern cities [1], such as recognition cameras on the trunk roads, smart-card readers in the buses, GPS equipped devices in taxis, RFID tags embedded on freights, inductive loops at the toll stations, and transducer in the power plants. The accumulated sensory data with attributes of space and time [2, 3] can reflect the urban rhythm.

Spatio-temporal data is a multidimensional continuum and always accumulated as big data. It is crucial to understand the dependencies across time and space [4] during the analysis. Considering their large amount and high velocity, the data analyses are intrinsically challenging. (1) First, traditionally it is a long cycle to describe requirements, complete programming, and plot the results. To balance the programming availability and analysis complexity is not trivial. It is urgent to find ways to depict

X. Wang et al. (Eds.): CollaborateCom 2019, LNICST 292, pp. 91–100, 2019.
https://doi.org/10.1007/978-3-030-30146-0_7

requirements easily in domain specific manner. (2) Second, for common descriptive statistical analysis, it is inconvenient to configure multiple steps of preprocessing, statistical processing and visualization. Each step would contain various configurable parameters. (3) Third, for such a statistical analysis application, collaborate multiple steps in a complete and rigorous manner is inefficient. In current solutions, only limited capabilities (e.g., processing through Hadoop MapReduce and storage on NoSQL) are supported, but the association with preprocessing or visualized is ignored.

In this paper, SMART is presented for typical descriptive statistics with the following contributions. (1) In view of the process of spatio-temporal data preprocessing, descriptive statistic and visualization, this paper summarizes the extraction of different types of service requirements, and on this basis, the corresponding service model is designed. (2) A method of implementing spatio-temporal descriptive statistic service in Hadoop environment is proposed. Statistics service program based on big data environment can be realized automatically through configuration. (3) The spatio-temporal data service composition language and the implementation engine are used to constrain and describe the behavior of the spatio-temporal data service composition.

The organizational structure of this article is as follows. Section 2 introduces the related work. Section 3 introduces the system structure of SMART. Section 4 presents the implementation of specific cases.

2 Related Work

Web service is a technology based on standard network protocol, which is an important means to realize the mutual access operation of application services between heterogeneous systems on the Internet. In the implementation of Web services, because the implementation and operation of REST-based Web services are easier and simpler than those based on SOAP and XML-RPC [5], REST has attracted wide attention in the industry since it was proposed. Amazon has also put REST principles into practice. It has implemented RESTFul services with XML as data exchange format [6], as well as social platform FaceBook and Paypal, which provide REST-style Web services.

Current research on service composition is mostly based on service discovery and service composition of the Internet of Things, and few studies are focused on service discovery and modeling under the background of multidimensional analysis of large spatial and temporal data. The literature [7] proposes a four-tier architecture, namely, storage layer, online and historical data processing layer, analysis layer and decision layer. This architectural approach can be used to handle large static data streams as well as large online data streams.

3 Architectural Design and Realization

3.1 Architectural Design

The architecture of SMART is presented as Fig. 1. On the virtualized infrastructure, data analyses as configurable applications would be built as these steps. We will explore and study these three steps in the next work.

(1) In the process of spatio-temporal data analysis, this paper summarizes different types of services for data preprocessing, statistical calculation and visualization. On this basis, the service model is studied. Following the characteristics of spatio-temporal data analysis, the statistical application of spatio-temporal data can be constructed quickly.
(2) A method of spatio-temporal description of statistical service in Hadoop environment is proposed, and the statistics service based on big data environment can be quickly realized through configuration
(3) Declarative configuration languages can describe the multidimensional characteristics of spatio-temporal data. It can also be used to constrain the behavior of services and service composition.

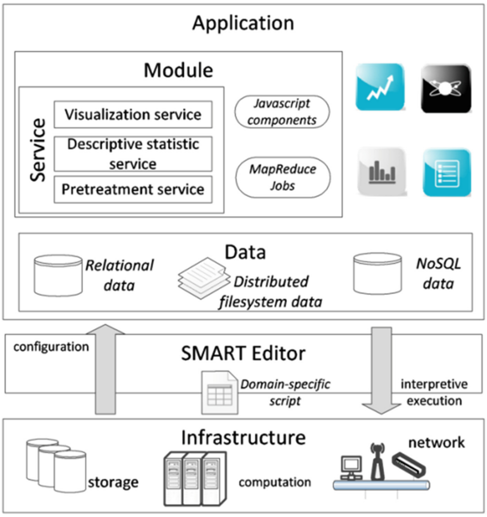


Fig. 1. SMART architecture

3.2 Spatio-Temporal Data Service Recognition and Modeling

According to the characteristics of spatio-temporal large data and multidimensional data, this paper studies the methods of spatio-temporal large data analysis, and preliminarily designs several common services: pretreatment service, descriptive statistic service and visualization service. The model of spatio-temporal data service is designed.

The service model of spatio-temporal data can be expressed as a quintuple <Prefix, Type, Service ID, Parameter, Result>:

Prefix: Universal prefix for basic services, defined here as http://ip:port/hn/jsps/service;

Types: Types of service. Three types are defined here: pretreatment service, descriptive statistic service and visualization service. These three types correspond to preprocess, statistics, visualized.

Service ID: Service ID, the unique service identification generated by a user when creating a service on a Web page.

Parameters: A list of parameters. The different parameters of the services accessed are different. Specific parameters are related to the content of each service. For example, the parameters of pretreatment service are optional processing methods, the parameters of descriptive statistic service are attributes of different dimensions such as space, time and object, and the parameters of visualization service are optional visual graphical effects.

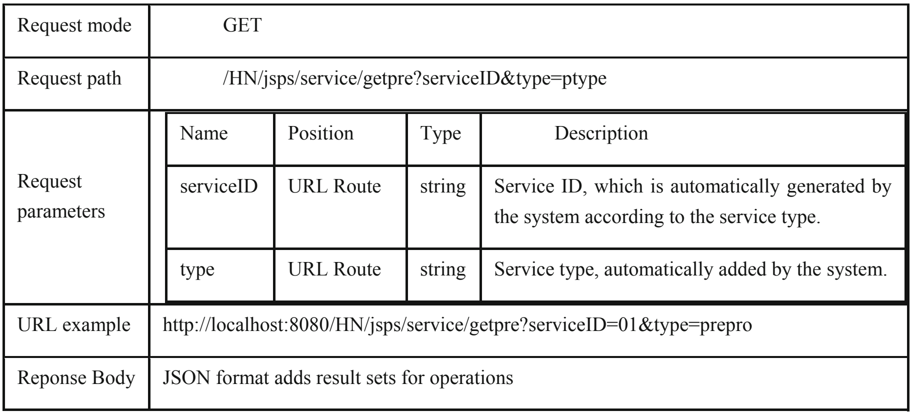
Result: The result returned by the service is displayed in the JSON string. Specific results are designed according to HTTP requests.

3.3 Fast Implementation of Descriptive Statistic Service Based on Big Data Environment

SMART provides RESTful style Web services for data communication and service invocation of spatio-temporal data services.
Pretreatment Service: Pretreatment service extracts some data from massive redundant spatio-temporal data and eliminates erroneous data. Pretreatment service mainly includes three functions. (1) Data integration: According to the needs of multidimensional analysis, extract part of the data from multidimensional space-time data. (2) Data revision based on spatio-temporal correlation: excluding data with cross temporal attributes, invalid spatial attributes and inconsistent spatio-temporal attributes. (3) Data filtering based on business rules: eliminating illegal attributes, invalid null attributes and duplicate redundant data.

The configuration information of the pretreatment service is obtained through the method of selecting preconfigured services by SMART. The preprocessing module obtains the URI of configuration information, executes corresponding preprocessing jobs according to configuration information, and sends the address where the results are stored to SMART. The pretreatment service API design is shown in Table 1 below.

Table 1. Pretreatment service API design



Request mode	GET			
Request path	/HN/jsps/service/getpre?serviceID&type=ptype			
Request parameters	Name	Position	Type	Description
	serviceID	URL Route	string	Service ID, which is automatically generated by the system according to the service type.
	type	URL Route	string	Service type, automatically added by the system.
URL example	http://localhost:8080/HN/jsps/service/getpre?serviceID=01&type=prepro			
Reponse Body	JSON format adds result sets for operations			

Descriptive Statistic Service: The function of descriptive statistic service is to analyze the data obtained by preprocessing and get the statistical results. After the pretreatment service is executed, the MapReduce model of Hadoop platform is used to realize the multidimensional descriptive statistical service. The descriptive statistic service API design is shown in Table 2 below.

Table 2. Descriptive statistic service API design

Request mode	GET			
Request path	/HN/jsps/service/getsta?serviceID&type=stype			
Request parameters	Name	Position	Type	Description
	serviceID	URL Route	string	Service ID, which is automatically generated by the system according to the service type.
	type	URL Route	string	Service type, automatically added by the system.
URL example	http://localhost:8080/HN/jsps/service/getsta?serviceID=11&type=statistics			
Reponse Body	Result Set from Data Statistical Job in JSON format			

The descriptive statistic service obtains the URI of the configuration information of the data statistics jobs configured by SMART, and executes the corresponding data statistics jobs according to the configuration information. The address stored in the statistical results is sent to SMART. Our innovative work is to design a data statistics template based on MapReduce model of Hadoop platform for multidimensional data statistics. When data statistics are needed, there is no need to repeat coding. We only need to get the parameter list of descriptive statistic service and then we can get the data statistics results in different dimensions through the template. The flow chart for MapReduce template execution is shown in Fig. 2.

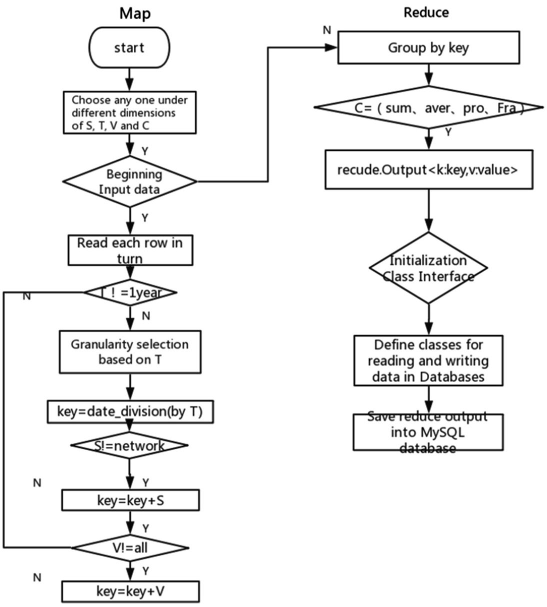

Map
start
Choose any one under different dimensions of S, T, V and C
Y
Beginning Input data
Y
Read each row in turn
T ! =1year
N
N
Granularity selection based on T
key=date_division(by T)
S!=network
N
Y
key=key+S
Y
V!=all
N
Y
key=key+V
Reduce
N
Group by key
C= (sum、aver、pro、Fra)
Y
recude.Output<k:key,v:value>
Initialization Class Interface
Define classes for reading and writing data in Databases
Save reduce output into MySQL database

Fig. 2. MapReduce diagram

Visualization Service: Each visualization service may correspond to multiple data statistics services. As required, we get the results of the descriptive statistic service and select the configurable component style to display the results. The visualization service API design is shown in Table 3 below.

Table 3. Visualization service API design

<table>
<tr><td>Request mode</td><td colspan="4">GET</td></tr>
<tr><td>Request path</td><td colspan="4">/HN/jsps/service/getview?serviceID&type=vtype</td></tr>
<tr><td rowspan="3">Request parameters</td><td>Name</td><td>Position</td><td>Type</td><td>Description</td></tr>
<tr><td>serviceID</td><td>URL Route</td><td>string</td><td>Service ID, which is automatically generated by the system according to the service type.</td></tr>
<tr><td>type</td><td>URL Route</td><td>string</td><td>Service type, automatically added by the system.</td></tr>
<tr><td>URL example</td><td colspan="4">http://localhost:8080/HN/jsps/service/getview?serviceID=21&type=view</td></tr>
<tr><td>Reponse Body</td><td colspan="4">JSON format to query data table fields, field types and other information result sets</td></tr>
</table>

3.4 Configurable Spatio-Temporal Data Service Composition

An application can be built using the services provided by SMART, such as Qingming Festival vehicle travel. The application involves three services: (1) pretreatment service extracts data from three-day traffic data and cleans data, extracts spatio-temporal related data, and eliminates erroneous data that do not conform to spatio-temporal correlation. (2) Descriptive statistic service performs corresponding data statistics jobs according to the dimension information of time, space and object selected by users, and obtains the data statistics results. (3) Visualization service obtains the results of descriptive statistic service and prediction service, and displays them visually.

SMART provides configurable options for three types of services. Users can select the specific information corresponding to the three types of services through SMART. The configuration files for each service are given below:

Configuration file1 Pretreatment service Information

```
<prepro>
 <period>2018-06-16--2018-06-18</period>
 <preproccess>Time attribute out of bounds</preproccess>
 <preproccess>Invalid space attribute</preproccess>
 <preproccess>Inconsistent time and space attributes</preproccess>
 <preproccess>Illegal license plate attribute</preproccess>
 <preproccess>license</preproccess>
 <data>
<url>http://localhost:8080/HN/jsps/service/enc?serviceID=10&type=encapsulate</url>
 </data>
</prepro>
```

Configuration file2 Descriptive statistic service Information

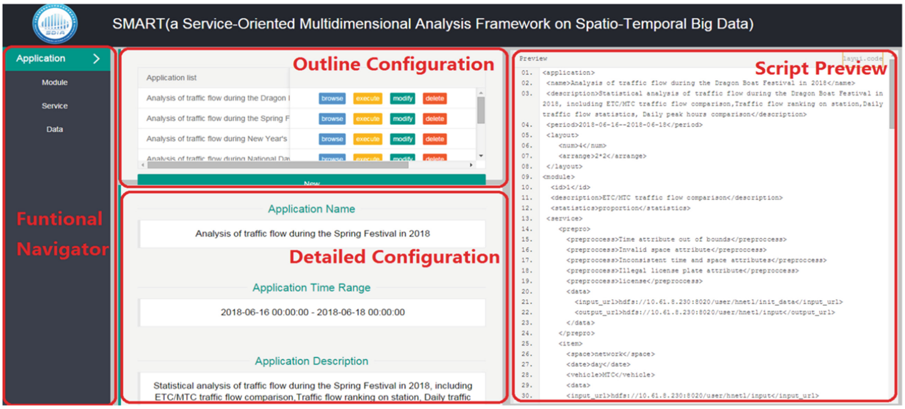
```
<item>
  <ID>d1</ID>
  <space>network</space>
  <date>day</date>
  <vehicle>MTC</vehicle>
  <data>
 <input_url>hdfs://10.61.8.230:8020/user/hnetl/input</input_url>
 <output_url>jdbc:mysql://10.61.4.120:3306/hnfreeway</output_url>
  </data>
 </item>
```

Configuration file3 Visualization service Information

```
<view>
 <form>pie plot</form>
 <serviceID>d1,d2</serviceID>
 <data>
<input_url>jdbc:mysql://10.61.4.120:3306/hnfreeway</input_url>
<tb_structure>tb_PerStaOneDayEM:stationID,date,type,volume</tb_structure>
 </data>
</view>
```

4 Case Study

Any service as a job runs on the infrastructural resources, and it is an instance parameterized from an abstract MapReduce template in our previous work [8]. Taking the mentioned highway domain as an example, an application for the traffic flow analysis during Dragon Boat Festival is presented as Fig. 3. The toll data used is typical spatio-temporal one with attributes vehicle license, entry (exit) timestamp, and entry (exit) station. The toll data of three-day vacation includes four modules: ETC and MTC traffic flow, daily traffic flow statistics, station traffic flow ranking, daily peak hours. Five corresponding MapReduce jobs are instantiated to execute, and the results are displayed as defined configuration scripts.



Analysis of traffic flow during the Dragon I
Analysis of traffic flow during the Spring F
Analysis of traffic flow during New Year's
Analysis of traffic flow during National Day

Application Name

Analysis of traffic flow during the Spring Festival in 2018

Application Time Range

2018-06-16 00:00:00 - 2018-06-18 00:00:00

Application Description

Statistical analysis of traffic flow during the Spring Festival in 2018, including ETC/MTC traffic flow comparison,Traffic flow ranking on station, Daily traffic

```
<application>
 <name>Analysis of traffic flow during the Dragon Boat Festival in 2018</name>
 <description>Statistical analysis of traffic flow during the Dragon Boat Festival in 2018, including ETC/MTC traffic flow comparison,Traffic flow ranking on station,Daily traffic flow statistics, Daily peak hours comparison</description>
 <period>2018-06-16--2018-06-18</period>
 <layout>
    <num>4</num>
    <arrange>2*2</arrange>
 </layout>
<module>
  <id>1</id>
  <description>ETC/MTC traffic flow comparison</description>
  <statistics>proportion</statistics>
 <service>
    <prepro>
      <preproccess>Time attribute out of bounds</preproccess>
      <preproccess>Invalid space attribute</preproccess>
      <preproccess>Inconsistent time and space attributes</preproccess>
      <preproccess>Illegal license plate attribute</preproccess>
      <preproccess>license</preproccess>
      <data>
        <input_url>hdfs://10.61.8.230:8020/user/hnetl/init_data</input_url>
        <output_url>hdfs://10.61.8.230:8020/user/hnetl/input</output_url>
      </data>
    </prepro>
    <item>
      <space>network</space>
      <date>day</date>
      <vehicle>MTC</vehicle>
      <data>
      <input_url>hdfs://10.61.8.230:8020/user/hnetl/input</input_url>
```

Fig. 3. SMART editor.

Next, we introduce the configuration process of the tool. (1) Fill in the basic information of the whole application on the web page, and apply the time range of title, description and statistics. (2) The number and arrangement of configuration modules and the statistical methods of each module. (3) The third step is to configure the services used for each module. Taking ETC and MTC traffic statistics as examples, we first choose the method of pretreatment service. Here we choose according to our needs. Then, in order to compare traffic flow, two traffic descriptive statistic services are needed: the spatial dimension of the whole network, the time dimension of one day and the object type of ETC vehicles and so on. MTC traffic statistics: the spatial dimension of the whole network, the time dimension of one day, the object dimension of MTC vehicles. Finally, visualization service method is selected as pie chart. After configuring the corresponding services of each module in turn, an application configuration file is obtained. (4) Submit a complete application configuration file to the engine for execution, and the results are shown in Fig. 4.

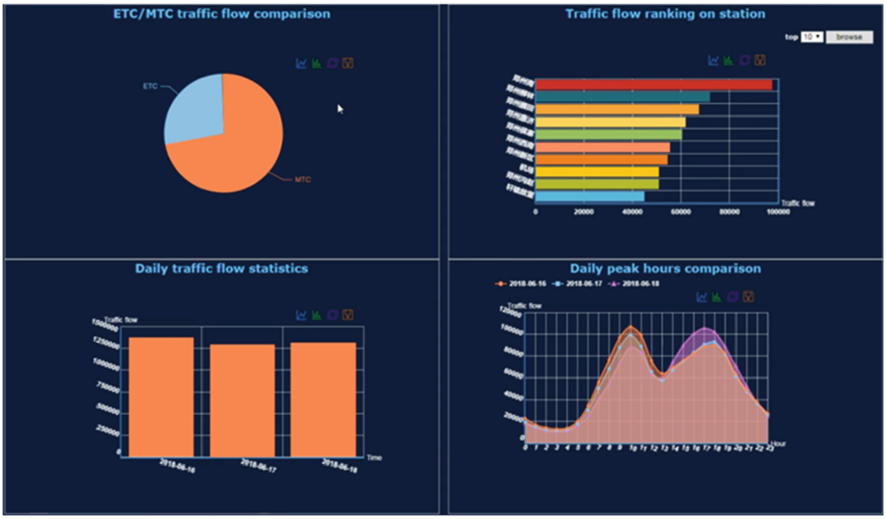


Fig. 4. An example data analysis application.

5 Summary

Three services are provided to support the multidimensional analysis of massive redundant spatial and temporal data. Pretreatment services extract keyword fields, eliminate erroneous data and correct available data. Descriptive statistic service conducts offline data statistics to obtain statistical results. Visualization services display statistical results with configurable composite visualized components.

A declarative language script is designed for describing the related information and service composition of three basic services. The script combines basic services into data analysis module, and data analysis module into data analysis application. The platform provides a series of data analysis, which can show some existing data analysis results, and can also be configured according to the platform to get new data analysis results.

SMART proposes the application of typical descriptive statistical analysis to spatio-temporal data only by configuration. Practice proves that the method is feasible and effective.

Acknowledgments. This work was supported by the Youth Program of National Natural Science Foundation of China (No. 61702014), Beijing Natural Science Foundation (No. 4192020), Youth Innovation Foundation of North China University of Technology (No. XN018022), and the R&D General Program of Beijing Education Commission (Grant No. KM201810009004).

References

1. Wang, S., Xu, J., Zhang, N., Liu, Y.: A survey on service migration in mobile edge computing. IEEE Access **6**, 23511–23528 (2018)
2. Zheng, Y.: Trajectory data mining: an overview. ACM Trans. Intell. Syst. Technol. **6**, 1–41 (2015)
3. Wikle, C.K.: Modern perspectives on statistics for spatio-temporal data. Wiley Interdisc. Rev.: Comput. Stat. **7**, 86–98 (2015)
4. Cressie, N., Wikle, C.K.: Statistics for Spatio-Temporal Data. Wiley, New York (2015)
5. Sheehy, J., Vinoski, S.: Developing RESTful web services with webmachine. IEEE Internet Comput. **14**(2), 89–92 (2010)
6. Maleshkova, M., Pedrinaci, C., Domingue, J.: Investigating web APIs on the world wide web. In: IEEE European Conference on Web Services. IEEE (2011)
7. A Four-Layer Architecture for Online and Historical Big Data Analytics. DASC/PiCom/DataCom/CyberSciTech (2016)
8. Ding, W., Zou, J., Zhao, Z.: A multidimensional service template for data analysis in highway domain. In: 11th International Conference on Service Science (ICSS 2018), Shanghai, China (2018)