https://doi.org/10.20965/jdr.2017.p0347

Paper:

Difference Operators in Simulation Data Warehouses

Jing Zhao†, Yoshiharu Ishikawa, Yukiko Wakita, and Kento Sugiura

Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
†Corresponding author, E-mail: zhao@db.ss.is.nagoya-u.ac.jp
[Received October 16, 2016; accepted March 3, 2017]

In analyzing observation data and simulation results, there are frequent demands for comparing more than one data on the same subject to detect any differences between them. For example, comparison of observation data for an object in a certain spatial domain at different times or comparison of spatial simulation data with different parameters. Therefore, this paper proposes the difference operator in spatio-temporal data warehouses, which store temporal and spatial observation data and simulation data. The requirements for the difference operator are summarized, and the approaches to implement them are presented. In addition, the proposed approach is applied to the mass evacuation of simulation data in a tsunami disaster, and its effectiveness is verified. Extensions of the difference operator and their applications are also discussed.

Keywords: data warehouse, difference operator, spatio-temporal databases, disaster information, simulation data

1. Introduction

As big data attracts attention in a variety of fields, research on data analytics for sophisticated analytic processing of a large amount of data in a database has gained popularity [6]. As for spatio-temporal databases, there are growing demands for analyzing large-scale spatio-temporal data in various domains such as mobility data, moving trajectory data, and scientific data. Our research group has been engaged in researches on spatio-temporal data warehouses where large-scale spatio-temporal simulation data are specially stored to enable interactive analyses, which are referred to as simulation data warehouses. In particular, the research is focused on analyzing simulation data on tsunami and earthquake disasters [14].

This paper specifically examines differences as one of the basic analytic requirements for spatio-temporal data warehouses, in which detection of temporal changes as well as differences between observation data with different parameters or conditions should be required. What types of difference operators are appropriate for the above-mentioned data analyses have still not been clarified. Various types of difference operators may be possible, depending on the properties or application purposes of object data. On the other hand, due to the challenges in detecting any remarkable changes within a large amount of data, it is necessary to develop some efficient algorithms by effectively using the latest database system technology. Based on the above-mentioned context, we proposes the general-purpose operators and the correspongding methods, as well as analyzing the requirements in order to detect any differences from the spatio-temporal data warehouses.

This paper consists of eight sections: Section 2 describes related studies; Section 3 describes analyses of the requirements for the difference operator; Section 4 shows specific images of the proposed difference operator; Section 5 provides definitions of the difference operator based on the preceding sections; Section 6 provides algorithms for constructing difference histograms; Section 7 describes implementations and experiments of the difference operator; and Section 8 contains discussions and summary.

2. Related Studies

Studies on data warehouses and OLAP (On-Line Analytical Processing), originally intended for business fields, have now been extended to the research and developments of spatio-temporal data.

The 11th chapter of [13] and [4] describe commentaries and studies on such research and development. Many of the approaches to spatio-temporal data warehouses use map data to analyze statistical information on maps (for example, population distributions) and other information. In some cases, they use temporal information in addition to spatial information. In that sense, the data warehouses for moving trajectory data [7] are technically deeply related to this study in that spatio-temporal information is collectively represented.

Simulations, an important research means in many scientific fields, produce a huge amount of data day by day. In this context, supporting simulation processes via database technology is a promising approach to more effectively supporting sophisticated analyses in scientific fields. [8] reports the development of a simulation database system with relatively simple simulation processes integrated in the database ready for execution. While that study aims to support simulation processes, the

technology to be developed in this study is aimed at post processing of simulations. Since it is challenging to directly link complex scientific simulations for which supercomputers are used to any database systems, this study develops a system technology that is rather focused on the storage, integration, and analyses of simulation-processed data. The difference operator described in this paper will be found necessary in such a context.

As data analytics has recently attracted increasing attention [6], data visualization is found so effective in supporting interactive users' analyses that many studies on that subject are now under way. From the database point of view, technologies that can instantly visualize large-scale data or select data to be visualized are important. For example, MuVE [3] visualizes data by bar graphs as a result of their consideration on the viewpoints that will concentrate data into specified conditions remarkably different from the whole data. SEEDB system for the visualization of databases, though intended for category attributes, is also closely related to this study [9].

3. Analyses of Requirements

The requirements for the difference operator for data with spatio-temporal characteristics are discussed in this section. As there could be many different object data or applications for the difference operator, it is assumed that one example for which the requirements for the difference operator are analyzed. The assumed example is user's location information in a two-dimensional space at each hour as acquired by GPS and portable devices. Such data can be represented by the type of relations (id, x, y, t). The object two-dimensional space is assumed to be spatially gridded. Grid cells corresponding to given points x, y are assumed easy to seek.

The following is an example of the requirements:

> Aggregate the number of moving users in each cell at time segments $I_1 = [t_1, t_2]$ and $I_2 = [t_3, t_4]$ and report any cells with remarkable differences

This requirement is to seek any differences in the distribution of the number of users between I_1 and I_2. Even such a simple example of requirements involves several considerations to be made as follows:

- What kind of aggregation is expected: The most common way of aggregation would be to count the numbers of records on the object cells and time segments (corresponding to SQL and SUM functions). To seek their distribution patterns, the frequency distributions as divided by the total number of records could be an option. For other kinds of object data, one could use aggregation functions such as AVG and MAX.

- How to detect "remarkable differences": It becomes necessary to formulate the differences. The requirements for the differences might be different with object data and their applications. This study focuses on the analysis of numerical data on the number of evacuees on spatio-temporal simulation data. If any differences in some of the cells in a certain spatial domain (for example, changes in the number of evacuees) are greater than those in the entire domain, then such differences are deemed remarkable. Whether differences are remarkable or not is to be determined by the thresholds specified by users. Among many different indices to measure differences, errors between the difference value of each cell and the distribution of differences in the entire domain are utilized. In this way, more resultant errors indicates more substantial differences. Definitions of specific difference errors are described in Section 5.

- Although it is assumed spatial grids are as presented in the past, appropriate grain sizes of grids must be selected in presenting the differences. As users are not always in a position to know appropriate grain sizes of grids in advance, the grain sizes of grids specified by users may be too fine or too coarse.

- Freedom in Specifying Time Segments: For instance, with T_2 denoting the entire time, one can measure any differences between a certain period of time and the entire period of time. Moreover, assuming time segments are not given, "which time segment in the time segment of width τ has the largest difference from the entire period of time" may become a possible requirement.

- In what form reports should be made: Reports may be data that are output in text and tabular formats or visualized another way. Assuming that reports are to be visualized, definitions of differences should be made to suit such visualization of reports.

Based on the above-mentioned analyses, the images of basic difference operators are presented as an example in the following section.

4. Images of Difference Operators

As an instance of the difference operator, the basic one is considered. **Fig. 1** shows the images of the basic difference operator. The figure on the left side of **Fig. 1** shows the aggregate results in time segment T_1. The shades of cells correspond to the sizes of aggregate values (the number of moving users in the cell during the period of time). The middle figure shows the aggregate results in time segment T_2. In actual use of the difference operator, one can choose to use the aggregate values at the two time segments T_1, T_2 as they are or to use the normalized aggregate values as divided by the total number, depending on object data and their applications. In this paper, it is assumed for the sake of simplicity that users have selected the former method.

The figure on the right side of **Fig. 1** approximately represented images of the differences between the aggregate

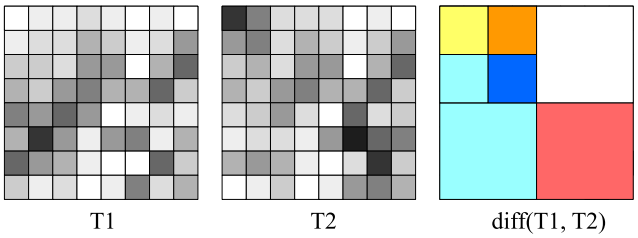


Fig. 1. Image of the difference operator.

results of T_1 and T_2 for the input data at the time segments T_1, T_2. In the heat map expressions, the more strongly red regions are, the aggregate value at the time segment T_2 is larger than that at the time segment T_1.

On the other hand, the more strongly blue regions, the aggregate value at the time segment T_1 tends to be larger than that at the time segment T_2. In order to represent a rough trend of differences, any adjacent cells with a similar trend of differences are expressed in a lump as one cell. Given quadtree-like spatial divisions to make the sides of a cell equal in size to the power of 2 in length.

Presentation of such output results enables users to easily grasp any differences between two different time segments. The above-mentioned explanation that the aggregate results at the time segments T_1 and T_2 should be generated before calculating their differences does not always apply to actual implementations, where there is still room for more efficient steps.

5. Settings of Difference Operators

The idea of spatial histograms [1, 2] is used to represent the results of the difference operator. Histograms are widely used for database query optimization and others [5], and their extensions to spatial databases make spatial histograms of approximated spatial frequencies of spatial data. The results of the difference operator are referred to as difference histograms.

Difference operators are formulated as follows. Symbols used for the formulation of the difference operator are shown in **Table 1**. First, a parameter n is given for grain sizes by which to divide the space. In the case of **Fig. 1**, $n = 3$ and the aggregation based on fine grain grids is processed on the grid structure divided into $2^n x 2^n = 64$. The set of total cells in the fine grain grid division is denoted by C_{base}. Then, $|C_{\text{base}}| = 2^n \times 2^n$.

There are many different approaches possible to construct spatial histograms. What is important in constructing spatial histograms is how to divide a space. One of such approaches is to create cells of $2^m \times 2^m$ $(m < n)$ in size (within the quadtree boundary) as shown in the figure on the right side of **Fig. 1**. Another approach is STHoles method [2] that allows the cells in a histogram to overlap each other. This paper utilizes the quadtree approach that is intuitively easy to learn as well as easy to formulate its construction processes.

In the difference histogram shown in the figure on the right side of **Fig. 1**, difference information is approximated by a limited number of cells. The set of cells in such a difference histogram is denoted by C_{hist}.

Table 1. Symbols and their meaning.

Symbol	Meaning
n	Parameters for fine grain divisions
C_{base}	Sets of cells in fine grain divisions
p	Number of quadtree partitioning times in constructing a difference histogram
C_{hist}	Set of cells in a difference histogram
B	Number of cells of a difference histogram
$\text{error}(c)$	Errors of Cell c of a difference histogram
$\text{total\_error}(C_{\text{hist}})$	Errors of difference histogram C_{hist}
$\text{count}(c)$	Aggregate value of Cell c
$\text{base\_cells}(c)$	Set of fine grain cells corresponding to cell c of a difference histogram
$\text{level}(c)$	Division level of Cell c

The total number of cells created in constructing a difference histogram by p times of the quadtree partitioning approach described below is $|C_{\text{hist}}| = 3p + 1$. In the case of the figure on the right side of **Fig. 1**, since quadtree partitioning processing has been executed twice, creating seven cells in total $3 \times 2 + 1 = 7$.

In this paper, it is assumued that the total number of cells which is denoted by B, in a difference histogram is specified by users. The parameter B needs to meet the constraint of $B = 3p + 1$ $(p = 0, 1, \ldots,)$ according to the definitions of difference histograms. Moreover, as the granularity of difference histograms cannot be finer than that of the original data, $B \leq |C_{\text{base}}|$.

Any difference histogram to be constructed should have as small errors from the original difference data as possible. Square errors are assumed for an error function. The aggregate value of a cell in the difference histogram or cell c divided by a fine grain size, that represent the number of data contained in the region of cell c, is denoted by $\text{count}(c)$. The set of fine grain cells in cell c of the difference histogram is denoted by $\text{base\_cells}(c)$. Then, errors from cell c of the difference histogram are defined by the following equation:

$$\text{error}(c) = \sum_{b \in \text{base\_cells}(c)} \left(\frac{\text{count}(c)}{4^{\text{level}(c)}} - \text{count}(b) \right)^2 \quad (1)$$

where b denotes the fine grain cells. $\text{level}(c)$ represents the level of cell c levels, where level 0 for fine grain cells and level 1 for cells with one-level coarser grains, raising the level number as the grain coarseness increases. The above-mentioned equation is to seek errors between the accurate aggregate value of fine grain cell c and the approximate aggregate value of cell c of the difference histogram. The aggregate value per area at a fine grain level is obtained by dividing the aggregate value by $4^{\text{level}(c)}$.

The total error of the difference histogram is defined by the following equation, using the above-mentioned func-

tion.

$$\text{total\_error}(C_{\text{hist}}) = \sum_{c \in C_{\text{hist}}} \text{error}(c) \quad \ldots\ldots (2)$$

In this paper, assume that each cell corresponds to a total number of data so that one can easily extend it to make the mean data value of a cell correspond to the particular cell. In constructing any difference histograms, they should be so constructed that errors *err* should be minimum under the given constraint B for the number of divisions.

6. Algorithm for Constructing Difference Histograms

At this point an algorithm based on an greedy approach is considered. The basic policy is to divide a data set top down. Algorithm 1 represents such an algorithm.

Algorithm 1: Histogram construction algorithm.

```
   Input: R: Root cells, B: Total number of cells of a
          histogram
   Result: C: Set of cells as divided
1  C ← children(R);
   // Set of four children cells
2  p ← (B − 1)/3;
   // Times of divisions (B > 1)
3  for i ← 2 to p do
4  |  max_Δerror ← 0;
5  |  for c ∈ C do
6  |  |  if is_base_cell(c) continue;
   |  |  // Cannot be divided
7  |  |  Δerror ← get_Δerror(c);
   |  |  // Improvements of errors
8  |  |  if Δerror > max_Δerror then
9  |  |  |  max_Δerror ← Δerror;
10 |  |  |  c_opt ← c;
11 |  |  end
12 |  end
13 |  C ← (C \ {c_opt}) ∪ children(c_opt);
14 end
15 return C;
```

Input R denotes grid cells corresponding to the object spatial regions. In the case of **Fig. 1**, R refers to cells corresponding to the entire space. On the other hand, one may specify some of the cells in the quadtree regions of the entire space by R as well. In this case,one can make the algorithm specific to differences in partial regions. Another input B denotes the total number of cells of a difference histogram. When $B = 1$, no divisions are executed, and required processing is so obvious that Algorithm 1 assumes $B > 1$.

Children (c) on lines 1 and 13 denotes a function that returns the set of four children cells of a given Cell c. get_Δerror(c) on line 7 denotes the function that returns possible improvements in error values when a given cell c is assumed to be partitioned. It is defined by the following equation:

$$\text{get\_}\Delta\text{error}(c) = \text{error}(c) - \sum_{c' \in \text{children}(c)} \text{error}(c') \quad (3)$$

In this algorithm, the outside loop is executed $p-1$ times. The first divisive processing (when $i = 1$) is automatically executed. Itinerating the process with increasing i, the size of C on line 5 is given as $|C| = 3i+1$, so that the total number of executions of the inside loop can be expressed by the following equation.

$$\sum_{i=1}^{p}(3i+1) = \frac{3}{2}p^2 + \frac{5}{2}p \quad \ldots\ldots\ldots\ldots (4)$$

The above-mentioned times of loop executions indicates computational complexity $O(p^2)$ and $O(B^2)$ as p is proportional to B.

In actual computations, however, the value of B will not be so large that though on the square order, computation time will not pose any serious problem. As for the count function that is called many times in the processing of the algorithm, it needs to aggregate actual data, in which computation of large-scale data may have a dominant impact. Actual computational costs could be reduced by taking advantage of the aggregate function of the database system as described below or by retaining already previously calculated values.

7. Implementation of Difference Operators and Experiments

7.1. Implement Environment

The difference operator is implemented by using SciDB [10–12], an array-oriented DBMS. Array-oriented DBMS is a DBMS specific to the management of large-scale array data and queries that appear in the scientific fields. Therefore, it could be an adaptive array-oriented DBMS for this paper, in which spatio-temporal simulation data are assumed stored in SciDB for analyses so that the difference operator can be applied to them as found necessary. The difference operator is implemented in R (programming language) and executed via the R interface provided on SciDB. The visualization function described below is also implemented in R.

In the processing of the difference operator, the `subset` operator and `aggregate` operator that are provided by the R interface of SciDB extract all aggregated information on corresponding cell sets based on the specified conditions for the spatio-temporal domains. In the process of taking direct differences between two array data, first calculate their connections by using the `merge` operator, and set any cell connections with a null value as 0. Next, generate difference arrays by taking differences in attributes, and then apply the above-mentioned algorithm, and the results are represented in raster graphics by the SciDB's `raster` operator and further processed into heat maps.

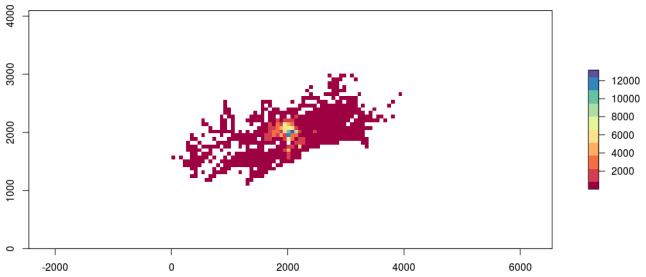


Fig. 2. Evacuation data distributions.

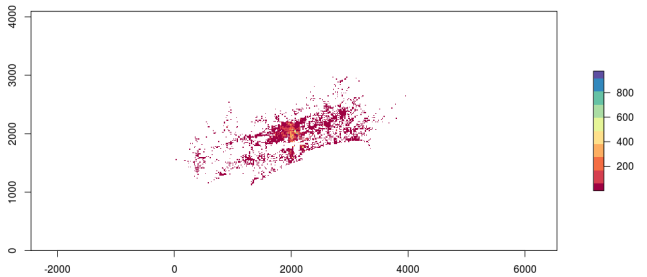


Fig. 3. Evacuation data at $T_1 = [9:00, 10:00]$.

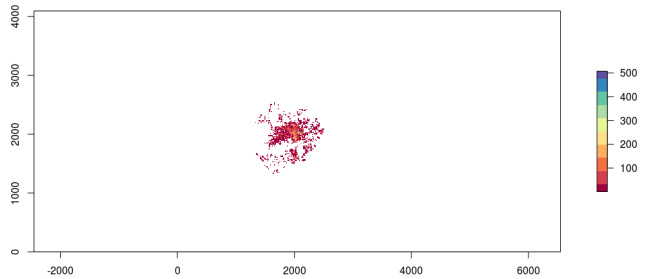


Fig. 4. Evacuation data at $T_2 = [11:00, 12:00]$.

7.2. Object Data Set

The object data set is mass evacuation simulation data in the event of a large-scale earthquake in Kochi City, which has been provided from the Sekimoto Research Laboratory of the University of Tokyo. The sample mass evacuation simulation data used in the experiments represents six-hour data compiled under the conditions in which an earthquake occurs at 9 a.m., and the mass evacuations peak sixty minutes after its occurrence: the simulation of about forty thousand people's evacuations is based on person trip data.

Simulation data are a collection of records in the format of $(id, time, x, y)$: id denotes user ID; $time$, a time stamp; x, y, the location of a user at $time$. Simulation data are static data that remain unchanged after the simulation. In this work, in the query processing on data warehouses, it is generally taking advantages of static data, to preprocess the data and make interactive processing more efficient. The above-mentioned mass evacuation simulation data is divided spatially by a maximum grain size of $4,096 \times 4,096$. Based on the fine grain spatial division, the number of evacuees in each cell is aggregated every minute. The aggregated data resulting from the preprocessing are loaded on SciDB and are subjected to the query processing of the difference operator.

Figure 2 shows an image of mass evacuation data, which represents the aggregated mass evacuation frequencies (the number of evacuees per area) at the time segments $[9:00, 15:00]$ for each cell after roughly dividing the areas around Kochi City subject to the simulations into 64×64 cells. The central part of the data corresponds to the central part of Kochi City with the coast line running below. The figure shows high evacuation frequencies in the vicinity of the city's central part. However, temporal changes in the mass evacuations cannot be identified from the figure.

The following query was utilized in the experiments.

> Seek the difference histograms of the differences in the distribution of evacuees at time segments $T_1 = [9:00, 10:00]$ and $T_2 = [11:00, 12:00]$ in the entire area of the evacuation simulation data.

The object space is the entire area subject to the simulations.

In actual applications of the algorithm, treating cells with a null value will be a problem. One idea may be to regard the values of cells with a null value, such as 0, and apply the above-mentioned algorithm; other possible approaches may be to leave cells with a null value out of consideration or to interpolate the values of cells with a null value by the values of nearby cells. If cells have no values due to topographical constraints (for example, sea regions in the mass evacuation data), external constraints could be an idea. The approaches or ideas that are appropriate depends on the user's intentions, nature of the data, and applications, and so it may be appropriate to treat those approaches as options of the difference operator.

In the implementation of the difference operator in the experiments, those optional approaches have been reviewed and the decision reached was to leave cells with a null value out of consideration. As for the aggregate values, specifically in Eq. (1) the domain should be divided by the total number of fine grain cells with no null values rather than by $4^{\text{level}(c)}$ (the total number of fine grain cells corresponding to c).

7.3. Experimental Results

Figures 3 and **4** show visualized data of the total numbers of evacuees at the time segments T_1 and T_2 respectively. They represent input data for the internal processing of the difference operator. These figures show that as compared with the distribution of evacuees at the time segment T_1, the total number of evacuees at the time segment T_2 is smaller and is differently distributed. However, a simple comparison of the two figures barely reveals any specific changes between them. Therefore, these two array data are subjected to difference processing to analyze the actual changes.

Figure 5 displays directly obtained differences in the

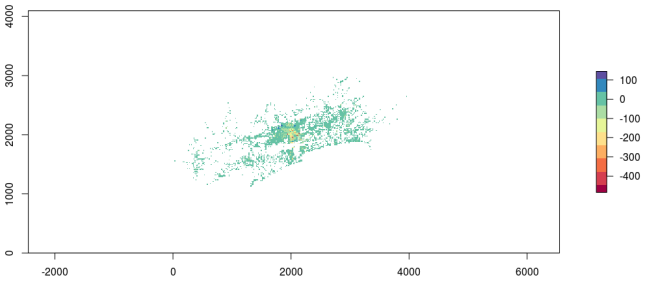


Fig. 5. Direct differences at time segments T_1 and T_2.

values of the two time segments. It shows detailed data than but is a little too finely grained to grasp an overall trend of mass evacuations.

Figures 6, **7** and **8** show the results of difference histograms with $B = 20$, $B = 50$ and $B = 100$, respectively. The results of the difference operator are visualized by heat maps.

In **Fig. 5**, the parts with significant difference values are first filtered with a minimum bounding box before constructing histograms that approximate their overall trends. Basically, the larger the number of cells B, the more approximated will be the results of difference histograms to actual data distributions. However, for the sake of visualization, cells with similar trends are displayed integrally and the parts with more remarkable differences are highlighted by histograms. For example, the histogram results in **Fig. 6** show the regions where the evacuees have most increased or decreased. If analyzers want to obtain more detailed information, they can just increase the value of B and return the histograms for the system to process. Because the more cells do not always generate better visualization effects, analyzers should carry out interactive analyses by coordinating object regions and settings of B until they can obtain useful information. For instance, as in **Fig. 8** ($B = 100$), the central part is not clearly displayed when B has a large value. In such cases, analyzers can iterate the difference operator limited to the green rectangular region R_1 for the number of cells $B = 100$. Then the difference histograms limited to the region R_1 can be obtained as shown in **Fig. 9**, from which more detailed difference distributions can be understood. Thus, the approach proposed in this paper proves to be capable of detecting remarkable differences in any local, large differences by seeking difference histograms of limited regions and visualizing them as an enlarged view.

Figure 10 shows the processing time of the difference operator. Red and green lines indicate the execution time from the query execution till the end of the difference operator and till their display, respectively. The execution time indicated by the green line includes not only the query execution time but also the time for subsequent visualization processing. As can be seen from **Fig. 10**, the larger the number of cells B, the longer the processing time of the difference operator. A comparison of the two lines shows that the overhead time for visualization processing is not so large.

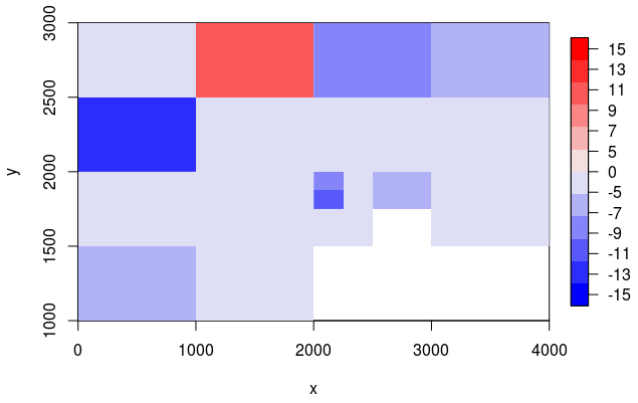


Fig. 6. Results of the difference operator ($B = 20$).

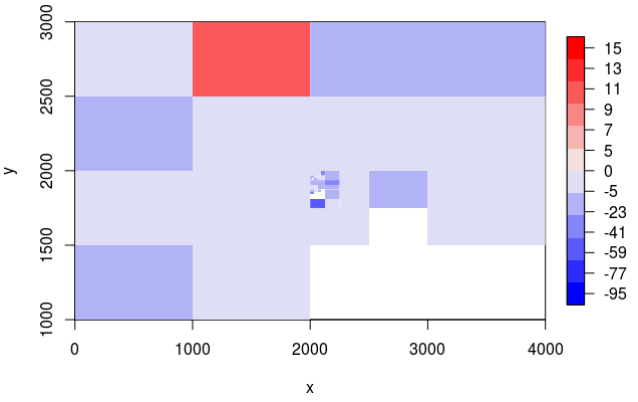


Fig. 7. Results of the difference operator ($B = 50$).

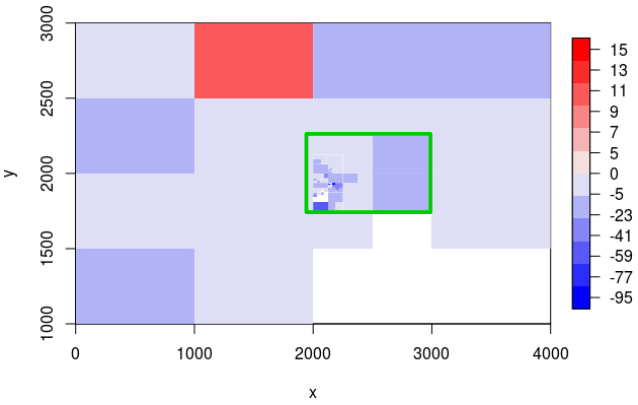


Fig. 8. Results of the difference operator ($B = 100$).

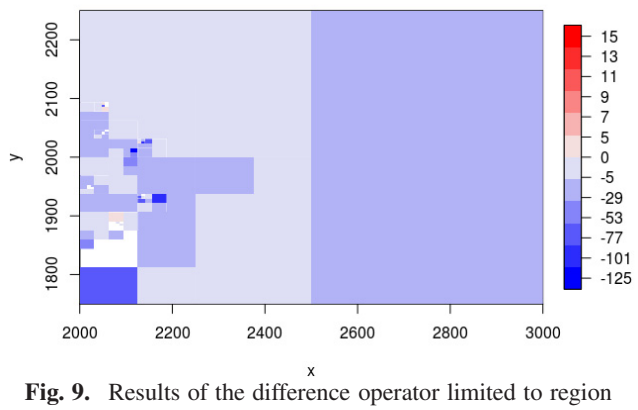


Fig. 9. Results of the difference operator limited to region R_1 ($B = 100$).

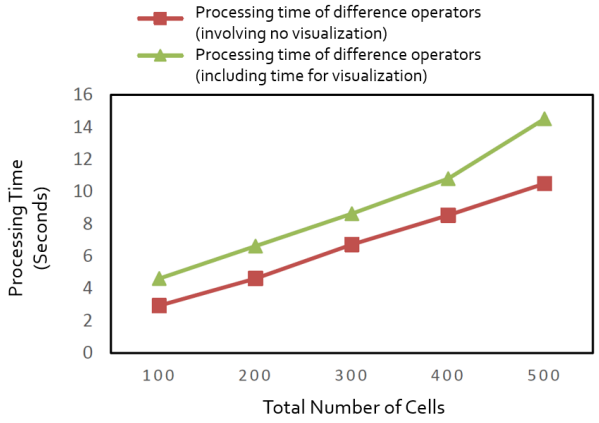


Fig. 10. Processing time of the difference operator.

8. Discussions and Summary

In order to realize sophisticated analytical functionalities, we proposed the difference operator to detect any differences in simulation data with spatio-temporal characteristics. We analyzed the requirements for such difference operators and proposed an approach to construct difference histograms. We also visualized the implementation of such difference operator, taking advantage of SciDB, an array-oriented DBMS. We discussed the effectiveness of the difference operator on the visualization results from the experimental results. As for the efficiency of the proposed approach, its processing time is approximately 2s if the total number of cells B is not very large (less than or equal to 100). A small overhead for the visualization processing proves that the proposed approach can enable interactive analyses.

The difference operator proposed in this paper assumes that users specify two time segments T_1 and T_2. Such an approach, however, is applicable only when the time segments to be specified by users are known in advance. For more useful and evolutionary approaches, it could be extended as follows:

1) Users can only specify the regions subject to analyses and the time segment width τ as an aggregate unit. The data sets can be aggregated in each time segment width τ and a sequence of aggregate results $T_1, T_2, \ldots, T_m$ can be constructed. Then, any differences between T_{i+1} and T_i $(1 \leq i \leq m-1)$ are determined and the top k pairs are selected in order of difference size from the largest one. In other words, selecting k pairs in order of the value of differences that users need to notice could effectively help users save their analyzing burden.
2) Similar to the above-mentioned idea, we have the following: 1) for selecting pairs by comparing two adjacent time segments, select top k pairs by comparing the trend of each time segment T_i $(1 \leq i \leq m$ with that of the entire time segment as an alternative. In other words, pairs are selected on the degrees of deviation from the general trend rather than on changes at a certain point.

The above-mentioned approaches for extended operations, particularly approach 1), involve high costs. Therefore, we need to review how to reduce the involved costs.

As for time segments τ, instead of user's arbitrary selections, users could alternatively be allowed to select only powers of 2 such as $1, 2, 4, 8$ and so on in the same way as that for spatial data. In such cases, if object data are static, they can be aggregated in advance, using the data warehouse technology [13] to reduce the processing time for executing the difference operator. That should be a realistic solution for realizing interactive processing.

Another future issue required to be addressed could be a semantic extension of the difference operator. Although in this study we have simply noticed nothing but the difference sizes, we hope to develop the difference operator on information such as whether differences have an increasing tendency or a decreasing one and the speeds at which they increase or decrease. The way to visualize such differences is another important matter of consideration. Therefore, we intend to review how to visualize them. We also plan to develop an implementation technology making the most of the functions of the array-oriented DBMS.

Acknowledgements

This study was partly supported: by the Grants-in-aid for Scientific Research (16H01722, 26540043), CREST: "Creation of Innovative Earthquake and Tsunami Disaster Reduction Big Data Analysis Foundation by Cooperation of Large-Scale and High-Resolution Numerical Simulations and Data Assimilations."

References:

[1] S. Acharya, V. Poosala, and S. Ramaswamy, "Selectivity estimation in spatial databases," In ACM SIGMOD, pp. 13–24, 1999.
[2] N. Bruno, S. Chaudhuri, and L. Gravano "STHoles: A multidimensional workload-aware histogram," In ACM SIGMOD, pp. 211–222, May, 2001.
[3] H. Ehsan, M. A. Sharaf, and P. K. Chrysanthis, "MuVE: Efficient multi-objective view recommendation for visual data exploration," In ICDE, pp. 731–742, 2016.
[4] L. Gómez, B. Kuijpers, and B. Moelans, "A survey of spatio-temporal data warehousing," International Journal of Data Warehousing and Mining, Vol.5, No.3, pp. 28–55, 2009.
[5] Y. Ioannidis, "The history of histograms (abridged)," In VLDB, pp. 19–30, 2003.
[6] Y. Ishikawa, "Research trend and future prospects for large-scale data analytics," IEICE Trans. on Information and Systems (Japanese Edition), J97-D(4), pp. 718–728, 2014 (in Japanese).
[7] L. Leonardi, G. Marketos, E. Frentzos, N. Giatrakos, S. Orlando, N. Pelekis, A. Raffaetà, A. Roncato, C. Silvestri, and Y. Theodoridis, "T-Warehouse: Visual OLAP analysis on trajectory data," In Proc. ICDE, pp. 1141–1144, 2010.
[8] H. Lustosa, F. Porto, P. Blanco, and P. Valduriez, "Database system support of simulation data," PVLDB, Vol.9, No.13, pp. 1329–1340, Sept. 2016.
[9] A. Parameswaran, N. Polyzotis, and H. Garcia-Molina, "SeeDB: Visualizing database queries efficiently," Proceedings of the VLDB Endowment, Vol.7, No.4, pp. 325–328, 2013.
[10] Paradigm4: Creators of SciDB a computational DBMS, http://www.paradigm4.com/ [accessed October 1, 2016]
[11] M. Stonebraker, P. Brown, A. Poliakov, and S. Raman, "The architecture of SciDB," In SSDBM, volume 6809 of LNCS, pp. 1–16, 2011.
[12] M. Stonebraker, P. Brown, D. Zhang, and J. Becla, "SciDB: A database management system for applications with complex analytics," IEEE Computational Science & Engineering, Vol.15, No.3, pp. 54–62, 2013.

[13] A. Vaisman and E. Zimányi, "Data Warehouse Systems: Design and Implementation," Springer, 2014.
[14] J. Zhao, K. Sugiura, Y. Wang, and Y. Ishikawa, "Simulation data warehouse for integration and analysis of disaster information," Journal of Disaster Research, Vol.11, No.2, pp. 255–264, 2016.

Name:
Jing Zhao

Affiliation:
PhD Candidate, Graduate School of Information Science, Nagoya University

Address:
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Brief Career:
2013- Master Student, Graduate School of Information Science, Nagoya University
2015-PhD Candidate, Graduate School of Information Science, Nagoya University

Selected Publications:
• "A Density-based Approach for Mining Movement Patterns from Semantic Trajectories," The IEEE Region 10 Conference (TENCON 2015), November 2015.

Academic Societies & Scientific Organizations:
• Database Society of Japan (DBSJ)

Name:
Yoshiharu Ishikawa

Affiliation:
Professor, Graduate School of Information Science, Nagoya University

Address:
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Brief Career:
1994- Assistant Professor, Nara Institute of Science and Technology
1999- Assistant Professor, University of Tsukuba
2003- Associate Professor, University of Tsukuba
2006- Professor, Nagoya University

Selected Publications:
• "Probabilistic Range Querying over Gaussian Objects," IEICE Transactions on Information and Systems, Vol.E97-D, No.4, pp. 694-704, April 2014.

Academic Societies & Scientific Organizations:
• Association for Computing Machinery (ACM)
• IEEE Computer Society
• Information Processing Society of Japan (IPSJ)
• Institute of Electronics, Information and Telecommunication Engineers (IEICE)
• Database Society of Japan (DBSJ)

Name:
Yukiko Wakita

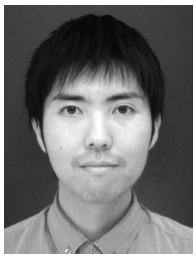
Affiliation:
Research Associate, Graduate School of Information Science, Nagoya University

Address:
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Brief Career:
2011- Part-time Lecturer, Nagoya College
2013- Part-time Lecturer, Chukyo University
2014- Researcher, Nagoya University
2016- Research Associate, Nagoya University

Selected Publications:
• "Traffic Network Design by Cellular Automaton-based Traffic Simulator," Computer Assisted Methods in Engineering and Science, Vol.22, No.1, pp. 51-61, 2015.

Academic Societies & Scientific Organizations:
• Information Processing Society of Japan (IPSJ)



Name:
Kento Sugiura

Affiliation:
PhD Candidate, Graduate School of Information Science, Nagoya University

Address:
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Brief Career:
2013- Master Student, Graduate School of Information Science, Nagoya University
2015- PhD Candidate, Graduate School of Information Science, Nagoya University

Selected Publications:
• "Grouping Methods for Pattern Matching in Probabilistic Data Streams," The 20th International Conference on Database Systems for Advanced Applications (DASFAA 2015), pp. 92-107, April 2015.

Academic Societies & Scientific Organizations:
• Information Processing Society of Japan (IPSJ)
• Database Society of Japan (DBSJ)