Research on Data Mining Models for the Internet of Things

Shen Bin[#], Liu Yuan[*], Wang Xiaoyi[*]

[#]Ningbo Institute of Technology, Zhejiang University
Ningbo, China

**College of Management, Zhejiang University*
Hangzhou, China

tsingbin@zju.edu.cn

Abstract— **In this paper, we propose four data mining models for the Internet of Things, which are multi-layer data mining model, distributed data mining model, Grid based data mining model and data mining model from multi-technology integration perspective. Among them, multi-layer model includes four layers: 1) data collection layer, 2) data management layer, 3) event processing layer, and 4) data mining service layer. Distributed data mining model can solve problems from depositing data at different sites. Grid based data mining model allows Grid framework to realize the functions of data mining. Data mining model from multi-technology integration perspective describes the corresponding framework for the future Internet. Several key issues in data mining of IoT are also discussed.**

Keywords— **Internet of Things, data mining models, RFID technology**

I. Introduction

The Internet of Things (IoT) refers to the next generation of Internet which will contain trillions of nodes representing various objects from small ubiquitous sensor devices and handhelds to large web servers and supercomputer clusters [23]. It is the next technological revolution after the revolution of computer and Internet. It integrates the new technologies of computing and communications (eg. sensor networks, RFID technology, mobile communication technologies, real-time localization, ubiquitous computing and IPV6 etc.) and builds the development direction of the next generation of internet. IoT is the core of Smart Planet that is proposed by IBM Corporation. Smart objects of the Internet of Things (eg. sensor inputs, actuators etc.) are able to communicate via the internet based on the new technologies of information and communication.

S. Haller et al.[2] have given the following definition of IoT : "A world where physical objects are seamlessly integrated into the information network, and where the physical objects can become active participants in business process. Services are available to interact with these 'smart object' over the Internet, query their state and any information associated with them, taking into account security and privacy issues."

Prof. Liu [3] has provided the ideas of IoT from the aspects of technology and economy respectively: "From the viewpoint of technology, IoT is an integration of sensor networks, which include RFID, and ubiquitous network. From the viewpoint of economy, it is an open concept, which integrates new related technologies and applications, productions and services, R. & D., industry and market."

The Internet of Things will produce large volumes of data. Let us take a supermarket in a supply chain, which adopts RFID technology, as an example. The format of raw RFID data is of the form: EPC, location, time. EPC represents the unique identifier read by an RFID reader; location is the place where the reader is positioned; and time is the time when the reading took place. It needs about 18 bytes to save a raw RFID record. In a supermarket, there are about 700,000 RFID tag. So for a RFID data stream of a supermarket, if the supermarket has readers that scan the items every second, about 12.6 GB RFID data will be produced per second, and the data will reach 544TB per day. Thus, it is necessary to develop effective methods for managing, analyzing and mining RFID data. The data in the Internet of Things can be categorized into several types: RFID data stream, address/unique identifiers, descriptive data, positional data, environment data and sensor network data etc [1]. It brings the great challenges for managing, analyzing and mining data in the Internet of Things.

II. Related Works

As a completely new paradigm of Internet, researches on the Internet of Things are still at the preliminary stage. Currently, there are some works about data mining in the Internet of Things, which mainly include the following three aspects:

Some works focus on managing and mining RFID stream data. For example, Hector Gonzalez et al.[4] proposed a novel model (RFID-Cuboids) for warehousing RFID data, which preserves object transitions while providing significant compression and path-dependent aggregates. RFID-Cuboids maintain three tables: (1) Info table, which stores path-independent information about product, (2) Stay table, which stores information about items that stay together at a location, (3) map table, which stores path information for performing structure-aware analysis. Hector Gonzalez et al.[5] adopted FlowGraph to represent the transportation of commodity, and also used it for multi-dimensional analysis of commodity flows. In reference [6], Hector Gonzalez et al. proposed a kind of compressed probabilistic workflows that capture the movement and significant exceptions of RFID flows. Elio

978-1-4244-5555-3/10/$26.00 ©2010 IEEE

Masciari [8] researches outlier mining in RFID data stream.

Some works interest in query, analyze and mine moving object data which is generated by various devices of IoT, e.g., GPS devices, RFID sensor networks, RADAR or satellites etc. For example, Xiaolei Li et al. [7] put forward a new framework, called ROAM, which is used for anomaly detection in moving object. In reference [10], Jae-Gil Lee et al. developed a novel partition-and-detect framework for trajectory outlier detection of moving object. Jae-Gil Lee et al. [9] also put forward a new trajectory classification method called TraClass using hierarchical region-based and trajectory-based clustering. In reference [11], a partition-and-group framework is proposed for trajectory clustering of moving object.

Other works are knowledge discovery from sensor data. Sensor network has several characteristics, e.g., limited resources, easy deployment of sensors, no maintenance, multi-hop and mass data etc. So data mining in sensor networks has its own features. Joydeep Ghosh [12] proposed a general probabilistic framework that allows supervised learning under the constraints of computational/memory/power limitations. Betsy George et al.[13] put forward Spatio-Temporal Sensor Graphs (STSG) to model and mine sensor data. STSG models are able to discover different types of patterns: anomaly patterns, centralized locations at each time interval, and nodes of future hotspot. Parisa Rashidi et al. [14] developed a novel adaptive mining framework for pattern mining from sensor data, which adapt to changes in data.

Although there are several contributions towards data mining from IoT, they mainly focus on the rudiments of IoT, eg. sensor network, RFID etc. As a completely new paradigm of Internet, IoT is still lack of models and theories for directing its data mining.

III. Data Mining Models for the Internet of Things

A. *Multi-layer data mining model for IoT*

According to the architecture of IoT and data mining framework of RFID [15], we propose the following multi-layer data mining model for IoT as shown in Fig 1, which is divided into four layers: data collection layer, data management layer, event processing layer and data mining service layer.

Among them, data collection layer adopts devices, e.g. RFID Reader and sinks etc., to collect various smart object's data, which are RFID stream data, GPS data, satellite data, positional data and sensor data etc. Different type of data requires different data collection strategy. In the process of data collection, a series of problems, e.g., energy-efficiency, misreading, repeated reading, fault tolerance, data filtering and communications etc., should be well solved.

Data management layer applies centralized or distributed database or data warehouse to manage collected data. After object identification, data abstraction and compression, various data are saved in the corresponding database or data warehouse. Take RFID data as an example, the raw format of RFID data stream is (EPC, location, time), where EPC marks smart object's ID. After data cleaning, we can obtain Stay table which contains records as the format (EPC, location, time_in, time_out). Then we use data warehouse, called RFID-CUBOID, to save and manage the corresponding data, which includes Info table, Stay table and Map table. Based on RFID-CUBOID, users can online analytical process RFID data conveniently. Besides, XML language can be adopted for describing data in IoT. Smart objects are connected with each other via the data management layer in the Internet of Things.

Event is an integration that combines data, time and other factors, so it provides a high-level mechanism for data processing of IoT. Event processing layer is used to analyze events in IoT effectively. Thus we can perform event-based query or analysis in event processing layer. The observed primitive events are filtered, and then complex events or events that are concerned by user are obtained. Then we can aggregate, organize and analyse data according to events.

Data mining service layer is built based on data management and event processing. Various object-based or event-based data mining services, such as classification, forecasting, clustering, outlier detection, association analysis or patterns mining, are provided for applications, e.g., supply chain management, inventory management and optimization etc. The architecture of this layer is service-oriented.

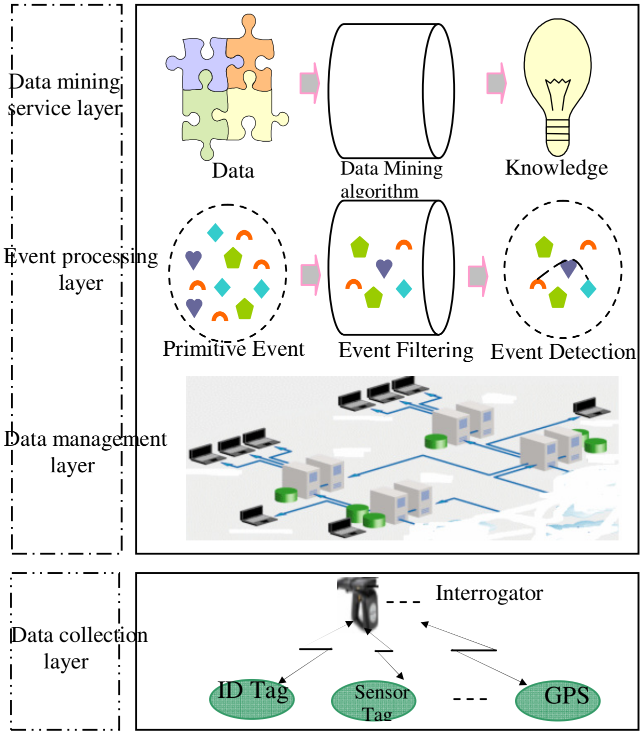

Fig. 1 Multi-layer data mining model for IoT

B. *Distributed data mining model for IoT*

Comparing with the common data, data in IoT has its own characteristics. For example, the data in IoT is always mass, distributed, time-related and position-related. At the same time, the data sources of IoT are heterogeneous, and the resources of nodes are limited. These characteristics bring several problems to centralized data mining architecture. At first, mass data of IoT is stored in different sites. Therefore, it is difficult for us to mine distributed data by centralized architecture. Secondly, data in IoT is mass and needs preprocessing in real time. So if we adopt central architecture, the requirement for hardware of central nodes is quite high. Thirdly, for the consideration of data security, data privacy, fault tolerance, business competition, legal constraints and other factors, the strategy of putting all relevant data together is often not feasible. Fourthly, the resources of nodes are limited. The strategy of sending all data to central nodes does not optimize the use of energy-costly transmissions. In most cases, the central nodes do not need all data, but some estimates of parameters. So we can pre-process the raw data in the distributed nodes, and then send the necessary data to the receiver.

Distributed data mining model for IoT is not only able to solve the problems brought by distributed storage of nodes, but also decompose the complex problems into simple ones. Thus the requirement of high performance, high storage capacity and computing power is reduced. In this paper, we propose a distributed data mining model for IoT, as shown in Fig. 2.

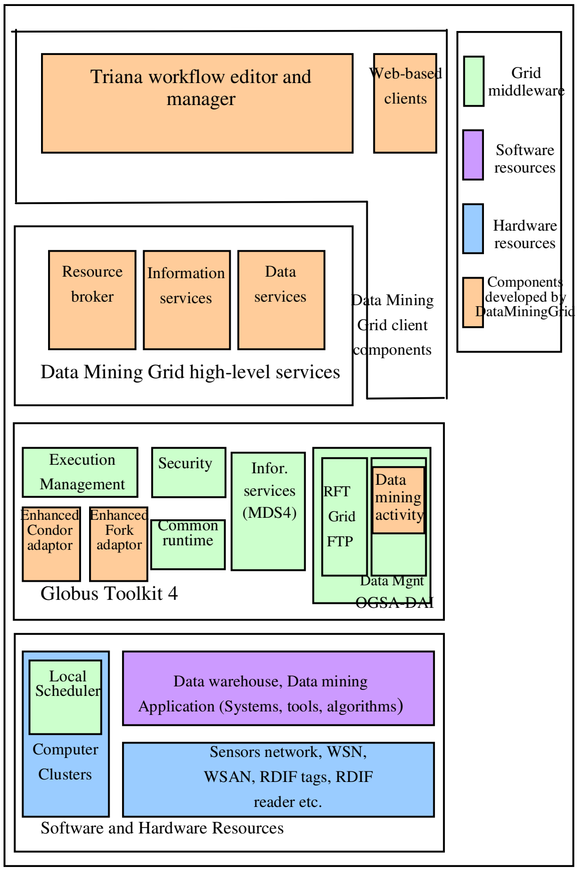

Fig. 2 Distributed data mining model for IoT

In this model, the global control node is the core of the whole data mining system. It chooses the data mining algorithm and the data sets for mining, and then navigates to the sub-nodes containing these data sets. The sub-nodes receive the raw data from various smart objects. These raw data is pre-processed by data filter, data abstraction and data compression, and then is saved in the local data warehouse. Local models are obtained by event filtering, complex event detection and data mining in local nodes. According to the demand of the global control node, these local models are submitted to the global control node and aggregated together to form the global model. Sub-nodes exchange object data, process data and knowledge with each other. The whole process is controlled by the multi-agent based collaborative management mechanism.

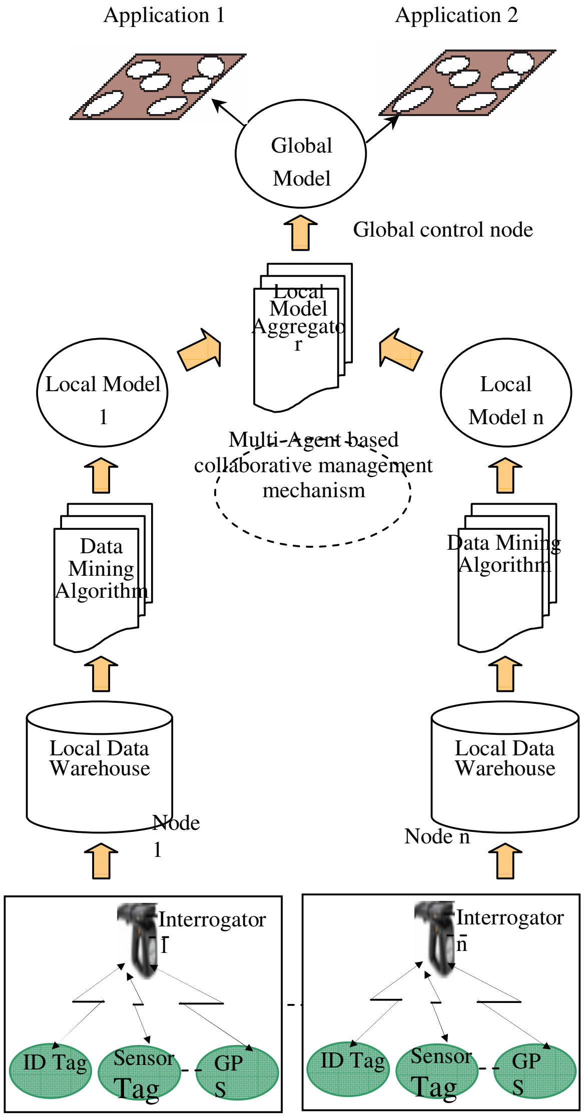

Fig. 3 Grid based data mining model for IoT

C. Grid based data mining model for IoT

Grid computing is a novel computing infrastructure, which is able to implement heterogeneous, large scale and high performance applications. As the same to IoT, Grid computing receives an increasing attention both from industry and the research community. The basic idea of Grid is that users can make use of the computation resources of Grid as the same as power resources. Various computing resources, data resources and devices resources can be accessed or used conveniently. The basic idea of IoT is to connect various smart objects via internet. Thus smart objects become intelligent, context-awareness, and long-range operable. Therefore we may regard smart objects of IoT as a kind of resources for Grid computing, and then use data mining services of Grid to implement the data mining operations for IoT.

P. Brezany et al. [19] proposed a fundamental infrastructure called GridMiner, which supports distributed online analytical processing and data mining. In reference [20], A. Congiusta discussed design aspects and implementation choices of WSRF-compliant Grid Services. In this paper, based on DataMiningGrid [21] which was put forward by Stankovski, V. et al., we propose a Grid-based data mining model for IoT, as shown in Fig. 3.

The differences between DataMiningGrid-based data mining model for IoT and DataMiningGrid is the part of software and hardware resources. IoT provides more types of hardware, e.g., RFID tags, RFID Readers, WSN, WSAN and Sensor networks etc. It also offers various software resources, e.g., event processing algorithms, data warehouse and data mining applications etc. Globus Toolkit 4 is adopted to implement various services of Grid. We also can make full use of high-level services of DataMiningGrid, client components of DataMiningGrid for data mining in IoT.

D. *Data mining model for IoT from multi-technology integration perspective*

The Internet of Things is one of the most important development directions of the next- generation Internet. At the same time, there are still a number of new directions, e.g., trusted network, ubiquitous network, grid computing, cloud computing etc. Therefore, from the perspective of multi-technology integration, we propose the corresponding data-mining model for IoT, as shown in Figure 4.

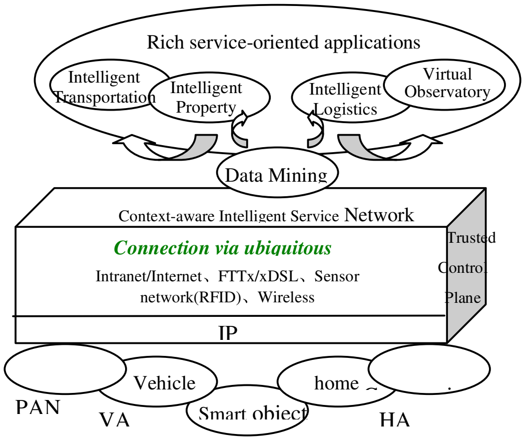
Rich service-oriented applications
Intelligent Transportation
Intelligent Property
Intelligent Logistics
Virtual Observatory
Data Mining
Context-aware Intelligent Service Network
Connection via ubiquitous
Intranet/Internet、FTTx/xDSL、Sensor network(RFID)、Wireless
Trusted Control Plane
IP
PAN
Vehicle
VA
Smart object
home
HA

Fig. 4 Data mining model for IoT from multi-technology integration perspective

In this model, data comes from the context-awareness of individuals, smart objects or the environment. 128-bit IPV6 address is adopted, and a variety of ubiquitous ways are provided for accessing to the future Internet, such as: Intranet/Internet, FTTx/xDSL, sensor devices, RFID, WLAN/WiMAX, 2.5/3/4G mobile access and so on. Trusted control plane is able to ensure credibility and controllability of data transmission. On this basis, we carry out data mining tools and algorithms, and submit gained knowledge to various service-oriented applications, such as intelligent transportation, intelligent logistics etc.

IV. Key Issues in Data Mining of IoT

A. *Data collection from smart objects of IoT*

When we conduct our data collection from smart objects of IoT, the special needs of smart objects should be taken into account. For example, if we want to collect data from distributed sensor networks, energy-efficiency, scalability and fault-tolerance should be considered. A series of strategies, e.g., data aggregation, can be adopted. Thus, the amount of transmission data is reduced, and the utilization of energy of sensor nodes is promoted. In reference [12], in order to reconcile the goals and conflicts in the process of sensor network data mining, Joydeep Ghosh proposed a general probabilistic framework under the constraints of computational/memory/power limitations.

B. *Data abstraction, compression, index, aggregation and multi-dimensional query*

The Internet of Things will produce a massive data of smart objects. Therefore, it is necessary to consider how to manage data of IoT effectively and how to implement online analytical query and processing conveniently [1, 5]. Data of smart objects has its own characteristics: (1) In the environment of IoT, devices such as RFID and sensors will produce massive data streams. (2) Data of smart objects is likely to inaccurate, and usually is time-related or location-related. (3) Data of smart objects tends to have its own implicit semantics. So it is necessary to recognize the implicit semantics of data. The characteristics of IoT data put forward new demands for data management and data mining. The key issues are includes:

1) Identification and addressing of smart objects: In IoT, the number of entities of smart objects will be billions. In order to query or interact with smart objects, it is necessary to realize smart objects' identification and addressing effectively.

2) Data abstraction and compression. Effective methods should be developed for filtering redundant data.

3) Data archive, index, scalability and access control for IoT data.

4) Data warehouse and its query language for multi-dimensional analysis.

5) Interoperability and semantic intelligibility for heterogeneous data of IoT.

6) Time-series level and event level data aggregation.

7) Privacy and protection problem in data management of IoT.

C. Event filtering, aggregation and detection

Event filtering and complex event processing are used to process simple events in data. The whole process includes the following steps. At first, data are aggregated according to events. The primitive events are filtered, and valuable events are obtained. And then, these simple atomic events are integrated into complex events. Thus we can detect the corresponding business logic by detecting complex events. For example, Tai Ku et al. [17] proposed a new complex event-mining network for monitoring RFID-enable application, and defined the fundamental concepts for the event management of supply chain, which uses RFID technology.

D. Centralized data processing and mining VS. Distributed data processing and mining

In different situations, centralized or distributed data processing and mining models can be adopted flexibility. Let's take distributed sensor network as an example. Under the constraints of nodes' computational/memory/power limitations, the strategy of sending all data to sink nodes does not optimize the use of energy-costly transmissions. In fact, in most cases we do not need all raw data, but are interested in some values of parameters. Therefore, a better approach is to pre-process data at each distributed nodes in advance. And then only necessary data is sent to sink nodes.

E. Research on data mining algorithms for IoT

Based on data management and event processing for IoT, a key issue is to study the novel data mining algorithms for IoT. The main tasks include classification, forecasting, clustering, outlier detection, association analysis, spatial and temporal patterns mining for IoT. For example, Chen Zhuxi et al.[16] proposed the frequent closed-path mining algorithm for RFID applications. Elio Masciari [8] studied outlier detection from RFID data stream.

F. Data mining towards the next generation of Internet

The next generation of Internet has many potential direction of development: IPV6 technology, ubiquitous networks, trusted network, semantic web, Grid (Semantic Grid, Data Grid and Knowledge Grid), service-oriented applications, optical transmission and cloud computing etc. In the next generation of Internet, these new technologies will integrate with IoT. Therefore, many new data mining problems need to be studied intensively. For example, semantic-based data mining from IoT, Grid-based data mining from IoT and Service-oriented data mining from IoT etc.

V. Conclusions and Future Works

As an important development direction of the next generation of Internet, the Internet of Things attracts many attentions by industry world and academic circles. IoT data has many characteristics, such as distributed storage, mass time-related and position-related data, and limited resources of nodes etc. These makes the problem of data mining in IoT become a challenge task.

In this paper, we propose four data mining models for the Internet of Things, which are multi-layer data mining model, distributed data mining model, Grid based data mining model and data mining model from multi-technology integration perspective. Among them, multi-layer model includes four layers (e.g. data collection layer, data management layer, event processing layer and data mining service layer). Distributed data mining model can well solve the problem arose from depositing data at different sites. At the same time, the complexity of problem is decomposed, and the requirements of high-performance, high storage capacity and high computing power for central nodes are reduced. Grid based data mining model adopts Grid framework to realize the functions of data mining. Data mining model from multi-technology integration perspective describes the corresponding framework for the future Internet. Several key issues in data mining of IoT are also discussed.

Possibilities for future work include: a) studying various data mining algorithms for IoT; b) implementation of Grid-based data mining systems and the corresponding algorithms.

Acknowledgment

This paper is supported in part by the National Natural Science Foundation of China under Grant No. 10876036 and 70871111. Besides, the paper is also supported in part by the Scientific Research Start-up Foundation of Ningbo Institute of Technology, Zhejiang University.

References

[1] Cooper J, James "A. Challenges for Database Management in the Internet of Things," *IETE Tech Rev*. 2009. 26:320-9.

[2] S. Haller, S. Karnouskos, and C. Schroth, "The Internet of Things in an enterprise context," *Future Internet Systems (FIS)*, LCNS, vol. 5468. Springer, 2008, pp. 14-8.

[3] Zhang Lin. "School of Management, Zhejiang University, Prof. Liu Yuan: The business scale of communications between smart objects is tens of times the scale of communications between persons," *Science Times*. 2009.11.16. (in Chinese)

[4] Hector Gonzalez, Jiawei Han, Xiaolei Li, Diego Klabjan. "Warehousing and Analyzing Massive RFID Data Sets," *ICDE* 2006: 83.

[5] Hector Gonzalez, Jiawei Han, Xiaolei Li. "FlowCube: Constructuing RFID FlowCubes for Multi-Dimensional Analysis of Commodity Flows," *VLDB* 2006: 834-845.

[6] Hector Gonzalez, Jiawei Han, Xiaolei Li. "Mining compressed commodity workflows from massive RFID data sets," *CIKM* 2006: 162-171.

[7] Xiaolei Li, Jiawei Han, Sangkyum Kim, Hector Gonzalez. "ROAM: Rule- and Motif-Based Anomaly Detection in Massive Moving Object Data Sets," *SDM* 2007.

[8] Elio Masciari. "A Framework for Outlier Mining in RFID data," *11th International Database Engineering and Applications Symposium (IDEAS 2007)*, pp.263-267, 2007.

[9] Jae-Gil Lee, Jiawei Han, Xiaolei Li, Hector Gonzalez: "TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering," *PVLDB* 1(1): 1081-1094 (2008)

[10] J.-G. Lee, J. Han, and X. Li. "Trajectory outlier detection: A partition-and-detect framework," *Proc. 24th Int'l Conf. on Data Engineering*, pages140-149, Cancun, Mexico, Apr. 2008.

[11] Jae-Gil Lee, Jiawei Han, Kyu-Young Whang. "Trajectory clustering: a partition-and-group framework," *SIGMOD* 2007: 593-604.

[12] Joydeep Ghosh. "A Probabilistic Framework for Mining Distributed Sensory Data under Data Sharing Constraints," *First International Workshop on Knowledge Discovery from Sensor Data*. 2007.

[13] Betsy George, James M. Kang, Shashi Shekhar. "Spatio-Temporal Sensor Graphs (STSG): A data model for the discovery of spatio-temporal patterns," *Intell. Data Anal.* 13(3): 457-475 (2009).

[14] Parisa Rashidi and Diane J. Cook. "An Adaptive Sensor Mining Framework for Pervasive Computing Applications," *2nd International Workshop on Knowledge Discovery from Sensor Data*, 2008.

[15] Joshua Huang. (2009) "RFID Data Mining: Opportunities and Challenges," [Online]. Available: http://homepage.fudan.edu.cn/~wdzhao/rfid.html.

[16] Chen Zhu-xi, HU Kong-fa, et al. "Frequency mining closed path algorithm based in the modern logistic management system," *Computer integrated manufacturing systems*, 15(4) 2009. (in Chinese)

[17] Tai Ku, YunLong Zhu, KunYuan Hu. "A novel complex event mining network for monitoring RFID-enable application," *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application.* 2008.

[18] JU Chunhua. "Research on distributed data mining model for data stream of dynamic chain enterprises," *Management World*. 2008.12 (in Chinese)

[19] P. Brezany, I. Janciak, and A. M. Tjoa. "GridMiner: a fundamental infrastructure for building intelligent Grid systems," *Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, IEEE press, 200, pp. 150~156.

[20] A. Congiusta, D. Talia, and P. Trunfio. "Distributed data mining services leveraging WSRF," *Future Generation Computer Systems*, 23(1), 2007. pp.34-41.

[21] Stankovski, V. Swain, M. "Digging Deep into the Data Mine with DataMiningGrid," *IEEE Internet Computing*. 12(6), 2008.

[22] Dun Haiqiang, Zhao Wen, Deng Pengpeng et al. "A Commodity workflow mining approach based on RFID datasets," *ACTA ELECTRONICA SINICA*. 2008.12 (in Chinese)

[23] Anne James, Joshua Cooper, Keith Jeffery, and Gunter Saake. "Research Directions in Database Architectures for the Internet of Things: A Communication of the First International Workshop on Database Architectures for the internet of things (DAIT 2009)," *BNCOD* 2009: 225-233.

.