A Wavelet-Based Approach to Improve the Efficiency of Multi-Level Surprise Mining*

Cyrus Shahabi[1], Seokkyung Chung[1], and Maytham Safar[2]

[1] Department of Computer Science
University of Southern California
Los Angeles, California 90089–0781
{shahabi, seokkyuc}@usc.edu
[2] Department of Computer Engineering
Kuwait University, Kuwait
maytham@eng.kuniv.edu.kw

Abstract. Due to the large amount of the collected scientific data, it is becoming increasingly difficult for scientists to comprehend and interpret the available data. Moreover, typical queries on these data sets are in the nature of identifying (or visualizing) trends and surprises at a selected sub-region in multiple levels of abstraction rather than identifying information about a specific data point. In this paper, we show how a wavelet-based data structure, 2D TSA-tree (stands for Trend and Surprise Abstractions Tree) can be utilized efficiently to detect surprises on spatio-temporal data at different levels. Furthermore, we show how to find surprises within a specified period of time at different levels of abstraction (e.g., weekly, or monthly) by constructing 1D TSA-tree. To demonstrate the effectiveness of our proposed methods, we evaluated our 2D TSA-tree using real and synthetic data. The results indicate that 2D TSA-tree approach can be used to visualize different kinds of surprises effectively.

1 Introduction

Rapid growth in remote sensing systems has made it possible to obtain data about nearly every part of our larger world, including the solid earth, ocean, atmosphere and the surrounding space environment. To illustrate a sample application, consider a joint project that we defined with Jet Propulsion Laboratory (JPL) for NASA[1]. After three levels of off-line pre-processing, the temperature,

* This research has been funded in part by NSF grants EEC-9529152 (IMSC ERC) and ITR-0082826, NASA/JPL contract nr. 961518, DARPA and USAF under agreement nr. F30602-99-1-0524, and unrestricted cash/equipment gifts from NCR, IBM, Intel and SUN.

[1] The project is entitled GENESIS: GPS environmental and earth science information system (see http://genesis.jpl.nasa.gov/html/index.shtml). In this project, signals from GPS satellites are processed and analyzed to extract global atmospheric and ionospheric data.

water vapor, and refractivity of certain coordinates on earth can be extracted for every half-hour at different heights. These data can be stored in database server(s) and accessed by users via the Internet. Typical queries on these data sets are, however, different than conventional point queries. For example, a query to acquire the temperature of a specific location at a specific time and date is rare. The more frequent queries are in the nature of identifying (or visualizing) trends and surprises of a *selected sub-region* at *multiple levels of abstraction*.

In [7], we proposed a versatile wavelet-based data structure, termed TSA-tree (stands for Trend and Surprise Abstractions Tree) for finding trends and surprises in time-series data. In [8], we extended TSA-tree to 2D to enable efficient multi-level *trend* mining on spatial data. We ignored surprise mining in [8] because detecting surprises at different dimensions was not of interest to the applications dealing with 2D spatial data (e.g., atmospheric data on 2D area). However, with spatio-temporal data, surprises in horizontal and vertical dimensions become important.

In this paper, we show how to support multi-level surprise queries on 2D spatio-temporal data where the space is conceptualized through a single dimension (e.g., ground instrument locations) and time as the second dimension. Using the detail values of 2D TSA-tree, we identify two kinds of surprises, which are referred to as *horizontal* and *vertical surprises*. By precomputing different levels of 2D TSA-tree, we do not need any reconstruction overhead nor the entire data for answering multi-level surprise queries. We also show how 2D TSA-tree can be used to support two different user interactions with 2D spatio-temporal data for surprise mining: 1) finding *horizontal* and *vertical* surprises at different levels, and 2) finding *horizontal* surprises within a given time interval at different levels by constructing 1D TSA-tree when 2D wavelet transform is not applicable.

With traditional outlier analysis, the first step is to inspect plots of the original data. Such plots help one define the nature of the data. Thus, the plot of the data against time is often sufficient to identify *outliers* [5,3,4]. However, the outlier detection techniques are expensive when compared with a wavelet-based method such as our 2D TSA-tree. Furthermore, it cannot capture surprises at multiple levels of abstraction.

The remainder of this paper is organized as follows. In Sec 2, we explain basic concepts of 2D TSA-tree. Sec. 3 shows how user can interact with 2D spatio-temporal data for surprise mining. Sec. 4 provides an analysis of our technique and Sec. 5 concludes the paper.

2 2D TSA-tree

In this section, we explain TSA-tree for time-series databases and how it can be extended to 2D TSA-tree for 2D spatial data mining. Throughout this paper we use Haar wavelet filter for our discussions.

In [7], we proposed a novel tree-like data structure termed *TSA-tree* (stands for Trend and Surprise Abstractions) for efficient management of time-series

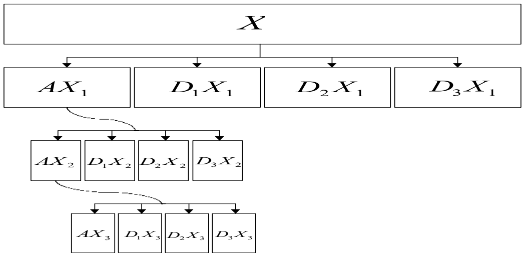


Fig. 1. 2D TSA-tree

databases. To provide efficient support of multi-level queries (e.g., within a week, a month or a year), TSA-tree precomputes trends and surprises at different levels and store them in a tree. By utilizing the wavelet transform, we can naturally split a time-series sequence into two nodes where one node captures the trends and the other the surprises within the original sequence. Then, the nodes of TSA-trees can be used to visualize trends and surprises at different levels. For detailed information, see [7].

We extended TSA-tree model to 2D, termed 2D TSA-tree, to enable efficient multi-level trend detection on 2D spatial data [8]. In order to construct 2D TSA-tree, we introduced two operations termed *split* and *merge*. *Split* is the operation that generates a multi-level tree, where each node contains the wavelet coefficients of the corresponding multi-level trends and surprises, while *merge* is the inverse operation of *split*. The *Split* operation can be defined using the $Down{-}Sampling$ and $Convolution$ operations along X-axis and Y-axis. A 2D TSA-tree is constructed by applying *split* operation on AX_i's repeatedly. This procedure repeats k times in order to construct a 2D TSA-tree with $k+1$ levels. Fig. 1 shows the structure of a general 2D TSA-tree. Original data is contained in the root node. An AX_i node (average values) contains information about trends, while D_1X_i (D-horizontal), D_2X_i (D-vertical) and D_3X_i (D-diagonal) are the detail values that contain information about surprises. Reversibly, for four equi-sized data AX_i, D_1X_i, D_2X_i and D_3X_i, *merge* operation can be applied to obtain the original data X from the average and detail values. The *merge* operation can be defined using the $Up{-}Sampling$ and $Convolution$ operations along X-axis and Y-axis. Note that the complexity of both *split* and *merge* operations is $O(nl)$ where n is the size of data and l is the length of filter.

3 Surprise Mining for Spatio-Temporal Data

While finding surprises in different directions was not of interest to the applications dealing with 2D spatial data, with spatio-temporal data, surprises in both dimensions become important. For example, consider an application that manages traffic flow of highways (i.e., traffic flow of highways at different locations are collected at different times). For any specific region, we can find the time that

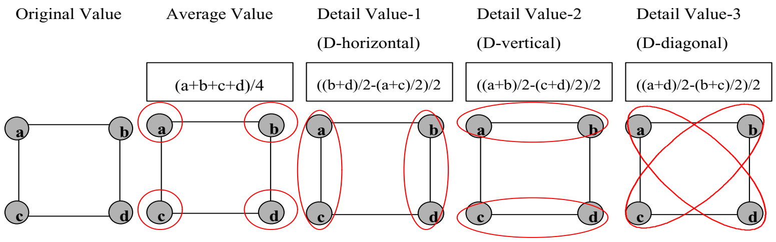


Fig. 2. Illustration of 2D wavelet transforms on 4 data points

the area has heavy/light traffic through detecting surprises in the time dimension. Similarly, at a specific time interval, we can identify the region or location that has heavy/light traffic through surprise in the spatial dimension. Consider another application that deals with atmospheric data at different heights at a fixed latitude/longitude location collected over different times. Again, the surprise through time dimension and spatial dimension are both interesting in this scenario.

In Sec. 3.1, we explain the basic concepts of surprises. In Sec. 3.2, we show how user can interact with 2D spatio-temporal data for surprise mining.

3.1 Basic Concepts of Surprises Mining

In this section, we show how the D-horizontal and D-vertical values of 2D TSA-tree can be utilized to identify two kinds of surprises, which are referred to as *horizontal* and *vertical* surprises, respectively. *Diagonal* surprise is of no interest to our applications and is not discussed in this paper any further. For the following discussion, assume that the spatio-temporal data set is defined as points in a 2D mesh. Without loss of generality, we also assume that time is X-axis and spatial dimension is Y-axis. First, we define cell and cover, respectively.

[**Cell (C)**] Each point (P) in the mesh is a cell by itself. A set of points is a cell only if they are used together by 2D wavelet transforms to compute a wavelet coefficient in some resolution.

Four floating points are associated with each cell, one for the average signal, and three for detail signals. Fig. 2 shows how a cell at level 1 can be obtained from 4 data points using 2D wavelet transforms. With 2D, each adjacent four points in a discrete plane can be replaced by their average value and three detail values. The detail values (D-horizontal, D-vertical, and D-diagonal) correspond to the average of the difference of: 1) the summation of the rows, 2) the summation of the columns, and 3) the summation of the diagonals. A cell at level i can be generated from average values of 4 adjacent cells at level $i-1$ using the same procedure. In Fig. 3, we show an example of points that are grouped as a cell. For example, a is not a cell since the points that are grouped are not used

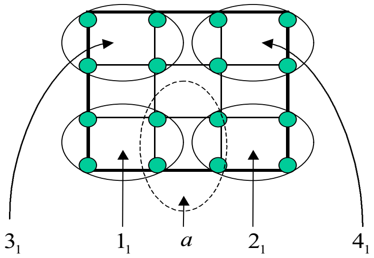


Fig. 3. Illustration of 2D cells

together to generate any wavelet coefficients. On the other hand, 1_1, 2_1, 3_1 and 4_1 are considered as cells at level 1.

[**Cover (CV)**] Let CV be a set of cells at level i, SA be a selected area and X be a set of all points inside SA, respectively. Then CV is a cover for SA at level i if and only if the following relation holds: ($\forall\ c \in CV$, if $p \in c$ then $p \in X$) and ($\forall\ p$, if $p \in X$ then $\exists\ c \in CV$, $p \in c$); where p is a point in a 2D mesh.

For example, in Fig. 4, $\{5_1, 6_1, 9_1, 10_1\}$ is a cover for $Selection{-}2$ at level 1 while there is no cover for $Selection{-}1$ (note that numbers inside the ellipses represent the cell numbers). Now, we provide definition for multi-levels surprises as follows:

[**Horizontal Surprise**] Given a 2D mesh, let CV be a cover for a selected area (SA) at level i. Then, *horizontal* surprise for a cell $C \in CV$ at level i is defined as the D-horizontal detail value of C. *Horizontal* surprise for SA is defined as a set of *horizontal* surprises for the cells in CV.

[**Vertical Surprise**] Given a 2D mesh, let CV be a cover for a selected area (SA) at level i. Then, *vertical* surprise for a cell $C \in CV$ at level i is defined as the D-vertical detail value of C. $Vertical$ surprise for SA is defined as a set of *vertical* surprises for the cells in CV.

The horizontal surprise at level i can be obtained from D_1X_i nodes of 2D TSA-tree while the vertical surprise at level i can be obtained from its D_2X_i nodes. Now, we show how horizontal and vertical surprises can be used for surprises mining. We assume that users can select a rectangular area for mining purpose if there exists a cover CV for that area. Hence, $Selection{-}2$, $Selection{-}3$ and $Selection{-}4$ are valid selections while $Selection{-}1$ is not (Fig. 4).

In Fig. 4, the temperature value $Temp(t, h)$ is the temperature at a specific height h and a particular time t on a fixed latitude/longitude grid. A sample query for $Selection{-}2$ can be "*Find the height interval where temperature has sudden changes within a specific time interval*", while for $Selection{-}3$ can be

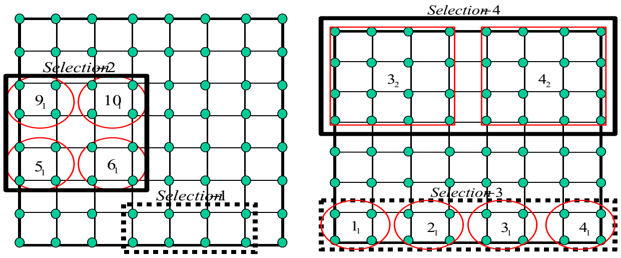


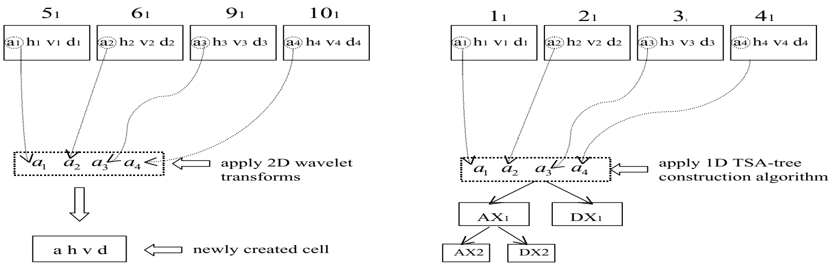
Fig. 4. Examples of selection areas

"Find the time interval when temperature has sudden changes within specific height interval over entire time period". The first query can be discovered by analyzing the vertical surprise while the second one can be answered through the horizontal surprise.

3.2 Mining Multi-Resolution Surprise

In a typical web-based environment, the client can be a GUI that provides the users with an interface for data analysis/visualization while a server contains scientific observation data. Since the human's eyes are restricted to distinguish tiny differences (which actually corresponds to the resolution of the screen), surprises at a low resolution might be visually good enough for visualization of surprises in a region, while the size of data used is much smaller than that of the original data (at a higher resolution). Hence, when the client requests to visualize the surprises for some large size of data, if the resolution of client's display is low or the user wants to see coarser surprises, it is possible to achieve good performance by sending much less amount of data while not sacrificing the user's visual impact.

2D TSA-tree has the capability of multi-resolution analysis in which the nodes of the tree can be immediately used to visualize surprises at different resolutions. The user can select a predefined sub-region (i.e., a cell) from a space. In such case, the query is answered using the information stored in the 2D TSA-tree nodes directly (e.g., D_1X_i's) and we can use nodes at different levels of the tree to obtain surprises at different resolutions. On the other hand, if the user selects an area that is defined by a cover (CV), the original 2D TSA-tree does not have the wavelet coefficients for the CV. However, it stores the wavelet coefficients for all the cells in CV. Therefore, to find the surprises for such area at different resolutions, the system has to compute new wavelet coefficients on-the-fly (i.e., create a customized 2D TSA-subtree). There are two cases we should consider when computing wavelet coefficients. First, consider the case when the shape of the cover at the lowest resolution (CV) is represented as a square. For simplicity, we assume that the size of the square corresponds to integral power of 4. If a user is interested in the horizontal/vertical surprises at the level greater



(a) *Selection-2*

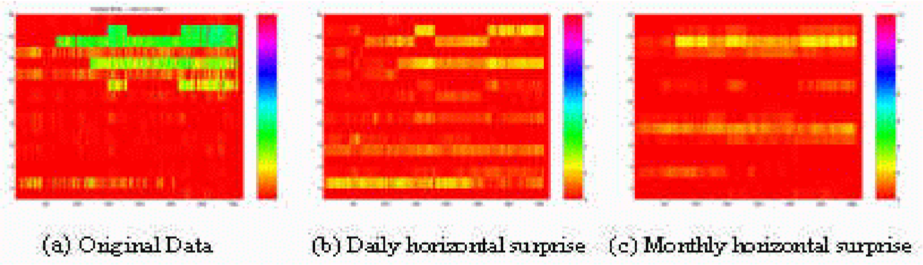
(b) *Selection-3*

Fig. 5. Computation of coefficients for a selected sub-region

than or equal to i, then we can use the values contained in the detail nodes. However, if the user wants the surprises at the resolution lower than i, then we need to compute new wavelet coefficients. To this end, we merge the average values of each cell in CV into one list and then apply 2D wavelet transforms to the list to obtain a new cell in the desired customized 2D TSA-subtree. For example, consider $Selection{-}2$ depicted in Fig. 4. If a user wants the vertical surprise at level 2, we should extract the average value of each cell at level 1 and apply 2D wavelet transforms to 4 cells (see Fig. 5(a)). A D-vertical detail value for the newly created cell (i.e., v) can then be used to visualize surprises for the selected region.

The second case is when a user submits a query to discover surprises within a specified period of time for different levels of abstraction (e.g., weekly, or monthly), but 2D wavelet transform is not applicable. That is, since the shape of the selected area is not a square, we cannot apply 2D wavelet transform to the selected area. In this case, we reformulate the data into a single 1D time-series. The stored *average* values in the cells (AX_i) of 2D TSA-tree are considered as the components to formulate the time-series. Subsequently, we apply 1D wavelet transform on the time-series data to obtain the required surprises. To this end, we use the 1D TSA-tree construction algorithm (see [7] for further details) to obtain the surprises. For example, with $Selection{-}3$, when horizontal surprise value at level 1 is H=(2 8 0 0), if a user submits a query like "*Find the horizontal surprise of the selected area*", then we can use H. Furthermore, if the user is interested in 1D surprise visualization, we consider the average values in the cells $\{1_1, 2_1, 3_1, 4_1\}$ as a time-series data set and apply 1D TSA-tree construction algorithm (see Fig. 5(b)).

The last thing we should consider is how to fetch those cells, which correspond to the selected area. In [8], we defined an index structure for 2D TSA-tree whose purpose is to group together those cells that are at the same level. For details, see [8].



(a) Original Data (b) Daily horizontal surprise (c) Monthly horizontal surprise

Fig. 6. Surprises at multiple levels of abstraction using Haar wavelet filter

4 Performance Analysis

Our application can be defined as visualizing surprises for some selected areas of interest to users at different resolutions when data for the entire region is available. We conducted some experiments to evaluate the above and the results are shown in Sec. 4.2

4.1 Experimental Setup

For all the experiments, we used both real-life data obtained from our NASA sponsered GENESIS project and synthetically generated data. With real data, ground station receivers of GPS signals are put along the San Andreas fault in Southern California and their height variations from day to day over one year are recorded. The large height variation (i.e., surprises in our case) can indicate possible earthquakes.

To generate synthetic data, we employed Burger's equation. The Burger's equation is used to model street traffic flow [1]. We used Burger's equation because it can generate several temporal surprises in finite steps while preserving the spatial correlation. Let $u(x,t)$ be the density of cars at location x and time t. Then, the following represents Burger's equation:

$$u(x,0) = f(x) \tag{1}$$

$$\frac{du}{dt} = -u\frac{du}{dx} \tag{2}$$

In (1), $f(x)$ is one-dimensional spatial data for initial condition. We then use a discrete version of the partial differential equation to create a time series (i.e. create a two dimensional data).

4.2 Visual Verification of Surprises Mining

Fig. 6 depicts horizontal surprises of real data at different levels using Haar wavelet filter. X-axis represents days in 1999 and Y-axis represents the ground

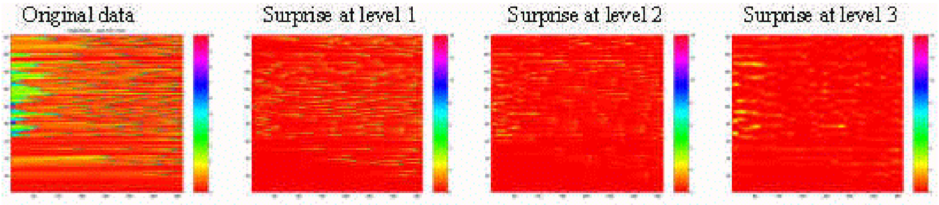


Fig. 7. Horizontal surprises of synthetic data using db6 wavelet filter

stations. In order to obtain daily and monthly surprises, we use the nodes of 2D TSA-tree which are close enough to query aggregation level (i.e. nodes at level 1 and level 5). Fig. 6 illustrates the existence of horizontal surprises. Those horizontal surprises localized temporally, may indicate the occurrences of earthquakes during the corresponding time interval. We performed cross-validation in order to verify our observation with real data on occurrences of earthquakes[2]. For example, a monthly surprise is detected from ground station 14 (which is located at 34.1^o in latitude and 243^o in longitude) in November. Cross-validation with SCEDC data shows that earthquakes indeed occurred very frequently at this location during this month. That is, for the area of station 14, although 2 earthquakes occur within a typical month, there were more than 8 earthquakes within November. Meanwhile, as depicted in Fig. 6(b), even though daily surprises are detected for ground station 5 (which is located at 34.5^o in latitude and 239.3^o in longitude), no monthly surprise is detected in this area since the number of earthquakes per month was never significantly more than 1 (which is the typical number of earthquakes in the area of station 5). This verifies our multi-level surprise abstraction property: sudden changes in a day or two are important for daily surprises while they can be ignored when considering monthly surprises. Similar results are observed in Fig. 7 for synthetic data. Again, more details are captured by the higher levels (e.g., level 1), while lower levels (e.g., level 3) are abstract. At higher level, since the size of scale is small, small changes are considered as surprises. However, as we go to lower levels, small changes are considered as noises since the size of scales become large. Even though surprise at higher level has more details while surprise at lower level is more abstract, we can observe that the general surprises for all levels are the same.

Different wavelet filters have different extent of smoothness and complexity. Thus, users can choose different wavelet filters for mining purposes depending on how much smoothness and complexity their applications desire. If the display resolution of the client's monitor is low, one can send much less amount of data while not sacrificing the user's visual quality. For example, as shown in Fig. 7, even though the user requests surprise at level 2, surprise at level 3 might be

[2] The data on occurrences of earthquakes (used for cross-validation) were obtained from SCEDC (Southern California Earthquake Data Center) at http://www.scecdc.scec.org.

visually good enough while its size is 1/4th of the requested data. Thus, it is necessary to enable users to provide the size of the scale that they can tolerate.

5 Conclusion

We presented techniques and data structures vital to applications that can benefit from visualizing surprises in different directions and at different levels of abstractions. We explained how the detail values (D_1X_i, and D_2X_i) that are stored in the nodes of 2D TSA-tree can be utilized to support multi-level surprise queries (horizontal and vertical surprise) on 2D spatio-temporal data. By precomputing different levels of 2D TSA-tree, we can answer multi-level surprise queries without any reconstruction overhead. In addition, we discussed how 2D TSA-tree can be used to support two different types of user interactions for 2D spatio-temporal surprise mining. Our experimental results indicate that 2D TSA-tree is suitable for visualization of surprises at multiple levels of abstraction. However, the results provided in this paper are at preliminary stage. Although it shows our approach is promising, more experiments need to be conducted to correspond real-world changes with our definitions of surprises.

6 Acknowledgement

We would like to thank Dr. George Hajj and Dr. Brian D. Wilson from JPL for providing the data used for our experiment. We would also like to acknowledge Rolfe R. Schmidt for suggesting the Burger's equation to generate synthetic data.

References

1. U.M. Ascher, S.J. Ruuth, and B.T.R. Wetton.: Implicit-explicit methos for time dependent partial differential equations. SIAM J. Numer. Anal. (1995) 32:797-823.
2. K. Chan and A. W. Fu.: Efficient time series matching by wavelets. ICDE. (1999) pages 126-133.
3. E. M. Knorr and R. T. Ng.: Algorithms for mining distance-based outliers in large datasets. VLDB. (1998) pages 392–403.
4. E. M. Knorr and R. T. Ng.: Finding intensional knowledge of distance-based outliers. VLDB. (1999) pages 211–222.
5. P. Raghavan A. Arning, R. Agrawal. A linear method for deviation detection in large databases. KDD. (1996) pages 164–169.
6. G. Sheikholeslami, S. Chatterjee and A. Zhang. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. *VLDB*, pages 428-439, 1998.
7. C.Shahabi, X.Tian, and W.Zhao. TSA-tree: A Wavelet-Based Approach to Improve the Efficieny of Multi-Level Trend and Surprise Queries. SSDBM. (2000) pages 55-68.
8. C.Shahabi, S. Chung, M.Safar and G.Hajj. 2D TSA-tree: A Wavelet-Based Approach to Improve the Efficieny of Multi-Level Spatial Data Mining. Technical Report 01-740, Department of Computer Science, University of Southern California. (2001) (URL: http://www.cs.usc.edu/tech-reports/technical_reports.html)