Spatiotemporal Data Cube Modeling for Integrated Analysis of Multi-Source Sensing Data

Jing Zhao, Peng Yue

School of Remote Sensing and Information Engineering, Wuhan University
129 Luoyu Road, Wuhan, 430079, China
* Corresponding author: pyue@whu.edu.cn

ABSTRACT

One of the significant challenges of spatiotemporal big data management is the inconsistent data structures and organization methods of different data sources, such as remote sensing raster data and location-aware social sensing data. Traditional spatial data management systems are unable to support the integrated process and analysis of those data with different data types. Currently, in the field of remote sensing data management and processing, the data cube model is promising for constructing "analysis ready" remote sensing big data. This paper is initiated by the data cube model from data warehousing research area, and investigates the integrated data model and organization methods for remote sensing data and location-based social sensing data based on spatiotemporal data cube. Moreover, we implement the spatiotemporal data cube model on an array DBMS, and show a disaster analysis scenario for integrated analysis of multi-source spatiotemporal big data.

Index Terms **—** remote sensing data, social sensing data, integrated model, spatiotemporal data cube, disaster analysis scenario

1. INTRODUCTION

As big data attracts attention in a variety of fields, research on management and analysis of large-scale scientific data has gained popularity. As one of the common data types of big data, spatiotemporal data has been widely applied in various domains such as scientific research and mobile applications [1]. For instance, in GIS field, analyzing earth observation (EO) data like remote sensing images collected by sensors carried by satellites, airplanes, etc., are widely used for urban computing, disaster management, etc. Meanwhile, with the advancement of ubiquitous wearable sensors (such as smartphones), human-centric sensing, also called as social sensing, has generated massive location-aware sensing data, such as check-in data, moving trajectories, IC card records, etc. Such sensing data which contains spatial and temporal information of target objects, has characteristics of large volume and multiple data types, and makes the integrated and interactive analysis of them be more challenging.

In this paper, in order to integrate different spatiotemporal sensing data, such as remote sensing images and location-aware social sensing data, we propose a unified and flexible spatiotemporal data cube

978-1-5386-9154-0/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: University of Guelph. Downloaded on July 06,2020 at 17:35:26 UTC from IEEE Xplore. Restrictions apply.

model, as well as its organization methods on the array database. We implement the data cube model on an open source array DBSM, SciDB [6, 9], to enable the query processing of multi-source spatiotemporal data.

The remainder of the paper is structured as follows. Section 2 introduce related work on data cubes. The proposed data model is illustrated in Section 3. Section 4 presents the implementation details and integrated analysis scenarios. Finally, Section 5 presents conclusions.

2. RELATED WORK

Data cube is one of the most typical data models for data analysis, and widely applied to business intelligence, spatial data and EO data modeling. In traditional RDBMS (Relational DBMS), data cube is organized by tables based on data schemas, such as star schema and snowflake schema [4]. As an extension to spatial database, [3] is the first ones to propose a framework for spatial data warehouses based on the star-schema, in which the cube dimensions can be both spatial and non-spatial, while the measures are regions in space, in addition to numerical data. As another data organization type of data cube, array-based structure is developed in [7, 8] for efficient computing of large multidimensional arrays.

In order to satisfy the requirement of EO data analysis, Open Data Cubes (ODC) [2] provides a system architecture for storage and analysis of time-series spatial data. As the implementation of its concept, Australian Geoscience Data Cube (AGDC) [5] aims to realize the usage of EO data, especially remote sensing imagery data by big data approaches. The data organization of AGDC utilizes the combination of relational database and NoSQL index for raster data.

Totally speaking, existing data organization methods of data cubes are designed for specific target data types, without consideration of a unified model for multiple data types. In this paper, we aim to overcome the difficulties of integrating multi-source spatiotemporal data, i.e., remote sensing data and social sensing data, and propose a unified data cube model, as well as its implementation on an array-oriented database.

3. METHOD

The proposed spatiotemporal data cube model is shown in Fig. 1. The 3-D spatiotemporal data cube shown left consists of three dimensions, i.e., x, y coordinates and time t, which is designed for remote sensing raster data. The attribute of each cell corresponds to the remote sensing data like gray scales. To enable integrated analysis of remote sensing raster data and discrete location-aware social sensing data, we transfer the GPS trajectories into aggregated values, corresponding to a multi-layer grid structure (shown on the middle) of target spatial area. The 2-D array on the right is constructed for discrete GPS trajectory data. The two dimensions are time and AreaID, with which the maximum bounding box of the corresponding area is obtained using hash map. The attribute of the 2-D array is the aggregated values (e.g., sum) of GPS points on the target area at a specific time interval.

The incentive of using a hash table to locate maximum bounding box of each grid area in the multi-layer grid is the data sparsity of discrete GPS points, which leads to difficulties on data partitioning, and bad performance on query processing. Based on the proposed data model, we are able to easily and efficiently get the aggregated values of GPS trajectories for each area and time, by the functionality of array

Authorized licensed use limited to: University of Guelph. Downloaded on July 06,2020 at 17:35:26 UTC from IEEE Xplore. Restrictions apply.

DBMS. Using hash map, we can retrieve the maximum bounding box of the target grid area efficiently, and further access the related remote sensing data for integrated analysis.

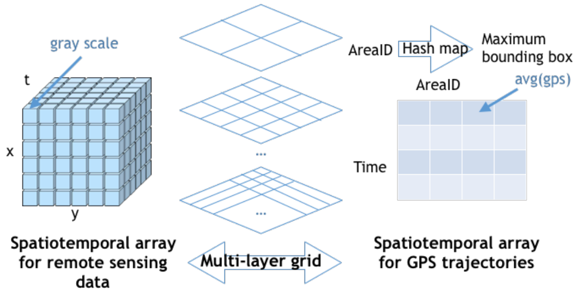

Fig. 1. Spatiotemporal Data Cubes for Remote Sensing Images and GPS Trajectories

4. IMPLEMENTATIONS AND INTEGRATED ANALYSIS SCENARIOS

Some array DBMSs such as RasDaMan [7], SciDB [6, 9], etc., are developed to support scientific computing and online analytical processing (OLAP) for multi-dimensional scientific data. In this work, we implement the proposed data cube model on SciDB, one of the state-of-the-art array DBMS.

SciDB supports a variety of operators for array data, such as data extraction (e.g., slice, between) and aggregation (e.g., aggregate, regrid). For aggregate operators, a lot of aggregate functions such as sum, min, and max are supported. SciDB also support a python interface, SciDB-Py, that leverages parallel computing and visualization of scientific data on Python, while SciDB acts as a back-end database server.

In what follows, we show a disaster analysis scenario using remote sensing data and GPS trajectory data. The analysis purpose is to dig out the relevance of the distribution of persons with the flood inundation area during a specific time interval.

Given two datasets, one of which is Normalized Difference Water Index (NDWI) calculations over a subset of satellite imagery data, and the other one is evacuees' GPS trajectories. We model the NDWI dataset as a 3-D array as the following schema:

Array_ndwi<ndwi>[x, y, t].

The ndwi calculations are stored as an attribute in the array with three dimensions, i.e., coordinates *x*, *y*, and time *t*. The GPS trajectories are modeled as a 2-D array, the schema of which is:

Array_GPS<num>[AreaID, Time].

The attribute *num* is the aggregated values of GPS points, such as average, maximum or minimum numbers.

In the following, we show a three-step query processing by SciDB-Py for the integrated analysis:

1) **Filtering**: Filter the 2-D array Array_GPS by the condition of num>100, which is conducted by the following query sentence: **Array_GPS [num > 100].toarray()**. The result array, represented as Array_AreaID_T, contains the list of AreaID and Time;
2) **Localization**: Compute the maximum bounding box of the retrieved AreaIDs list by time intervals using hash map;
3) ***Dice* operator**: Get the remote sensing data corresponding to the retrieved maximum bounding box and time intervals by *dice* operator on Array_ndwi.

By the three-step query processing described above, analysts are able to get an insight of the flood inundation situation at human-dense area, and further

Authorized licensed use limited to: University of Guelph. Downloaded on July 06,2020 at 17:35:26 UTC from IEEE Xplore. Restrictions apply.

support disaster decision making, such as choosing shelter places and organizing human relief activities.

5. CONCLUSIONS

In this paper, we designed a unified and flexible spatiotemporal data cube model for integrated analysis of remote sensing raster data and location-aware social sensing data. The proposed data cubes are organized as multi-dimensional arrays, which are relevant with each other by multi-layer grid and hash map. We implemented the data model on an array DBMS, SciDB, in which built-in operators are available for scientific computing and integrated analysis of target data. We also shown a disaster analysis scenario, which provides a possible solution for integrated process and analysis of multi-source spatiotemporal data.

ACKNOWLEDGEMENT

The work was supported by Major State Research Development Program of China (No. 2017YFB0503704), National Natural Science Foundation of China (No. 41722109), Hubei Provincial Natural Science Foundation of China (No. 2018CFA053), and Wuhan Yellow Crane Talents (Science) Program (2016).

6. REFERENCE

[1] A. Eldawy and M. F. Mokbel. "The era of big spatial data: A survey". Found. Trends databases, 6(3-4):163–273, 2016.

[2] K. Brian. Overview of the Open Data Cube Initiative. In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, 8629-8632, 2018.

[3] J. Han, N. Stefanovic, K. Koperski. Selective Materialization: An Efficient Method for Spatial Data Cube Construction. In Research and Development in Knowledge Discovery and Data Mining, 144–158, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[4] K. Ralph, R. Margy. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. Wiley Publishing, 3rd edition, 2013.

[5] A. Lewis, S. Oliver, L. Lymburner, et al. The Australian Geoscience Data Cube- Foundations and Lessons Learned. Remote Sensing of Environment, vol. 202, no. Supplement C, 276- 292, 2017.

[6] M. Stonebraker, P. Brown, A. Poliakov, et al. The Architecture of SciDB. In Proc. SSDBM, volume 6809 of LNCS, 1-16, 2011.

[7] P. Baumann, A. Dehmel, P. Furtado, et al. The Multidimensional Database System RasDaMan. In Proc. ACM SIGMOD, 575–577, 1998.

[8] S. Sarawagi, M. Stonebraker. Efficient Organization of Large Multidimensional Arrays. In Proc. ICDE, 328–336, Washington, DC, USA, 1994. IEEE Computer Society.

[9] M. Stonebraker, P. Brown, D. Zhang, et al. SciDB: A database management system for applications with complex analytics. IEEE Computational Science & Engineering, 15(3):54–62, 2013.

Authorized licensed use limited to: University of Guelph. Downloaded on July 06,2020 at 17:35:26 UTC from IEEE Xplore. Restrictions apply.