Hair Data Model: A New Data Model for Spatio-Temporal Data Mining

Abbas Madraky, Zulaiha Ali Othman, Abdul Razak Hamdan

Data Mining and Optimization Research Group (DMO), Centre for Artificial Intelligence Technology (CAIT)

School of Computer Science, Faculty of Information Science and Technology

Universiti Kebangsaan Malaysia (UKM), Malaysia

{madraky,zao,arh}@ftsm.ukm.my

***Abstract*—Spatio-Temporal data is related to many of the issues around us such as satellite images, weather maps, transportation systems and so on. Furthermore, this information is commonly not static and can change over the time. Therefore the nature of this kind of data are huge, analysing data is a complex task. This research aims to propose an intermediate data model that can represented suitable for Spatio-Temporal data and performing data mining task easily while facing problem in frequently changing the data. In order to propose suitable data model, this research also investigate the analytical parameters, the structure and its specifications for Spatio-Temporal data. The concept of proposed data model is inspired from the nature of hair which has specific properties and its growth over the time. In order to have better looking and quality, the data is needed to maintain over the time such as combing, cutting, colouring, covering, cleaning etc. The proposed data model is represented by using mathematical model and later developed the data model tools. The data model is developed based on the existing relational and object-oriented models. This paper deals with the problems of available Spatio-Temporal data models for utilizing data mining technology and defines a new model based on analytical attributes and functions.**

Keywords-hair data model; spatio-temporal data models; data warehouse model.

I. INTRODUCTION

Spatio-Temporal databases deal with data types are categorized by both spatial and temporal concepts. The Spatio-Temporal data consists two specifications. The first one describes the data in each position. For example in a weather map, each position has some information such as temperature, humidity, pressure, etc. These values are allocated to specific position. The second specification describes the data that related to time. For instance, values of temperature, humidity and pressure change during the time. In the other word, data values are changeable by passing time. The first specification is known as spatial data and the second is known as temporal data. The Spatio-Temporal data consists of these two kinds of data which related to each other.

These huge sets of data often hide interesting information which conventional systems and classical data mining techniques are unable to discover. New concepts and methods are needed to extract more complete and detailed information from the vast repositories of Spatio-Temporal data that are accumulating. Numerous articles in the field of Spatio-Temporal data analysis show that the topic has a great significance. Two main issues are important while mining the data. The first is on modelling the data and the second is the method to discover interesting, useful, and significant patterns from the data [1]. So, presenting new data models and methods for data analysing is absolutely essential.

In proposing a Spatio-Temporal model, it is natural to extend existing spatial data models with time. Throughout the relatively young history of research on Spatio-Temporal modeling, a substantial number of models have been presented in the last two decades and many Spatio-Temporal data models and corresponding query languages have been proposed. The main Spatio-Temporal data models generally include the following: Space Time [2][3][4], Snapshots[5], Simple Time-Stamping [6], Base State with amendments [7][8], Spatial Temporal Domain[8], Feature-based spatial-temporal data model [9], Event Based Spatio-Temporal Data Model [10], History Graph Model [11], Spatio-Temporal Entity Relationship (STER) Model [12], Object-Relationship (O-R) Model [13].

There are some problems in available Spatio-Temporal data models. The first, analytical process to be difficult in Spatio-Temporal information systems because there are few analytical attributes or functions in these models. The available analytical operations are limited and it is needed to use more analytical actions depending on information of each data group. The second problem is about using the object-oriented features in data models. Most of data models are not suitable for defining an object oriented system because they are designed based on relational models. The third, due to use complex data types, the user-friendly features in these models are weak and the users do not dominate to utilize them. These problems will be explained in the following paragraphs:

A. *Difficulties in analytical operations*

Users would like to analysis data values by using data characteristics in many data preprocessing tasks. There are few analytical parameters as a structural field in data warehouse models. For applying data mining technologies in spatial temporal systems some parameters are useful for analysing data mining results. Such of them are importance, orientation and neighbourhood of data.

978-1-4673-2718-3/12/$31.00 ©2012 IEEE

Data importance is about necessity of information for task success. For instance, in data analysing of gas wells, information of the main wells are more important. Orientation is useful for finding the answer of a query that maybe is not exactly in the stored data. For example, for finding gas zones in a geographic area some information about exploratory wells able to detect gas reservoirs. It also depends to neighborhood of data in the region. In gas case study, specifications of each well are related to other wells specially the neighbourhood wells.

In addition, data warehouse models have a few functions. These functions are roll-up, drill-down, slicing, and dicing [14]. The other functions are defined in application programs and they don't have a common usability for analyzing. Separation between functions and data during transferring, removal and update analytical information will cause problems. Some security functions are failed after changing the location of data. So, data warehouses need to have more functions to process and protect the information. For example, if we want to know about main fields in the gas exploration zone, it should be done by using separate programs for analysing exploration gas wells so there are not any functions inside database or data warehouse structure for finding data dependencies or defining their arrangements.

B. Poor compatibility with Object-Oriented Environments

Due to structural criteria, relational data models don't utilize the object-oriented features completely. Because, unlike the object oriented approaches, in relational data models, the data and functions are not bound together so it is needed to have various parameters and functions for data analysing with different objectives. Incomplete deployment of object oriented concepts is considered as a limitation in an environment. This limitation will be more egregious while we have more complex data types. In Spatial and temporal systems, data types and functions are more complex. However, the object-oriented model is suited to store and retrieve complex data type especially Spatio-Temporal data types. For example, graphical results are essential for image processing in the geological layers about recognition of gas reservoirs.

C. The user-friendly weakness

Nowadays, the software producing companies give to users some unclassified design information for more customers attracting. It causes that users have more identity about software abilities and the system. The most data models in Spatio-Temporal systems are not able to identify by users because the Spatio-Temporal data types are more complex. So, the users could not inform about facilities and features in data models or predict the model behaviour in variant situations.

The rest of this paper is organized as follows. In section II we present a new model which is categorized as structure and specifications.The new analytical tasks are also explained in this Section. In section III we define the mathematical relationships between analytical parameters and analytical functions. In Section IV we consider the contributions and benefits of the new model. Finally, in section V we discuss about future works for improvement and further studies.

II. THE MODEL DEFINITIONS

Human beings are interested in to inspire from nature for creating tools for meeting their requirement. This desire has also been included into design structures and methods. For instance, we can consider flying or moving underwater as a sensible case. Whenever, a pattern in nature is used to perform some job, there is fewer problems occurred in the implementation and applications. This is because in nature the structure and methods are coordinated during the time passing. Computer designing, manufacturing and improving are based on human structure and its performance. Artificial intelligence, artificial neural networks, robotics and many algorithms have inspired from human structures and thoughts.

As mentioned in the introduction section, Spatio-Temporal data models have two properties. These data models should be able to preserve spatial data and they could also consider a sequence or time stamp for data representation. The hair also has these specifications. Each hair is related to a location and it growths by the time sequent.

In hair data model, a hair is a set of data about specific location. Basic definitions of each hair are stored in the root or data catalog. These definitions are divided to two categories. Data definitions specify structure's information (type), location (position) and some functions about maintenance and security. The second category is about analytical definitions include the importance of data (strength), neighborhood of data and orientation (direction) as an analytical attributes and some analytical functions. These parameters explain the specifications and qualities of data. Like the natural system, these parameters can be changed due to spending time or different application. In figure 1 we determine these specifications and corresponding features in hair natural structure.

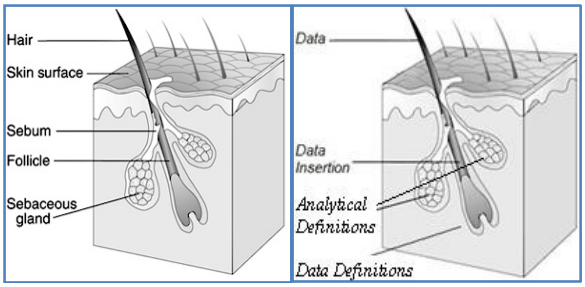


Figure 1. Compare of Hair natural structure and Hair Data Model

Furthermore, Analytical tasks are defined based on hair natural specification. Each task uses the data values and analytic parameters and changes them if it is needed. Some tasks based on natural specification of hair are showed in Table I. The first part of tasks is about insertion or deletion of data like Growth, Cutting, Falling and Planting. In this model data are inserted from the root side. So, the new data are near to root and the old data are in the end of the hair. By using Growth function the new data are inserted to hair and by using Cutting the old data are removed. Over time, the importance of data (or strength of the hair) may be changed and the data erased by using Falling function. The Planting function creates a new set of data and definitions. This function is opposite of Falling function.

The second part of tasks is about analysing. Combing and Plaiting functions are used for analytical processing. By using Combing, the orientation or classification of data is changed and we can define multi usability for a part of data. We can also arrange the data by this function. Routing in a GIS environment and finding the answer that is not exactly in stored data are some advantages of this function. Plaiting is used for clustering or grouping a set of data. Obviously, this grouping is dynamic and can be variable depending on the request. Each plate has analytical parameters such as orientation or plait neighborhood

TABLE I. HAIR NATURAL SPECIFICATIONS AND HAIR DATA MODEL FUNCTIONS

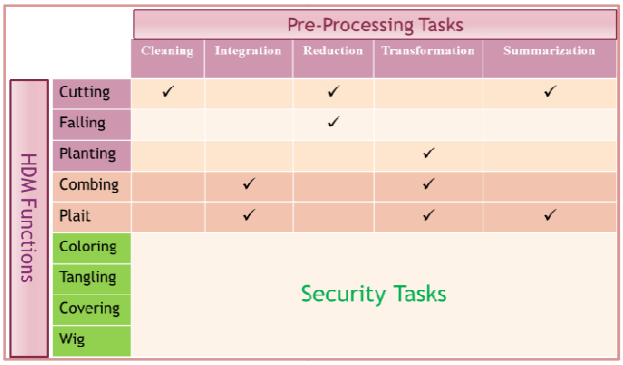
Row	Function Name	Specifications in the Real Mode	Specifications
1	Growth	To Increase the length of hair body	To insert data to hair from the root side.
2	Cutting	Shorten the length of hair.	To remove of unnecessary redundant or useless information without structure modification. Summarization also is done by this task.
3	Falling	Hair loss caused by the time or illness.	To erase data and definitions by weakening importance over time.
4	Planting	A surgical technique for increasing the number of hair	To create of a new set of data and definitions and it is inserted into a specific location.
5	Combing	Arrange the hair by the hair comb.	To specify orientation of Information based on specific application. Classification or Clustering of neighbored data is done by Combing.
6	Plaiting	A plait is a complex structure by intertwining three or more strands of human hair.	Used for Data grouping. Each plate consist integrated data.
7	Colouring	the practice of changing the color of hair.	To change some attributes of data for better data presentation or security without changing data definition.
8	Tangling	Messy and disheveled hair.	To convert data to sloppy case for preserving security. (Opposite Combing)
9	Covering	To cover the hair by veil, hat and etc.	To protect data from unauthorized accessing or damage.
10	Wig	A wig is a head of hair made from synthetic materials, wool, human hair or …	To use non-real data for impossibility of identifying the main data by unauthorized accessing.

The third part of tasks includes Coloring, Tangling, Covering and Wig functions are defined for preserving security. Coloring doesn't create any changes in structure or data definitions but this function changes some attributes of data for virtual presentation In some cases for achieving more security the regularity of data would be reduced. Tangling function eliminates the arrangement of data. This function is opposite of Combing function. Covering built a protective layer for important data so that unauthorized person or software does not access the data. Number of protective layers can be more than one layer. The last security function is Wig. This function produces non real data and stores them like the main data. But the authorized user can recognize the real data and use it.

- We mentioned that some pre-processing is needed for converting raw data stored in databases into information in data mining system. The main tasks and their specifications are:
- Cleaning: To remove noise and correct inconsistencies in the data.
- Integration: To merge data from multiple sources into a coherent data store.
- Reduction: To reduce the data size by aggregating, eliminating redundant features, or clustering.
- Transformation: To improve the accuracy and efficiency of mining algorithms.
- Summarization: To create an overall picture of data.

The mapping of hair data model functions and pre-processing task is illustrated in Table II.

TABLE II. MAPPING PRE-PROCESSING TASKS &HAIR DATA MODELFUNCTIONS



III. MATHEMATICAL MODEL

Analytical parameters are identified based on their application in different fields. Relations of components also are identified. Analytic functions are defined inspired by natural model and its mapping to methods, functions and tools in data mining systems will be determined. A preliminary proposed data model is represented using basic mathematic expression as illustrated bellows:

This model is a interface between database and knowledge base. Hair Data Model (HDM) consist several data sets which named Hair. 'n' is the number of hair in HDM.

$$HDM = \{H_1, H_2, \ldots H_i \ldots H_n\} \quad (1)$$

Each hair includes two parts. R or root indicates definitions and B or body is about data which stored in the data model.

$$H=(R, B) \quad (2)$$

R consist two set of definition which are attributes and functions. Attributes determine data properties and functions define the tasks. m is the number of attributes and k is the number of function in root of hair.

$$R=\{A_1, A_2,\ldots A_i.\ldots A_m\}, \{F_1, F_2,\ldots F_i.\ldots F_k\} \quad (3)$$

Body of hair includes several cells. Each cell consist some values related to attributes that are defined by R.l is the number of cells in a Body and o is the number of values in a Cell.

$$B=\{C_1, C_2,\ldots C_i.\ldots C_l\} \quad (4)$$

$$C=\{V_1, V_2,\ldots V_i.\ldots V_o\} \quad (5)$$

We define two types of attributions. The first one is about definition attributes that define the structure of cells values, their types, domains and positions. The second part of attributes is analytic. These attributes determine the analytic parameters such as importance of values, orientation and neighborhood. The | sign is used for 'or' operator. For example each attribute (A) could be considered as a definition attributes (DA) or analytical attributes (AA).

$$A=(DA|AA) \quad (6)$$
$$DA=(Type, Position)$$
$$AA=(Strength|\ Direction|\ Proximity)$$

$$Strength\equiv Importance\ of\ data$$
$$Direction\equiv Orientation\ of\ data$$
$$Proximity\equiv Neighborhood\ of\ data$$

The functions in this model are divided to three categories. Some functions used for performing a query (QF) such as insertion, deletion and etc. Thesecond category is about analytical functions (AF) that related to data mining tasks and the last category is devoted to security functions (SF).

$$F=(QF|AF|SF) \quad (7)$$

Some functions based on Table 1 are bellows:

$$QF=(Growth|Cutting|Falling|Planting) \quad (8)$$
$$AF=(Combing|Plaiting)$$
$$SF=(Coloring|Tangling|Covering|Wig)$$

One cell is inserted to hair by The Growth Function. The cell adds to hair from the root side so sequence of insert is based on time sequence.Previous_B illustrate the state of body before insertion and Current_B is used for the state after insertion.

$$Growth(H,C): Previous\_B \rightarrow Current\_B \quad (9)$$
$$Previous\_B, Current\_B \in B$$
$$Current\_B=Previous\_B+C$$

The Cutting function is opposite of Growth function. This function deletes a cell from a hair. Previous_B illustrate the state of body before deletion and Current_B is used for the state after deletion.

$$Cutting(H,C) : Previous\_B \rightarrow Current\_B \quad (10)$$
$$Previous\_B, Current\_B \in B$$
$$Current\_B=Previous\_B-C$$

Falling and Planting functions are the rest of query functions which are about deletion and insertion a hair. Falling erase a hair with attributes and cells and Planting define a new set of attributes in a new hair.

$$Falling(H): Delete(H) \quad (11)$$
$$Planting(H): New(H)$$

Combing function define a direction for hair. Direction is an analytical attribute that determines the orientation of data. By orientation we can define the data groups. Data classification could be defined by this function.

$$Combing:H(Old\_Direction) \rightarrow H(New\_Direction) \quad (12)$$

Plating is another analytic function for clustering a set of data. This function defines a group of data and consists a hair to a group.

$$Plaiting(H,Group): H \in Group \quad (13)$$

Coloring, Tangling, Covering and Wig functions are security functions in hair data model. The definitions of these functions are stated in section II. All of these functions convert the state of data model depend on user identification. Security properties are different according to user access level. The definition of security functions are bellows:

$$Coloring:H(Old\_Properties) \rightarrow H(New\_Properties) \quad (14)$$
$$Tangling:H(Old\_Direction) \rightarrow H(Random\_Direction)$$
$$Covering:H(visible) \rightarrow H(Unvisible)$$
$$Wig: Real\_HDM \rightarrow Non\_Real\_HDM$$

IV. Novelty & Contributions

This paper proposed a new data model for Spatio-Temporal environment inspired by maintaining quality natural hair which consists of conditions and properties of the hair.

Some advantages of the proposed model are as follows:

- To define and store a group of cleaned data for ease of access and sorting. This ability is frequently used in GIS.
- Ability to allocate a different and independent structured attributes such as importance, orientation and neighborhood to each group of data. This attributes define the analytical information about stored values.
- Ability to define flexible data application by using orientation property.
- Better understanding for model definitions and behavior similar to other models or algorithms which are inspired from the nature.
- To define the common analytical parameters for Spatio-Temporal data can be reused.

- To define more security functions for preserving data from unauthorized access. Because discovered information or knowledge are more important to protect by security tools.

V. FUTURE WORKS

This paper is about designing of a proposed model for using data mining technology. It is desired to implement hair data model in a specific field by using the tools and programming modules. It will be evaluated by test data. After identifying model problems, the model will be revised by redefining components and functions. After implementation and evaluation, model strengths and weaknesses are identified based on performance parameters including integrity, time costs and other metrics of data analysis. Finally, the model should be improved and optimized by comparison with other analytical models in the various environments.

REFERENCES

[1] V. Bogorny, Sh. Shekhar**, "**Spatial and Spatio-Temporal Data Mining**",** IEEE International Conference on Data Mining,2010

[2] Hägerstrand, T, "What about people in Regional Science?",Papers of the Regional Science Association Volume 24, Issue 1, December 1970, Pages 6-21

[3] Szego, J., "Human cartography: mapping the world of man", Central Board for Real Estate Data, Gavle, Sweden, 1987

[4] X.Wu, Zh. Zhang, X. Hao, Z. Su, "Research on the feature and uncertainty based Spatio-Ttemporal data model" 2nd International Workshop on Intelligent Systems and Applications (ISA), 2010

[5] Ross, C.a, Guensler, R.b, Stevens, "Spatial and statistical analysis of commercial vehicle activity in metropolitan Atlanta",Transportation Research Record, Issue 1625, 1998, Pages 165-172

[6] Gary J. Hunter and Ian P. Williamson, "The Development of a Historical Digital Cadastral Database", Int. Journal of Geographic Information Systems, 4(2), 1990.

[7] Langran, G., "Tracing temporal information in an automated nautical charting system",Cartography & Geographic Information Systems. Volume 17, Issue 4, 1990, Pages 291-299

[8] Peuquet, D.J.,"It's about time: a conceptual framework for the representation of temporal dynamics in geographic information systems", Annals - Association of American Geographers, Volume 84, Issue 3, 1994, Pages 441-461

[9] Chu, W.W, leong, I.T.b, Taira, R.,"A semantic modeling approach for image retrieval by content", The VLDB Journal, Volume 3, Issue 4, October 1994, Pages 445-477

[10] D. Peuquet and N. Duan, "An Event-Based Spatio-Temporal Data Model (ESTDM) for Temporal Analysis of Geographical Data", Int. Journal of Geographical Information Systems, vol. 9, no. 1, pp. 7-24, 1995..

[11] A. Renolen, "History Graphs: Conceptual Modeling of Spatio-Temporal Data", In GIS Frontiers in Business and Science, Vol. 2, International Cartographic Association, Brno, Czech Republic, 1996.

[12] T. Hadzilacos and N.Tryfona, "An Extended Entity-Relationship Model For Geographic Applications", ACM SIGMOD Record, 26(3), 24-29., 1997.

[13] C.Claramunt, C. Parent, S.Spaccapietra, M. Theriault, "Database Modeling for environmental and Land Use Changes", Geographical Information and Planning, Chapter 20, Springer-Verlag, 1998.

[14] J.Han, M. Kamber, "Data Mining: Concepts and Techniques", University of Illinois at Urbana-Champaign, 2006.

[15] C.Claramunt, C. Parent, S.Spaccapietra, M. Theriault, "*Database Modeling for environmental and Land Use Changes*", Geographical Information and Planning, Chapter 20, Springer-Verlag, 1998.

[16] J.Han, M. Kamber, "*Data Mining: Concepts and Techniques*", University of Illinois at Urbana-Champaign, 2006.