Distributed Data Mining in Wireless Sensor Networks

https://doi.org/10.3991/ijoe.v12i11.6224

Du Juan[1], Wu Fenfen[2]
[1] Yellow River Conservancy Technical Institute, Henan, China
[2]Henan College of Transportation, Henan, China

***Abstract*—With the rapid development and wide application of sensor network technology, consequently a huge volume of data would be continuously generated and collected. In order to process the data and analyze the data more accurate and efficient, the paper proposed a distributed data mining method in wireless sensor networks. Thus Smart Octopus, an open framework for seamlessly integrating sensor network and data mining technology, so that both of the huge amounts of data resource collected in sensor networks and the powerful knowledge discovery capability of data mining could be effectively and efficiently utilized, is discussed in this paper.**

***Index Terms*—distributed data, data mining, integration, Hadoop, wireless sensor networks**

I. INTRODUCTION

Rapid advances in wireless communication, microelectronics and embedded systems during the past several decades have enabled the development of low-cost, low-power and multi-functional sensor nodes that are small in size and capable of untethered communicating within short distance. These tiny sensor nodes leverage the idea of new generation of massive-scale distributed sensor network technology, which is characterized with its scalability, distributed, self-organizing capability, easy-to-deploy and a couple of other features as well. The technology is therefore much appropriate to a wide range of application areas as briefly described in [1-2]. It thus promises to revolutionize the way we live, work and interact with the physical environment. Nowadays, sensor network had been accordingly recognized as one of the dominant technology trends in the coming decades[3].

For the problem of most of sensor data estimation methods did not consider the characteristics of sensor data, which may lead to high computational complexity, Li [4] propose a correlation analysis-based estimation framework of sensor data. For the problem of high computational complexity of SVR (Support Vector Regression)–based estimation method, based on the framework, he proposed correlation analysis-based LS-SVR (Least Square Support Vector Regression) sensor data estimation method called CALS-SVR (Correlation Analysis-based LS-SVR). In this method, he considers the characteristics of the sensor data of wireless sensor networks, and extracts the most correlated sensor variable to be used as the input of modeling and estimation. And also the author adopts the LS-SVR with low computational complexity to estimate the sensor data. So the estimation efficiency can be improved largely. The experiments results show that the proposed CALS-SVR has better estimation efficiency and higher estimation accuracy compared to present sensor estimation method based SVR and LS-SVR.

For the problem of high computational complexity and low estimation accuracy in existing sensor estimation method, and also considering the implementation problem of sensor data estimation method on the wireless sensor node of WSN, Reddy [5] propose the correlation analysis-based multiple linear regression sensor data estimation method called CA-MLR (Correlation Analysis-based Multiple Linear Regression). In this method, he explores the characteristics of sensor data of wireless sensor networks, and takes correlation analysis on them. And then combine with multiple regressions, which have low computational complexity. So the estimation efficiency can be improved largely. The experiments results show that the proposed method has better estimation efficiency and accuracy, so it is very suitable for applying on the wireless sensor node.

For the problem of uncertain sensor data processing, Wang [6] proposes an uncertain sensor data processing framework based on MVPCA (Multiple variable Principle Component Analysis). And for the problem of uncertain sensor data estimation, Sun [7] proposes an uncertain data estimation method based the framework. In this method, he uses the MVPCA to extract the intrinsic feature of the uncertain sensor data. In this way, most of the uncertainty can be eliminated. Then he adopts the multiple regressions based on correlation analysis to estimation the sensor data. So the uncertain sensor data can be estimated. The experiments results show that the proposed method can estimate the uncertain sensor data efficiently with high efficiency.

For the problem of estimation mode updated not in time and big estimation error of dynamic stream sensor data, Jing [8-9] proposes a sensor data stream estimation method, which is called KF-CAMLR (Kalman Filter-Correlation Analysis-based Multiple Linear Regression). In this method, the Kalman Filter is combined with multiple regressions to estimate dynamic sensor data stream. The Kalman filter adjusts its working states according the estimation error, and at the same time adjusts the model parameters of the multiple regressions, so the estimation model adjusts efficiently according to data model in the sensor stream. The estimation accuracy can be improved largely. The experiments results show that the proposed sensor data stream estimation method based on Kalman can estimate dynamic sensor data in sensor data stream efficiently with high estimation accuracy.

An integrated open framework (code named Smart Octopus hereafter) for seamlessly integrating sensor network and data mining technology, so that both of the precious huge volume of data resources collected in sensor networks and the powerful knowledge discovery capability of data mining technology could be best utilized for process control, decision making, business management, and the other purposes is discussed in this paper.

II. METHOD AND ALGORITHM

Data management issue including frameworks that support attribute-based data naming, the networking aggregation (or reduction) and routing is one of the key challenges that sensor network poses to researchers. In this subsection, three popular data management techniques, namely Directed Diffusion, Cougar and Tiny DB, are briefly described.

Directed Diffusion [8]: Directed Diffusion is a novel data centric and data dissemination paradigm for sensor networks. It has some remarkable features such as data-centric dissemination, reinforcement-based adaptation to the empirically best path, and in-network data aggregation and caching. These features can enable highly energy-efficient and robust dissemination in dynamic sensor networks at the same time minimizing the per node configuration that is characteristic of today's networks. More specifically, it has three key characteristics: localized algorithms, named data, and support for in-network processing. Direct Diffusion adopts a declarative, publish/subscribe API that isolates data producers and consumers from the details of the underlying data dissemination algorithms, and it is the business of the diffusion implementation to ensure that data travels from publisher to subscriber efficiently. Cougar; The Cougar device database system proposes distributing database queries across a sensor network as opposed to moving all data to a central site Queries are declarative, and users can issue queries without knowing how the data is generated in the sensor network and how the data is processed to compute the query answer. An underlying query optimizer decides on the placement of in-network processing, and is performed by a gateway node that has an idea of the status of nodes and links in the network. Such a centralized scheme can be inefficient if the status of nodes is rapidly changing e. g.， due to environmental dynamics), but can enable a potentially more efficient global optimization of in-network query processing.

Tiny DB: Tiny DB is a query processing system with small footprint intended for extracting information from highly resource-constrained sensor networks running Tiny OS. Tiny-DB does not require users to write embedded C code for sensors. Instead, Tiny DB provides a simple, SQL-like interface to specify the data users want to extract along with additional parameters, much as we would pose queries against a traditional database. Given a query specifying our data of interests, Tiny DB collects that data from motes in the environment, filters it, aggregates it together, and routes it out to an IC. Tiny DB does this via power efficient in-network processing algorithms.

Although the specific approaches have varied widely, the data management approaches to sensor networks have broadly followed similar themes. Among the three popular techniques described above, both Cougar and Tiny DB use a database approach towards sensor data management while Direct Diffusion falls into the other category. Unfortunately, there is no further research focus on sensor network data mining had been conducted as of today. The network topology is shown in figure 1.

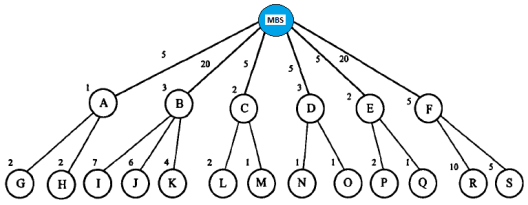

Figure 1. Network topology

The basic equation of the algorithm is shown in the following equation (1):

$$\hat{f}_H^{\alpha}(x) = \frac{1}{\Gamma(1+\alpha)} \int_{-\infty}^{\infty} \frac{f(t)}{(t-x)^{\alpha}} (dt)^{\alpha} = \frac{1}{\Gamma(1+\alpha)} \int_{-\infty}^{\infty} f(t) g(x-t) (dt)^{\alpha} = f(x) * g(x), \quad (1)$$

The equation is as follows:

$$\partial_j (C_{ijkl} \partial_k u_l + e_{kij} \partial_k \varphi) - \rho \ddot{u}_i = 0 \quad (2)$$

Under the linear theory, that is:

$$\partial_j (e_{ijkl} \partial_k u_l - \eta_{kij} \partial_k \varphi) = 0 \quad (3)$$

The linear equation can be expressed into the following simplified forms:

$$L(\nabla, \omega) f(x, \omega) = 0$$

$$L(\nabla, \omega) = T(\nabla) + \omega^2 \rho \mathsf{J} \quad (4)$$

In which,

$$T(\nabla) = \begin{Vmatrix} T_{ik}(\nabla) & t_i(\nabla) \\ t_k^T(\nabla) & -\tau(\nabla) \end{Vmatrix}, \ \mathsf{J} = \begin{Vmatrix} \delta_{ik} & 0 \\ 0 & 0 \end{Vmatrix},$$

$$f(x, \omega) = \begin{Vmatrix} u_k(x, \omega) \\ \varphi(x, \omega) \end{Vmatrix} \quad (5)$$

Consider delay, the L can be expressed as:

$$L^0 = \begin{Vmatrix} C_{ijkl}^0 & e_{kij}^0 \\ e_{ikl}^{0T} & -\eta_{ik}^0 \end{Vmatrix} \quad (6)$$

These functions can be expressed in the following form:

$$C(\mathrm{x}) = C^0 + C^1(\mathrm{x}), \ e(\mathrm{x}) = e^0 + e^1(\mathrm{x}),$$
$$\eta(\mathrm{x}) = \eta^0 + \eta^1(\mathrm{x}), \ \rho(\mathrm{x}) = \rho_0 + \rho_1(\mathrm{x}) \quad (7)$$

The value with superscript of 1 represents the difference below:

$$C^1 = C - C^0, e^1 = e - e^0,$$
$$\eta^1 = \eta - \eta^0, \rho_1 = \rho - \rho_0 \quad (8)$$

And local fractional integral of $f(x)$ defined by Eq.9.

$$
\begin{aligned}
{}_aI_b^{(\alpha)} f(t) &= \frac{1}{\Gamma(1+\alpha)} \int_a^b f(t)(dt)^{\alpha} \\
&= \frac{1}{\Gamma(1+\alpha)} \lim_{\Delta t \to 0} \sum_{j=0}^{j=N-1} f(t_j)(\Delta t_j)^{\alpha}
\end{aligned} \tag{9}
$$

Its local fractional Hilbert transform, denoted by $f_x^{H,\alpha}(x)$ is defined by

$$
\begin{aligned}
H_{\alpha}\{f(t)\} &= \hat{f}_H^{\alpha}(x) \\
&= \frac{1}{\Gamma(1+\alpha)} \int_R \frac{f(t)}{(t-x)^{\alpha}} (dt)^{\alpha}
\end{aligned} \tag{10}
$$

Where x is real and the integral is treated as a Canchy principal value, that is,

$$
\begin{aligned}
&\frac{1}{\Gamma(1+\alpha)} \int_R \frac{f(t)}{(t-x)^{\alpha}} (dt)^{\alpha} \\
&= \lim_{\varepsilon \to 0} [\frac{1}{\Gamma(1+\alpha)} \int_{-\infty}^{x-\varepsilon} \frac{f(t)}{(t-x)^{\alpha}} (dt)^{\alpha} + \\
&\frac{1}{\Gamma(1+\alpha)} \int_{x+\varepsilon}^{\infty} \frac{f(t)}{(t-x)^{\alpha}} (dt)^{\alpha}]
\end{aligned} \tag{11}
$$

To obtain the inverse local fractional Hilbert transform, write again Eq. (11) as

$$
\begin{aligned}
\hat{f}_H^{\alpha}(x) &= \frac{1}{\Gamma(1+\alpha)} \int_{-\infty}^{\infty} \frac{f(t)}{(t-x)^{\alpha}} (dt)^{\alpha} \\
&= \frac{1}{\Gamma(1+\alpha)} \int_{-\infty}^{\infty} f(t) g(x-t) (dt)^{\alpha} \\
&= f(x) * g(x),
\end{aligned} \tag{12}
$$

The equation of motion is as follows:

$$
\partial_j (C_{ijkl} \partial_k u_l + e_{kij} \partial_k \varphi) - \rho \ddot{u}_i = 0 \tag{13}
$$

Under the linear theory, that is:

$$
\partial_j (e_{ijkl} \partial_k u_l - \eta_{kij} \partial_k \varphi) = 0 \tag{14}
$$

Linear equation can be expressed into the following simplified forms:

$$
L(\nabla, \omega) f(x, \omega) = 0, \tag{15}
$$

$$
L(\nabla, \omega) = T(\nabla) + \omega^2 \rho \mathrm{J} \tag{16}
$$

In which,

$$
T(\nabla) = \begin{Vmatrix} T_{ik}(\nabla) & t_i(\nabla) \\ t_k^T(\nabla) & -\tau(\nabla) \end{Vmatrix}, \mathrm{J} = \begin{Vmatrix} \delta_{ik} & 0 \\ 0 & 0 \end{Vmatrix},
$$

$$
f(x, \omega) = \begin{Vmatrix} u_k(x, \omega) \\ \varphi(x, \omega) \end{Vmatrix} \tag{17}
$$

Consider delay, the L can be expressed as:

$$
L^0 = \begin{Vmatrix} C_{ijkl}^0 & e_{kij}^0 \\ e_{ikl}^{0T} & -\eta_{ik}^0 \end{Vmatrix} \tag{18}
$$

These functions can be expressed in the following form:

$$
C(\mathrm{x}) = C^0 + C^1(\mathrm{x}), \quad e(\mathrm{x}) = e^0 + e^1(\mathrm{x}),
$$

$$
\eta(\mathrm{x}) = \eta^0 + \eta^1(\mathrm{x}), \quad \rho(\mathrm{x}) = \rho_0 + \rho_1(\mathrm{x}) \tag{19}
$$

The value with superscript of 1 represents the difference below:

$$
\begin{aligned}
C^1 = C - C^0, e^1 = e - e^0, \\
\eta^1 = \eta - \eta^0, \rho_1 = \rho - \rho_0
\end{aligned} \tag{20}
$$

The whole function can be simplified into the following integral equation set:

$$
\begin{aligned}
f(x, \omega) &= f^0(x, \omega) + \int_V S(x - x')(\mathrm{L}^1 F(\mathrm{y}') \\
&+ \rho_1 \omega^2 \mathbf{g}(R) \mathrm{T}_1 f(\mathrm{y}')] S(\mathrm{y}') \mathrm{d}\mathrm{y}'
\end{aligned} \tag{21}
$$

$f(t)$ is defined as:

$$
[\frac{\partial}{\partial t} + \varepsilon]^2 f(t) = \delta(t) \tag{22}
$$

$$
\begin{aligned}
\overline{g}(k, t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{e^{-i\omega t} d\omega}{k^2 + (\varepsilon - i\frac{\omega}{c})^2} \\
&= c^2 \Theta(t) \frac{\sin(ckt)}{ck} e^{-\varepsilon t}
\end{aligned} \tag{23}
$$

$$
\overline{g}(r, t) = \frac{1}{(2\pi)^2} \int e^{-i\mathrm{kr}} \overline{g}(\mathrm{k}, t) d^2 k \tag{24}
$$

III. EXPERIMENT RESULT

The traditional data mining algorithms can deal with small-scale data, but they may not necessarily suitable for large-scale data processing. Under this requirement, parallel data mining algorithms have emerged. As an important parallel computing technology tool, Hadoop parallel framework has attracted the attention of the business community and academia. Hadoop Framework is an academic hot issue in data mining algorithm. Apriori algorithm is one of the most typical data mining algorithms, the technical bottlenecks in large-scale data mining is that a huge amount of data are always traversing many times, and it causes I/O bottlenecks, also increases computation time. There are a lot of the optimization algorithms of Apriori algorithm, which mainly are parallel algorithms including CD (count distribution), DD (data distribution), CaD (candidate distribution) algorithm. There is another optimization Apriori algorithm which is based on Hadoop. PageRank algorithm is the core algorithm of commercial search engine, facing to the soaring number of page data, it is difficult to avoid the overhead of processing time-consuming which happened in much iteration and traverse the page data. Issued to PageRank algorithm handling large data, scholars already have a lot of achievements, such as the PageRank algorithm that does not achieve the best results by transplanting directly PageRank algorithm onto Hadoop platform. This experiment focuses on the transplantation and optimization of Apriori algorithm and PageRank algorithm in Hadoop platform. Combined with Map Reduce framework of Hadoop distributed computing platform, the proposed algorithm use a parallel connection operations called Data Join to achieve the next computation at each iteration. The figure 2 shows the

comparison of average response time of different algorithm and the figure 3 shows the average number resource nodes task allocation of different algorithm.

Hadoop and HDD algorithms are efficient and scalable parallel methods applied in the discovery of association rules in the field of data mining. However, they become less effective due to the imbalance caused by distributing the candidates among the processors. Therefore, Hadoop and HDD are improved by means of introducing the approximate algorithms to solve the problem of load balance effectively. From the experiment result, we may conclude that the proposed algorithm has better performance in average response time and the average number resource nodes task allocation.

IV. CONCLUSION

The primary objective of this project is to develop an integrated open framework for mining sensor data, hence the evaluation of the system is focus on the ability and extent that it integrates sensor network with data mining and open to new technologies. Though that the objective is achieved could be guaranteed by adopting industry standards and/or the standards, the performance of sensor network data processing algorithms as well as the performance and accuracy of JDM API, on which our system is built, compared with current mainstream data mining products still needs to be carefully evaluated in the very near future.

In the second phase of the project, we plan to support more hardware platforms in the framework with the introduction of a HAL, which provides a uniform service and interface to the upper layer while keeping the hardware detail transparent to users. Our ultimate purpose is to propose a common architecture independent from any vendor specific hardware and/or software products while open to various new emerging techniques such as routing algorithms, data mining models and/or algorithms, and so forth.

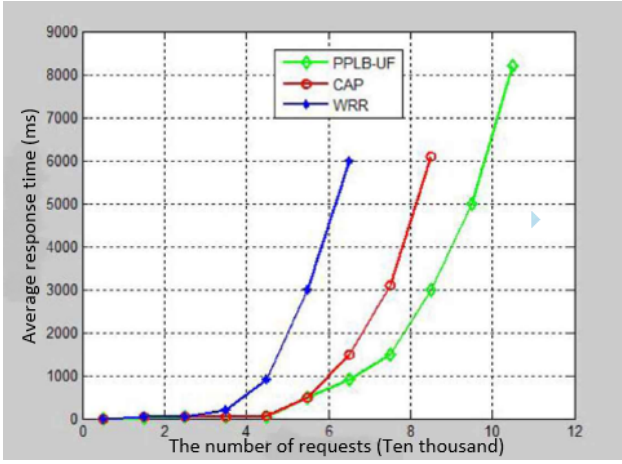

Figure 2. The comparison of average response time of different algorithm

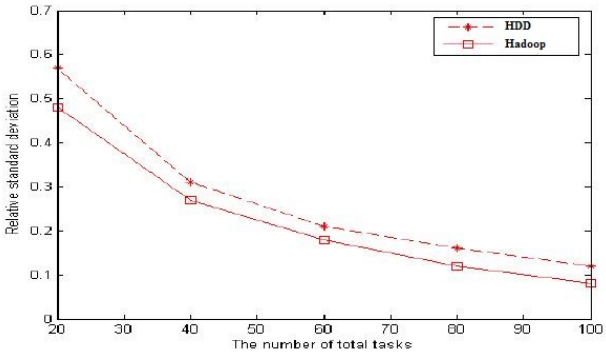

Figure 3. The average number resource nodes task allocation of different algorithm

REFERENCES

[1] A. Pisano, F. Bignami, and R. Santoleri, "Oil Spill Detection in Glint-Contaminated Near-Infrared MODIS Imagery," *Remote Sensing*, vol. 7, no.1, pp. 1112-1134, 2015. https://doi.org/10.3390/rs70101112

[2] Z. Lv, J. Chirivella, and P. Gagliardo, "Bigdata Oriented Multimedia Mobile Health Applications," *Journal of medical systems,* vol.40, no.5,pp. 1-10,May 2016. https://doi.org/10.1007/s10916-016-0475-8

[3] Z. Yang, and C. Yang, "Research on the Application of Intelligent Calibration Device for Nuclear Power Plants Based on Wireless Sensor Technology," *International Journal of Online Engineering*, vol. 12, no 05, pp. 22-26, May 2016. https://doi.org/10.3991/ijoe.v12i05.5731

[4] Y. Li, Y. Zhang, J. Chen, et al., "Improved Compact Polarimetric SAR Quad-Pol Reconstruction Algorithm for Oil Spill Detection," *IEEE Geoscience & Remote Sensing Letters,* vol. 7, no.1, pp. 1139-1142, 2014. https://doi.org/10.1109/LGRS.2013.2288336

[5] B. Reddy, "A Modified clustering for LEACH algorithm in WSN," *International Journal of Advanced Computer Science & Applications,* vol. 4, no.5, 2013.

[6] W. Wang, and Y. Peng, "LEACH Algorithm Based on Load Balancing," *Telkomnika Indonesian Journal of Electrical Engineering,* vol. 11, no.9, 2013.

[7] Z. Sun, Y. Zhang, Y. Nie, et al., "CASMOC: a novel complex alliance strategy with multi-objective optimization of coverage in wireless sensor networks," *Wireless Networks*, pp. 1-22, 2016. https://doi.org/10.1007/s11276-016-1213-3

[8] H. Jing, "Node deployment algorithm based on perception model of wireless sensor network," *International Journal of Automation Technology*,vol.9, no.3, pp. 210-215, April 2015. https://doi.org/10.20965/ijat.2015.p0210

[9] H. Jing, "Routing optimization algorithm based on nodes density and energy consumption of wireless sensor network," *Journal of Computational Information Systems*, vol. 11, no.14, pp. 5047-5054, July 2015.

AUTHORS

DU Juan is with Yellow River Conservancy Technical Institute, Henan, China (dujuan@chinauniedu.com).

WU Fenfen is with College of Transportation, Henan, China.

Submitted 09 September 2016. Published as resubmitted by the authors 10 October 2016.