SENSOR DATA MINING TECHNIQUES AND MIDDLEWARE STRUCTURE FOR USN ENVIRONMENT

Cheng Hao Jin, Yongmi Lee, Hi-Seok Kim*, Gouchol Pok**, Keun Ho Ryu
Dept. of Computer Sciience, Chungbuk National University, Korea
*School of Electronics and Information Engineering, Cheongju University, Korea
**Yanbian University of Science and Technology, China
{kimsungho, ymlee, khryu}@dblab.chungbuk.ac.kr
*khs8391@cju.ac.kr
**gcpok2000@gmail.com

ABSTRACT ... With advances in sensor technology, current researches on the pertinent techniques are actively directed toward the way which enables the USN computing service. For many applications using sensor networks, the incoming data are by nature characterized as high-speed, continuous, real-time and infinite. Due to such uniqueness of sensor data characteristics, for some instances a finite-sized buffer may not accommodate the entire incoming data, which leads to inevitable loss of data, and requirement for fast processing makes it impossible to conduct a thorough investigation of data. In addition to the potential problem of loss of data, incoming data in its raw form may exhibit high degree of complexity which evades simple query or alerting services for capturing and extracting useful information. Furthermore, as traditional mining techniques are developed to handle fixed, static historical data, they are not useful and directly applicable for analyzing the sensor data. In this paper, (1) describe how three mining techniques (sensor data outlier analysis, sensor pattern analysis, and sensor data prediction analysis) are appropriate for the USN middleware structure, with their application to the stream data in ocean environment. (2) Another proposal is a middleware structure based on USN environment adaptive to above mining techniques. This middleware structure includes sensor nodes, sensor network common interface, sensor data processor, sensor query processor, database, sensor data mining engine, user interface and so on.

KEY WORDS: sensor data mining, USN middleware, sensor stream data.

1. INTRODUCTION

USN (ubiquitous sensor network) is drawing a lot of attention as a method for realizing a ubiquitous society. USN affixes sensor nodes to everything at anywhere and collects situational and environmental information. By doing this, it provides us more convenient and safer life. With fast development in sensor nodes and sensor network techniques, we can collect data from sensor nodes so that can speed up the implementation of ubiquitous computing life. Now, there are many kinds of USN application domain: physical distribution, circulation, environment, disaster prevention, traffic, home network, automation, security and so on.

Data collected from the sensor nodes in USN environment application is a kind of stream data, hence it has the characteristics of stream data as follows: continuity, expiration, infinity, time-sensitivity, approximation, adjustability. For sensor data application, the speed of incoming data is very fast and the arrival rate is not constant, also the volume of data is usually too huge to be stored on permanent devices or to be scanned thoroughly for more than once.

In such sensor data stream, there may be some hidden pattern or other useful information that we can't directly get from simple queries. However, because of the characteristics of sensor data, traditional data mining techniques which use finite and statistical data sets can't be directly applied to sensor data. So in order to extract such useful information, we need to research new data mining techniques which are fit to the sensor data.

In the future, with emergence of much more complex USN applications, we need a USN middleware which can analyze the sensed data to generate the query result and provide sensor data mining techniques to extract useful information. So in this paper, we propose (1) sensor data mining techniques which are fit to USN environment application and (2) USN middleware structure that can provide above sensor data mining techniques.

The rest of this paper is organized as follows. In section 2, we review related work. The classification of the sensor data mining techniques and their detail definitions, which are fit to USN environment applications, are described in section 3. In section 4, we present our proposed USN middleware structure. Finally, our conclusion and future work is given in section 5.

2. RELATED WORK

In recent years, with fast development in communication device and sensing technique, it enables us to collect the sensing data in real time. However, much research is limited in monitoring fields focused on collecting data and algorithms and query framework related to the applications such as web click and traffic monitoring are mainly developed. Therefore, there is nearly no research about sensor data mining techniques which are fit to USN application.

MobiMine[Kargupta et al., 2002] is the project of University of Maryland Baltimore County and it is called the first ubiquitous data stream mining system. MobiMine is a distributed data mining environment that allows "intelligent" monitoring of stock market data from mobile devices. It facilitates the monitoring process by identifying the interestingly behaving stocks and detecting their causal relationship with different features characterizing the stocks. By doing this work, it draws the attention of the user to time critical "interesting" emerging characteristics in the stock market. But because of the limitation of mobile devices such like low storage, battery that the traditional data mining techniques which reduce the disk I/O are not fit in mobile environment.

Much research about stream data has been widely studied over the last decade and many efficient algorithms have been developed and many prototype systems for stream query processing have been built, such as STREAM, Cougar, Aurora, Hancock, Niiagara, OpenCQ, Telegraph, Tradebot, Tribeca, etc. However, based on the best our knowledge, there is no existing stream data mining system or system prototype that can integrate multi-dimensional OLAP stream data analysis with long-term multi-resolution stream book-keeping and perform multiple stream data mining functions. So NCSA and department of Computer Science, the University of Illinois at Urabana-Champaign developed MAIDS[Cai et al, 2004] system that can perform stream classification, clustering, frequent pattern analysis and so on. MAIDS system can be applied to network intrusion detection, credit card flaw prevention, and web click streams analysis and so on.

3. SENSOR DATA MINIING TECHNIQUES

Considering time aspect, we can classify sensor data mining techniques into 3 classes: sensor data outlier analysis, sensor data pattern analysis and sensor data prediction analysis. Further detailed classification is presented in figure 1. In snapshot aspect, there are simple outlier analysis, spatial outlier analysis and spatial sensor pattern. Considering the distance between time point and time point, as mining techniques of time interval and sliding window, there are time sensor pattern analysis, continuous outlier analysis, continuous pattern analysis and prediction based on pattern analysis.

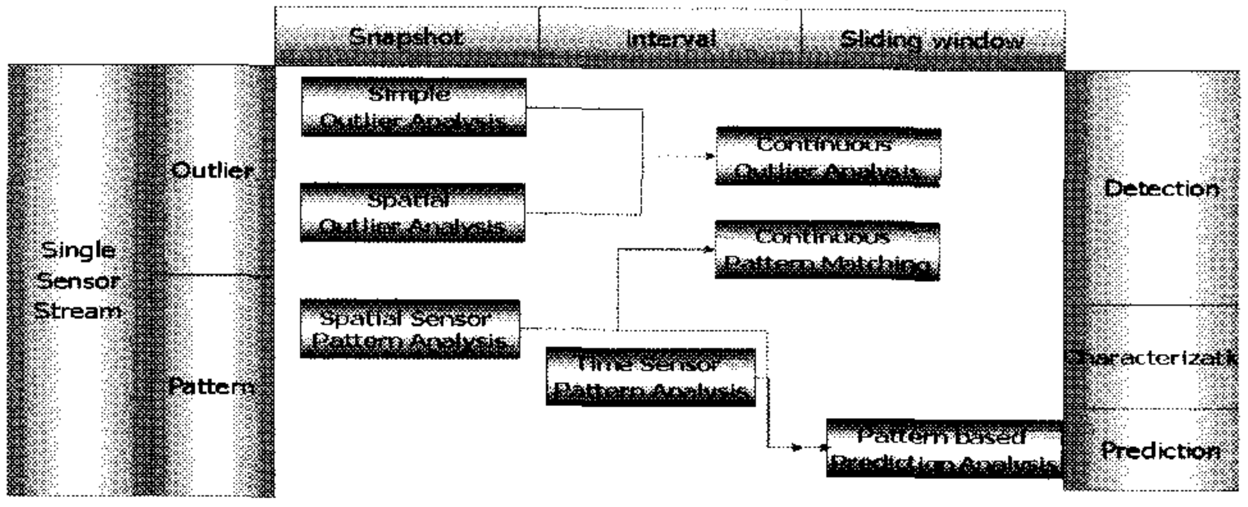
Snapshot
Interval
Sliding window
Single Sensor Stream
Outlier
Pattern
Simple Outlier Analysis
Spatial Outlier Analysis
Spatial Sensor Pattern Analysis
Time Sensor Pattern Analysis
Continuous Outlier Analysis
Continuous Pattern Matching
Pattern based Prediction Analysis
Detection
haracterizati
Prediction

Figure 1. The classification of sensor data mining classification

To give the details of their classification and definition, we choose ocean environment, one of the USN environment applications, and describe them as below.

3.1 Sensor Data Outlier Analysis

Outlier[Subramaniam et al., 2006] is a sensor data object that does not comply with the general behaviour of the data, though included in database. It can be considered as noise or exception, but is quite useful in some applications.

3.1.1 Simple Outlier Analysis: simple outlier is a sensor data sensed at a specific time point with characteristics that are considerably different than most data set collected from sensor networks.

In ocean environment, assume there is a requirement that "find out the area where its seawater density is quite different than usual today." We can see today as a specific time point and get the result using simple outlier analysis. In the sea, oil leakiness may cause serious ocean pollution, so we can prevent the occurrence of it by doing simple outlier analysis.

3.1.2 Spatial Outlier Analysis [Lu et al.,2003]: A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from the values of its neighbourhoods. Identification of spatial outliers can lead to the discovery of unexpected, interesting and useful spatial patterns for further analysis. The main difference between spatial outlier and simple outlier is that in spatial outlier analysis, we consider the location property.

In ocean environment, assume there is a requirement that "In the East Sea, classify the area where its change of seawater temperature is quite strange." Here, East Sea is the location property and we can use this property to do spatial outlier analysis. In ocean, the temperature distribution is different between each different area. Most fishes take action to seawater temperature and move their habitat even though there is a small range of change in seawater temperature. So if we know in which place what kinds of fishes live is very useful to improve our water resources.

3.1.3 Continuous Outlier Analysis: Continuous outlier is an outlier detected from sensor networks not at a specific time point but during a time interval.

In ocean environment, assume there is a requirement that "find the phenomenon which lasts 30 days with the temperature of seawater 2~10 ℃ higher than the average temperature." Here, we can consider 30 days as a time interval and apply continuous outlier analysis to this example. El Nino phenomenon lasts some time and by doing continuous outlier analysis, we can detect and predict its occurrence and reduce the loss from damage.

3.2 Sensor Data Pattern Analysis

Pattern is a behaviour that repeats periodically in a specific form. In USN environment, sensor pattern can be divided into two classes. One is time sensor pattern and the other is spatial sensor pattern.

3.2.1 Time Sensor Pattern Analysis: Time pattern is an iteration or trend in data collected from sensor networks during time intervals.

In ocean environment, assume there is a requirement that "find out the red tide occurrence pattern when the range of DO concentration change is above 3mg/L." Using time pattern analysis, we can find the pattern of red tide because when DO concentration is below 3mg/L, the probability of occurrence of red tide is very high. There is no much oxygen when red tide occurs, so fish can't breathe and then die. It causes lots of damage to us, so know its occurrence pattern is very necessary.

3.2.2 Spatial Sensor Pattern Analysis: Spatial sensor pattern analysis considers the corresponding sensor node's location property in the data collected from sensor networks.

In ocean environment, assume there is a requirement that "find the distributed pattern of red tide using the data of seawater change from the east part of the East Sea to the west part of the East Sea." If red tide occurs, may cause a lot of fish's death. Hence, using both east and west of the East Sea location properties to do spatial sensor pattern analysis can get the area where red tide always takes place. So spatial sensor pattern analysis helps us reduce loss from red tide.

3.3 Sensor Data Prediction Analysis

Predictions are probabilistic estimates of future occurrences or trend based upon many different estimation methods, including past patterns of occurrence and statistical projections of current data. In USN environment, prediction can be divided into two classes. One is continuous pattern matching and the ohter is prediction based on pattern.

3.3.1 Continuous Pattern Matching: Continuous pattern matching is the act of checking for the presence of the constituents of a given pattern during a specific time interval.

In ocean environment, assume there is a requirement that "analyze past ocean current data and find the area that environmental abnormality lasts 7 days with similarity up to 80%." Every year, ocean current occurs almost the same pattern as the previous ones and has much influence to the temperature of seawater, DO concentration and so on. So in this example, doing continuous pattern matching analysis with 80% similarity can predict the occurrence of ocean current. By this, we can prevent much loss from natural hazard like ocean current.

3.3.2 Prediction based on Pattern: Prediction based on pattern is a kind of prediction query using already mined time pattern during time intervals.

In ocean environment, assume there is a requirement that "predict the temperature of seawater, DO concentration based on the data past 10 year's temperature of seawater and DO concentration." It is obvious that if the temperature is high, the DO concentration is low and if temperature is low, the DO concentration is high. In this example, we can predict temperature of seawater and DO concentration based on such this kind of pattern. If know more about the weather, we can manage water resource more efficiently and more useful to navigation and other fields.

4. USN MIDDLEWARE STRUCTURE

USN middleware, which is suitable for USN environment and supports the sensor data mining techniques, is located between sensor network and USN application service. It analyzes the sensed data collected from sensor networks and returns query result or extracted useful information to the client user. Also, It should take a rapid response to the required queries and provide multiple kinds of queries that are possible in USN environment. They are snapshot query, continuous query and event query. Now, it is a trend that many sensor networks are integrated to construct a system in USN application domain. There is no problem when all sensor networks use the same kind of sensor nodes. However, if the nodes used in different sensor networks are made in different companies and perform different functions and use different communication protocol, then we should provide common interface to use commonly. We call it a sensor network common interface that is located between sensor networks and USN application to integrate them.

Moreover, among the data collected from sensor networks, there may be some unexpected values or hidden pattern that we can't directly get from simple query. In order to provide useful information to client user, USN middleware should support sensor data mining techniques that are fit to sensor data characteristics. So in this paper, we propose a USN middleware including these functions and it is presented in Figure 2.

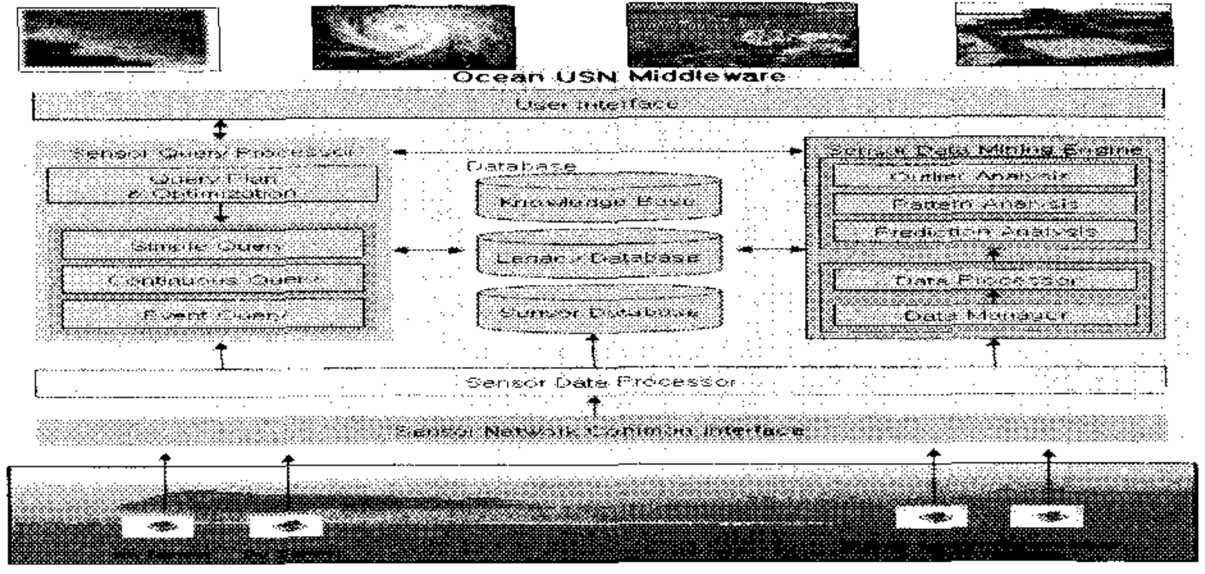

Figure 2. USN middleware structure

USN middleware is composed of sensor network, sensor network common interface, sensor data processor, sensor query processor, sensor data mining engine, user interface and so on. The functions of each component and their relationship are described as below.

Sensor network common interface defines the common message of the commands, report, requirements and responses which are used to communicate between sensed data and USN middleware. Sensor network common interface and sensor data processor communicate with these common messages.

Sensor query processor receives the query from client user and supports multiple kinds of queries in order to satisfy the user's requirements. There are 3 kinds of queries; they are snapshot query, continuous query and event query. Snapshot query requires real time response to the incoming data, continuous query requires the sensed information in period time and event query requires sensing information only when there is an event or specific situation occurs. After receiving the query from client, the query processor decides which kind of query to execute and then perform query plan and optimization.

Sensor data mining engine as mentioned in section 3 includes sensor data outlier analysis, sensor data pattern analysis and sensor data prediction analysis. Here we can get useful hidden information in data collected form sensor networks and send results to the client through user interface.

In order to process past sensed information or perform sensor data mining techniques, USN middleware needs to store the sensor data efficiently in the databases. We call database sensor database, legacy database and knowledge database. Sensor database is used to store sensed data, legacy database is used to store historical data and knowledge database is used to store useful information extracted from sensor data mining engine.

5. CONCLUSIONS AND FUTURE WORK

We can see sensor data collected from sensor networks as stream data and the research on stream data mining has been done in many fields, such as network monitoring, traffic monitoring, web click and so on. However, there is nearly no research or system about sensor data mining based on USN environment. In USN environment, data is collected from sensor nodes and there may be hidden pattern or some other useful information that we can't directly get from query processing. In order to extract this useful information, we need mining techniques, but the traditional data mining techniques which focus on historical and static data set can't be directly applied. Because of the characteristics of sensor data, we need sensor data mining techniques instead of traditional data mining techniques. So in this paper, we proposed sensor data mining techniques based on USN environment and USN middleware structure supporting these sensor data mining techniques. Sensor data mining techniques are categorized into sensor outlier analysis, sensor pattern analysis and sensor prediction analysis. Their detail classification and definition are also described. USN middleware which supports these sensor data mining techniques consists of sensor network, sensor network common interface, sensor data processor, sensor query processor, sensor data mining engine, user interface and so on. Among many USN application domains, in this paper we chose ocean domain and described them.

In the future, we will model the sensor data mining techniques based on USN environment and also implement detail algorithms of each sensor data mining techniques.

REFERENCES

Lu, C.T., Chen, D., Kou, Y., 2003 Algorithms for Spatial Outlier Detection, Proceeding of 3rd International Conference on Data Mining, Melbourne, Florida, pp. 597-600, Nov. 19-22.

Kargupta, H., Park, B., Pittie, S., Liu, L., Kushraj, D., Sarkar, K, 2002, *MobiMine: Monitoring the Stock Market from a PDA*. ACM SIGKDD Explorations, 3(2). ACM Press 37-46.

Subramaniam, S., Palpanas ,T., Papadopoulos ,D., Kalogeraki ,V and Gunopulos ,D., 2006, Online Outlier Detection in Sensor Data Using Non-Parametric Models. In *VLDB*, Seoul, Korea.

Cai, Y. D., Clutter, D., Pape, G., Han, J., Welge, M., and Auvil ,L., June 2004, MAIDS: Mining Alarming Incidents from Data Streams, (system demonstration), Proceedings of ACM-SIGMOD International Conference on Management of Data (SIGMOD'04), Paris, France.

ACKNOWLEDGEMENTS

This research was financially supported by the Ministry of Commerce, Industry and Energy (MOCIE) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation.