Mining Data Correlation from Multi-faceted Sensor Data in the Internet of Things

Cao Dong[1,2], Qiao Xiuquan[2], Judith Gelernter[1], Li Xiaofeng[2], Meng Luoming[2]

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, 15213, USA

[2] State Key Lab of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China

Abstract: Sensors are ubiquitous in the Internet of Things for measuring and collecting data. Analyzing these data derived from sensors is an essential task and can find the useful latent information besides the data itself. Because the Internet of Things contains sorts of sensors, the measurement data collected by these sensors are multi-type data, sometimes containing temporal series information. If just separately dealing with different sorts of data, we will miss useful information. This paper proposes a method to learn the correlation among multi-faceted data, which contain many types of data with temporal information, and our method can simultaneously deal with multi-faceted data. We transform high dimensional multi-faceted data into lower dimensional data which are set as multivariate Gaussian Graphical Models, then mine the correlation among multi-faceted data by learning the structure of the multivariate Gaussian Graphical Models. With a real data set, we verified our method, and the experiment demonstrated that the method we propose can correctly find the correlation among multi-faceted measurement data.

Key words: Multi-faceted data, Sensors, Internet of Things, Gaussian Graphical Models

I. INTRODUCTION

The Internet of Things (IoT) is a world-wide network of interconnected objects uniquely addressable, based on standard communication protocols [1]. The IoT is also a heterogeneous network [2]. There are very large numbers of heterogeneous sensor devices and sensor networks being deployed to support the vision of the IoT [3]. Sensors have become ubiquitous in the IoT because they are tiny, inexpensive to produce, and most are reliable and fairly robust under varying ambient conditions. A sensor network requires much less bandwidth than does a wired network [4]. A summary of network types and applications is provided by Yick et al. [5].

These sensors are used to monitor and collect data. We find several features of sensor data in the IoT. (1) The data derived from sensors in the IoT are massive because there are a large number of sensors for various applications. (2) The sensor data in the IoT are multi-type. There are two meanings of the term "multi-type": (a) the data derived from sorts of sensor networks are multi-type data because the IoT is a heterogeneous network which includes many kinds of sensor networks. (b) the data collected from a kind of sensor network are multi-type data. This paper focuses on this kind of multi-type sensor data. (3) the sensor data are time-sensitive. At different times, the sensors collect different measured values. Cooper et al. [6] also mentioned that sensor data is multidimensional time series data.

Sensors are also for providing useful latent information behind the data. For example, we want to know the correlation among temperature values at different times in order to predict the temperature in the future.

Some recent works concern the analysis and mining of sensor network data [7] [8] [9] [10]. Indeed, this is an emerging field with its own dedicated annual ACM workshop SensorKDD [11]. These research projects, however, do not explicitly address the multi-faceted nature of the data derived from sensors. Some of them just

consider the correlation among the data for data gathering.

Unlike others, we simultaneously consider the multi-type data and the temporal information. For example, we want to know whether the humidity at one time will affect the carbon dioxide content at a later time. In this case, humidity, carbon dioxide content and temporal information need to be processed simultaneously. If we just consider parts of types of these data, then we can't get a good answer. We set multi-faceted data as multivariate Gaussian Graphical Models and find whether there are correlations among multi-faceted data by estimating the structure of Gaussian Graphical Models.

The remainder of this paper is organized as follows. Section 2 introduces the related work on the data analysis of sensor network. Section 3 explains how to learn the correlation of multi-faceted data derived from sensors. Experimental results are presented in Section 4. We summarize our work and discuss directions for future research in Section 5.

II. RELATED WORK

Recently analyzing sensor network data is concerned by researchers. Jindal et al. [12] presented a model of spatially correlated sensor network data. The proposed model is Markovian in nature and can capture correlation in data irrespective of the node density, the number of source nodes, or the topology. They created tools that can be easily used by researchers to synthetically generate traces of any size and degree of correlation. Their work is similar with ours, but they just focus on the spatial correlation about single type data. Gupta et al. [13] proposed the algorithm to exploit data correlations in sensor data, but the aim of their research is to minimize communication costs incurred during data. Their proposed approach is to select a small subset of sensor nodes that may be sufficient to reconstruct data for the entire sensor network. They defined the problem of selecting such a set of sensors as the connected correlation-dominating set problem, and formulated it in terms of an appropriately defined correlation structure that captures general data correlations in a sensor network.

McGuire et al. [14] gave the method to discover spatiotemporal neighborhoods in sensor datasets where a time series of data is collected at many spatial locations. Bhattacharya et al. [15] thought the modeling of high-level semantic events from low-level sensor signals is important. So they considered the problem of distributed indexing and semantic querying over such sensor models. In fact, they mined the correlation among the data by querying, which is different from ours. Safarinejadian et al. [16] proposed a distributed variational Bayesian algorithm for density estimation and clustering in sensor networks. This algorithm produces an estimate of the density of the sensor data without requiring the data to be transmitted to and processed at a central location. Alternatively, this algorithm can be viewed as a distributed processing approach for clustering the sensor data into components corresponding to predominant environmental features sensed by the network. There are still many other methods for analyzing the data derived from sensors. But few of research simultaneously consider multi-faceted measurement data.

III. MINING DATA CORRELATION FROM MULTI-FACETED SENSOR DATA

A. Statement of Problem

The problem is as follows: there is a sensor network with N sensors deployed. Each sensor has M types of measurement data. For each type of measurement data for one sensor, there are T measured values collected from T different times $t_{i,i=1,\cdots,T}$. Figure 1(a) describes the structure of the multi-faceted measurement data. For one type of measurement data, we use $N \times T$ matrix $\mathbf{U}$ to express the measured values derived from N sensors gotten at T different times.

$$\mathbf{U}_{N\times T} = (U_1, U_2, \cdots, U_T) \quad (1)$$

For all types of measurement data, we get M types of data at the same time for N sensors, which is expressed by $N \times M$ matrix $\mathbf{Q}$.

$$\mathbf{Q}_{N\times M} = (Q_1, Q_2, \cdots, Q_M) \quad (2)$$

B. The Method for Finding Correlations among Multi-faceted Data

We can extend (1) and (2), and then combine them together. We will get the expression of multi-faceted data:

$$\begin{aligned} X_{N\times(M\times T)} &= \left(\mathbf{U}^{(1)}_{N\times T}, \mathbf{U}^{(2)}_{N\times T}, \cdots, \mathbf{U}^{(M)}_{N\times T}\right) \\ &= \left(\mathbf{Q}^{(1)}_{N\times M}, \mathbf{Q}^{(2)}_{N\times M}, \cdots, \mathbf{Q}^{(T)}_{N\times M}\right) \end{aligned} \quad (3)$$

The transformation procedure is shown in Figure 1. Now we get the new matrix $X_{N\times K}$ about multi-faceted data if we set $K = M\times T$.

$$X_{N\times K} = (X_1, X_2, \cdots, X_K) \in \mathbb{R}^K \quad (4)$$

We set $X_k \in X_{N\times K}$ as a feature of the multi-faceted data and then there are K features in the multi-faceted data as shown in Figure 1(b).

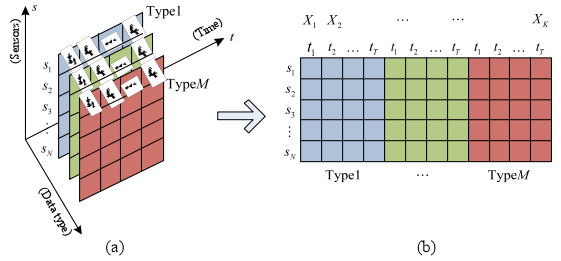


Fig. 1 Transforming the high dimensional data into lower dimensional data

If we set these features of multi-faceted data as random variables, the correlation among these features can be described by the dependence or independence of random variables. We suppose that the distribution of these features (random variables) is a Gaussian distribution, which will make the calculation easier. So the distribution of $X_{N\times K}$ can be regarded as multivariate Gaussian Graphical Models.

The distribution of $X_{N\times K}$ is:

$$p_X(x_1, x_2, \cdots, x_K) = \frac{1}{(2\pi)^{K/2}|\Sigma|^{1/2}} \exp\left(\frac{-(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}\right) \quad (5)$$

where, Σ is the covariance matrix, μ is the known mean vector. The precision matrix of p_X is

$$\Omega = \Sigma^{-1} = [\omega_{ij}] = (\omega_1 \omega_2 \cdots \omega_K) \quad (6)$$

In order to specify the problem more clearly, we show the Gaussian Graphical Models in Figure 2. Let $G = (V, E)$ is an undirected graph, which corresponds to the Gaussian Graphical Models. The nodes V represent features (random variables) in $X_{N\times K}$. And the edges E correspond to pairs of nodes. $(X_i, X_j) \notin E$ in case $\omega_{ij} = 0$, which means X_i and X_j are not connected by the edge if $\omega_{ij} = 0$. In the Gaussian Graphical Models, any two nodes X_i and X_j are conditionally independent given the other coordinates of $X_{N\times K}$ if these two nodes are not connected by an edge, which means $\omega_{ij} = 0$. So the precision matrix Ω can encode the graph.

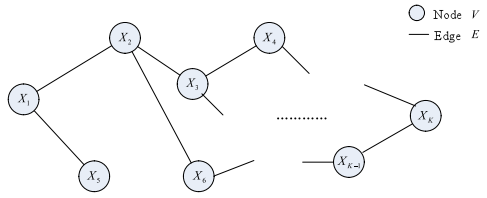


Fig. 2 Gaussian Graphical Models for the multi-faceted data

Finding the correlation among data can be transformed into finding the conditional independence among random variables in Gaussian Graphical Models. Further more, the precision matrix Ω encodes all the conditionally independent or dependent relation among all random variables, our target is to calculate the precision matrix Ω from measured (observed) values.

We use maximum likelihood estimation for log-likelihood of Ω to calculate the estimated value of Ω .

$$\ell(\Omega) = \frac{n}{2}\log|\Omega| - \frac{n}{2}\mathrm{trace}(\Omega\hat{\Sigma}) - \frac{np}{2}\log(2\pi) \quad (7)$$

where $\hat{\Sigma}$ is the sample covariance matrix as follows.

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \frac{1}{N}\sum_{i=1}^{N}X_i\right)\left(X_i - \frac{1}{N}\sum_{i=1}^{N}X_i\right)^T \quad (8)$$

[17] [18] shows that if the number of samples N is small compared to the number of features K , the sample covariance matrix $\hat{\Sigma}$ may not be invertible, which also makes the estimated value of Ω invertible. Our paper adopt the algorithm proposed in [19] to solve the estimation of the precision matrix Ω when $N < K$.

For convenience, we can get (9) by removing constants of (7)

$$\ell(\Omega) = \log|\Omega| - \mathrm{trace}(\hat{\Sigma}\Omega) \quad (9)$$

Next adding the ℓ_1 penalty $\|\Omega\|_1$ to the log-likelihood of Ω with a positive regularization parameter λ [20], then the estimator $\Omega'(\lambda)$ is obtained by minimizing the

regularized negative log-likelihood:

$$\Omega'(\lambda) = \arg\min_{\Omega \succ 0} \left\{ -\ell(\Omega) + \lambda \|\Omega\|_1 \right\} = \arg\min_{\Omega \succ 0} \operatorname{trace}(\Sigma\Omega) - \log|\Omega| + \lambda \|\Omega\|_1 \quad (10)$$

At last, we can get the estimation $\Omega'(\lambda)$ of the precision matrix Ω.

IV. EXPERIMENTS

A. Data Set

We downloaded the sensor data set provided by the Intel Research Berkeley Lab from [21] because it is a popular data set in this area. The data set has 54 nodes sensor network which collect temperature, humidity, and light data.

B. Experiments and Analysis

For each type of measurement data: temperature, humidity, and light, we select first 60 time slot measurement values for 54 sensors, which means $N = 54$, $T = 60$, $M = 3$, to find correlations among the same type of data at different times. Then we separately get the Gaussian Graphical Models for temperature, humidity, and light, which are named $X_{Temperature}$, $X_{Humidity}$ and X_{Light}.

1. Experiment for the temperature measurement data

According to the result of (11), we get the estimated precision matrix $\hat{\Omega}(\lambda)$ of $X_{Temperature}$. Finally we get the graphical model for temperature measurement data in Figure 3.

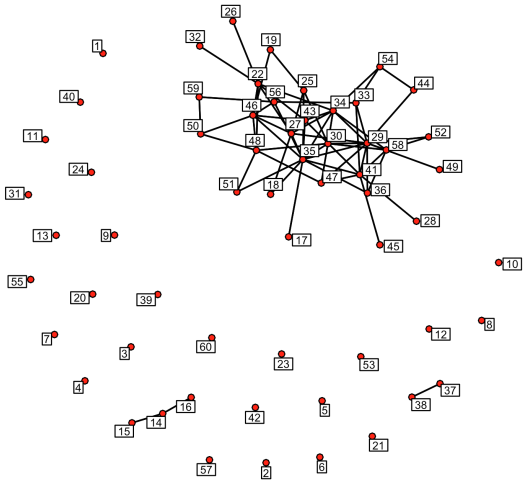
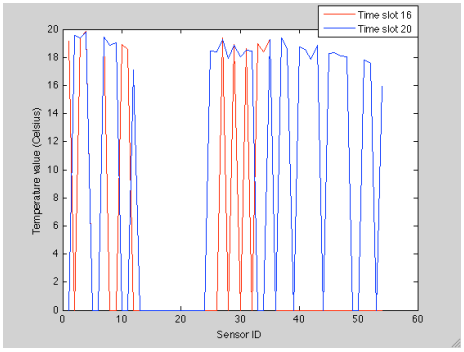
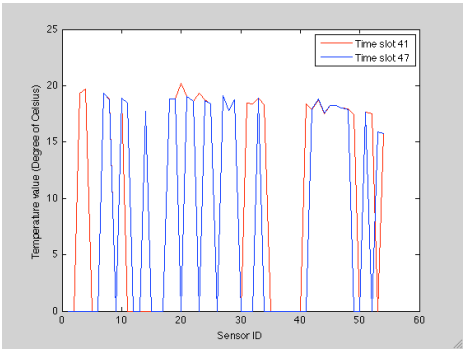


Fig. 3 Graphical models for the temperature measurement data



(a)



(b)

Fig. 4 Comparison of conditional independent variables and dependent variables in temperature data

We learn from Figure 3 that the nodes which are not connected by the edges are conditionally independent. For example, node 16 and node 20 are not connected by an edge, which means the measurement data in the time slot 16 are conditional independent with the data in the time slot 20. In order to verify the conditionally independent of these two time slot, we show the measured values for 54 sensors in the time slot 16 and 20 in Figure 4(a). The X-axis shows 54 different sensors and the Y-axis is the measured values. The red curve is the measured values in the time slot 16 and the blue one is the measured values in the time slot 20. From Figure 4(a) we found the curve of measured values in the time slot 16 is vastly different from the curve of values in the time slot 20.

Meanwhile, we find node 41 and 47 are connected by an edge in Figure 3, which means the measurement data in the time slot 41 are dependent with the data in the time slot 47. We also give the measured values for all sensors in the time slot 41 and 47 in Figure 4(b).

We found the similar results for humidity, and light with temperature.

2. Experiment for the multi-faceted measurement data

Now we estimate the precision matrix about the multi-type data with temporal information. The aim of the experiment is to verify whether our algorithm can find the correlation among different type of data in different times. For three types of measurement data: temperature, humidity, and light, we select first 25 time slot measurement data of 54 sensors, which means $N = 54, T = 25, M = 3$. Then we got the Gaussian Graphical Models for all types of data. According to (10), we got the estimated precision matrix. The graphical model for all types of data is in Figure 5.

In order to determine whether all three types of data are similarly correlated at same time, we average the measured values in each time slot over all sensors, the result of which is shown in Figure 6. We find that the curves of three type measurement data have similar shape, which tell us that the three types of measurement data have relevance in the same time slot.

At last, we find whether there is the correlation among data collected from both different times and different data types. And we found an interesting thing that in Figure 5 there is no edge between any two nodes which derived from both different times and different data types. That means any two data are conditional independent when these two data collected from both different times and different data types in this data set we used.

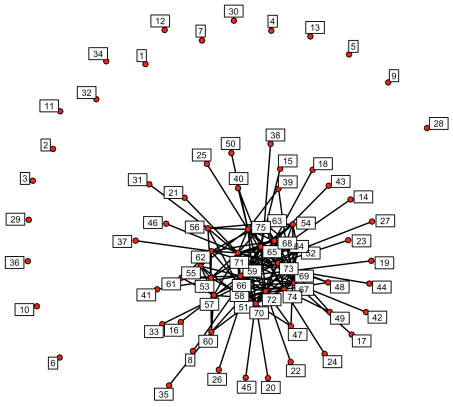


Fig. 5 Graphical models for the multi-faceted data

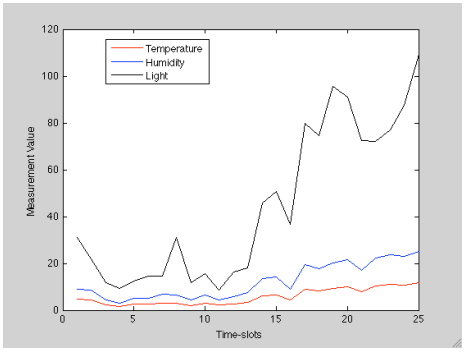


Fig. 6 Comparison of three types of measurement data

V. CONCLUSIONS

We proposed a method to detect whether there can be found a correlation among multi-faceted sensor data. Because the multi-faceted data are high dimensional, we transformed them into lower dimensional data. Each sensor produces more than one type of data, making the amount of data produced more than the number of sensors, so we estimated the precision matrix of the Gaussian Graphical Models with ℓ_1 penalty. We ran experiments to verify whether our method can discover dependency among temperature, humidity and light data. Results demonstrate that our method is valid for the temperature, humidity and light data.

Acknowledgements

This work has been performed in the Project "The Basic Research on Internet of Things Architecture" supported by National Key Basic Research Program of China (973 Program) (2011CB302704), and partly supported by National Natural Science Foundation of China (No. 60802034), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20070013026), Beijing Nova Program (No.2008B50) and "New generation broadband wireless mobile communication network" Key Projects for Science and Technology Development (No. 2011ZX03002-002-01).

References

[1] "Internet of Things in 2020: Roadmap for the Future," 2008, online, http://www.smart-systems-integration.org /public/internet-of-things

[2] N. Gershenfeld, R. Krikorian, D. Cohen, "The Internet of Things", Scientific American, 2004.

[3] A. Katasonov, O. Kaykova, O. Khriyenko, "Smart

semantic middleware for the internet of things", Proc. of the Fifth International Conference on Informatics in Control, Automation and Robotics, pp. 169-178, 2008.
[4] J.F. Martínez, P. Castillejo, M. Zuazua, "Wireless sensor networks in knowledge management", Procedia Computer Science, 1(1), pp. 2291-2300, 2010.
[5] J. Yick, B. Mukherjee, D. Ghosal, "Wireless sensor network survey", Computer Networks, 52(12), pp. 2292-2330, 2008.
[6] J. Cooper, A. James, "Challenges for Database Management in the Internet of Things", IETE Tech Review, 26 (5), pp. 320-329, 2009.
[7] M. Zhao, Y. Yang, "An Optimization Based Distributed Algorithm for Mobile Data Gathering in Wireless Sensor Networks", INFOCOM 2010, pp. 501-505, 2010.
[8] S. Yoon, C. Shahabi, "The Clustered AGgregation (CAG) technique leveraging spatial and temporal correlations in wireless sensor networks", ACM Transactions on Sensor Networks, 3 (1), pp. 1-39, 2007.
[9] V. Rajamani, S. Kabadayi, Julien, C., "An interrelational grouping abstraction for heterogeneous sensors", ACM Transactions on Sensor Networks, 5(3), pp. 1-31, 2009.
[10] J. Wang, Y. Liu, S.K. Das, "Energy-efficient data gathering in wireless sensor networks with asynchronous sampling", ACM Transactions on Sensor Networks, 6 (3), pp. 1-37, 2010.
[11] O.A. Omitaomu, A.R. Ganguly, J. Gama, "Knowledge Discovery from Sensor Data (SensorKDD)", SIDGKDD Explorations 11(2), pp. 84-87, 2009.
[12] A. Jindal, K. Psounis, "Modeling spatially correlated data in sensor networks", ACM Transactions on Sensor Networks, 2 (4), pp.466-499, 2006.
[13] H. Gupta, V. Navda, S. Das, "Efficient gathering of correlated data in sensor networks", ACM Transactions on Sensor Networks, 4(1), pp. 1-31, 2008.
[14] M. McGuire, V.P. Janeja, A. Gangopadhyay, "Spatiotemporal Neighborhood Discovery for Sensor Data". KDD Workshop on Knowledge Discovery from Sensor Data, pp. 203-225, 2008.
[15] A. Bhattacharya, A. Meka, A. K. Singh, "MIST: distributed indexing and querying in sensor networks using statistical models", Proc. of the 33rd international conference on Very large data bases (VLDB), pp. 854-865, 2007.
[16] B. Safarinejadian, M.B. Menhaja, M. Karrari, "Distributed variational Bayesian algorithms for Gaussian mixtures in sensor networks", Signal Processing, 90 (4), pp. 1197-1208, 2010.
[17] O. Baneriee, L.E. Ghaoui, A. d'Aspremont, " Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data", Journal of machine learning research, 9, pp. 485-516, 2008.
[18] S. Amizadeh, M. Hauskrecht, "Latent Variable Model for Learning in Pairwise Markov Networks", Twenty-Fourth AAAI Conference on Artificial Intelligence, pp. 382-387, 2010.
[19] J. Friedman, T. Hastie, R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso", Biostatistics, 9 (3), pp. 432-441, 2008.
[20] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso", Journal of the Royal Statistical Society, Series B, pp. 267-288, 1996.
[21]http://www-2.cs.cmu.edu/~guestrin/Research/Data/, download at October 2010.

Biographies

Cao Dong is doctoral student who has been visiting Carnegie Mellon University since 2008. He is enrolled at the Beijing University of Posts and Telecommunications in China. He received his undergraduate degree at the China University of Mining and Technology in 2005. His current research interests include Next Generation Network, information retrieval and data mining.

Qiao Xiuquan is born in 1978, Associate Professor. His main research interests include the intelligent theory and technology of network services.

Judith Gelernter is a postdoctoral fellow in the School of Computer Science of Carnegie Mellon University, with research projects also in the university's Institute for Software Research. She received a BA degree from Yale University, an AM in Fine arts from Harvard University, and a PhD in Information science from Rutgers University. Her current research is in text mining, social network analysis and geo-informatics.

Li Xiaofeng, born in 1950, Professor. Her main research interests include intelligent network and communication software.

Meng Luoming, born in 1955, Professor, Ph.D. supervisor. His main research interests include network management and communication software.