Information Technology and Quantitative Management (ITQM2013)

A New Data Processing Algorithm Based on Model Fitting for Wireless Sensor Networks

Guimin Huang a, Yuanyuan Chen a, Ya Zhou b, Pingshan Liu a, Siyun Liu a

“Research Center on Data Science & Social Computing, Guilin University of Electronic Technology, Liulin 541004, China

Abstract

Wireless sensor networks (WSNs) are data-centric networks. In recent years, with the continuous development of WSNs technology, the characteristics of node storage capacity-constrained and node energy resource-constrained are prominent in WSNs technology. How to effectively compress data is an important issue in WSNs. Especially, the issue becomes a more challenging problem when the data monitored by nodes have no spatial and temporal correlation or stable correlation. To settle the above problem, this paper proposes a new data processing algorithm based on B-spline fitting model, called BFM algorithm. The BFM algorithm can compress data effectively, and thus reduce the amount of data transmitted. This can save the energy consumption of WSNs and prolong the lifetime of WSNs. Through experiments, the BFM algorithm is compared with the Tiny DB model. The results of the experiments show the BFM algorithm outperforms the Tiny DB model in terms of data transfer ratio.

© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords: Wireless Sensor Networks; Data Compression; Cubic B-spline Interpolation

1. Introduction

Wireless sensor network (WSN) is a collection of sensor, embedded computing, distributed information processing and wireless communication technology. It contains a series of wireless sensor nodes with wireless RF transmitter module, perception module for data and forms a multi-hop self-organizing network system through the wireless communication. In recent twenty years, digital data collection devices and data storage technology have great development [1]. With the rapid development of wireless communication technology and microelectronics technology, WSNs have been widely used in environmental monitoring, military applications, medical health and many other fields.

At present, WSNs are resource constraint: limited power supply, bandwidth for communication and difficult to

---

This work is supported by the Research Foundation of Key Laboratory of Wireless Broadband Communication and Signal Processing, No.11108

Corresponding author. Tel.: 18777395269; fax: 07732291153.
E-mail address: phoneixtvctfy@163.com

1877-0509 © 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management
doi:10.1016/j.procs.2013.05.147
adapt to the needs of amount of data transmission. Kimura and Latifi [2] points out that the energy consumption of data communication is much higher than data calculation. The energy consumption of data transmission accounts for the total energy 70 percent. If the data collected by WSNs can be processed and reduce the volume of data transmitted, the energy of WSNs will be reduced and thus the network lifetime will be prolonged. Therefore, how to effectively save the energy of WSNs and prolong the lifetime of WSNs is a significant topic we are facing today.

WSNs have the disadvantages that the energy is limited and hard to recharge. To reduce energy consumption for WSNs, data fusion technology is proposed. Data fusion technology focuses on how to effectively fuse and process large amounts of data collected or received in sensor nodes. In WSNs, data fusion technology can remove redundant information and reduce the amount of data transmission in the process of data collection. The data fusion technology makes use of the local computing and the storage capacity of network nodes to reduce the energy consumption of data transmission. Data fusion can also be combined with multiple protocol layers of the sensor network. It has been widely used in target tracking, automatic target recognition and many other fields.

Currently, most of the research has focused on using spatial and temporal correlation among nodes to save energy. Chen et al. [3] presented a wavelet compression algorithm, based on the temporal correlation of sensing data and the available communication bandwidth of nodes. The algorithm selects the wavelet coefficients, ready to be transmitted, determines quantization bits, coding method and adaptively adjusts compression rate of sensor. Thus the amount of output data is controlled within the acceptable error limits. In [4], Cao et al. used the part values of monitoring to extrapolate another part of the monitoring values based on the spatial correlation of sensor network nodes. Nodes, in the correlation, are working and dormancy in turn to save energy. Zhou et al. [5] present a progressive data compression algorithm based on wavelet transform. This algorithm selects sensor nodes for the using of transmitting data according to the spatial correlation of sensing data. With this method, the data unit of progressively transferring would be able to generate a large amount of coding gain. The data compression algorithm in Zhou's theory has higher compression efficiency, therefore reducing the energy consumption of network. In [6], He et al. proposed a method based on hierarchical cluster for compressing data in the process of transmission. They used different models of wavelet transformation to compress data. With theoretical analysis and experimental stimulating, their proposed algorithm has effective performance of approximation and compression of data. Eventually, the data compression algorithm will reduce the amount of data transmission for prolonging the lifetime of the whole network.

However, the algorithms for saving energy which are mentioned above cannot be applied to the case when the monitoring values of nodes have no the spatial and temporal correlation or unstable correlation. Therefore, in this paper we propose a new data processing algorithm based on B-spline fitting model, called BFM algorithm, which is independent with the spatial-temporal correlation. The BFM algorithm condenses the monitoring data collected from nodes of sensor network. We will find a best fitting model which could replace the trend of the change of large amount of data. Then, the fitting model could be transmitted by local nodes to Sink node. By using BFM algorithm, we can achieve the goal of reducing the energy consumption of data transmission and prolonging the network lifetime.

2. The Theoretical Basis for BFM Algorithm

2.1. The spline function [7]

In mathematics, a spline is a sufficiently smooth polynomial function which is piecewise-defined, and possesses a high degree of smoothness at the places where the polynomial pieces connect (which are known as knots) [12].

A Spline (function) is a piecewise polynomial, which also has some characteristic of connection among polynomials on the adjacent segment. Thus, the spline not only maintains simple feature for polynomial and feasibility of approximation, but also independent in partial properties among segments. Therefore, a spline is a particularly effective tool for approximation.

We are using a given function $s(x)$ to introduce a mathematical concept for the spline as following.
We can assume a given series of nodes $-\infty = x_0 < x_1 < \cdots < x_N < x_{N+1} = \infty$ (1)

When $s(x)$ meets two conditions as following:
1. $s(x)$ is a real coefficient algebraic polynomial whose number of times do not exceed $n$ for each interval $[x_i, x_{i+1}]$ ($i = 0, \cdots, N$).
2. $s(x)$ has been to the n-minus-oneth order derivative among $(-\infty, \infty)$.

We claim $y = s(x)$ is $n$-th spline function, normally $n$-th spline function overall based on formula (1) for the node is denoted as $S_n(x_1, x_2, \cdots, x_N)$. $x_1, \cdots, x_N$ are called spline nodes.

2.2 The Cubic B-spline [8]

In the mathematical subfield of numerical analysis, a B-spline is a spline function which has minimal support with respect to a given degree, smoothness, and domain partition. B-splines were investigated as early as the nineteenth century by Nikolai Lobachevsky. A fundamental theory states that every spline function of a given degree, smoothness, and domain partition can be uniquely represented as a linear combination of B-splines of that same degree and smoothness, and over that same partition [13].

We assume that $\{a = t_0 < t_1 < \cdots < t_N = b\}$ denoted as $N + 1$ nodes.
Let $U = \{t_0, t_0, t_0, t_1, t_1, \cdots, t_{N-1}, t_{N-1}, t_N, t_N, t_N\}$.
Suppose that $U$ is the node vector parameters of cubic B-spline curve which meet the tolerance requirements.
Because $t_0 = a$, $t_N = b$.
Let $r = \{r_1, r_2, \cdots, r_{N-1}\}$, $0 < r_i < 1$, $i = 1, 2, \cdots, N - 1$

Then $t_i = r_i t_{i-1} + (1 - r_i) t_{i+1}$, $i = 1, 2, \cdots, N - 1$ (2)

Finishing a tridiagonal equations of diagonally dominant as following

\[
\begin{pmatrix}
-1 & 1 - r_1 & 0 & 0 & 0 \\
0 & -1 & 1 - r_2 & 0 & 0 \\
0 & 0 & \ddots & \ddots & \ddots \\
r_{N-2} & 0 & \ddots & -1 & 1 - r_{N-2} \\
0 & r_{N-2} & \cdots & 0 & -1 \\
\end{pmatrix}
\begin{pmatrix}
t_1 \\
t_2 \\
\vdots \\
t_{N-2} \\
t_{N-1} \\
\end{pmatrix}
= 
\begin{pmatrix}
-r_1 a \\
0 \\
\vdots \\
0 \\
-(1 - r_{N-1}) b \\
\end{pmatrix}
\] (3)

The problem of the distribution for spline nodes is transformed into a tridiagonal matrix structure as formula (3).

Clearly, if the choice of $r = \{r_1, r_2, \cdots, r_{N-1}\}$ is right. $r$ can be substituted into equation (2), we can be easily obtained the internal node $\{t_1, \cdots, t_{N-1}\}$ by chasing method. Therefore, for a given curve, cubic B-spline interpolation curve has a good approximation.

3. The Model Building

3.1 The Problem Definition

In this paper, sensing data are collected by HANBACK electronic Zigbex wireless sensor module (HBE-USN_E Ubi-Zigbex). These types of dataset include temperature, humidity, and light. We can carry out one type of dataset to definition the problem.

For the collected light data, we select portions of the data sequence,
Set a given range \([a, b]\), and \(a = x_0 < x_1 < \cdots < x_N = b\), the sequence numbers are from 0 to \(N\).

Any given set of monitoring data sequence \(L(x) = y_0, y_1, \cdots, y_N\) with setting specified tolerance, which based on the idea of the B-spline interpolation \([7]\), can be used to find a \(s(x)\). \(s(x)\) is satisfied the following conditions:

1. \(s(x) \in L(x_0, x_1, \cdots, x_N)\)
2. \(s(x_i) = y_i, i = 0, 1, 2, \cdots, N\)
3. \(\max_{a \leq x_i \leq b} |L(x) - s(x)| \leq \varepsilon\)

\(s(x)\) is known as the best model \(M\).

### 3.2 FM Algorithm

FM algorithm bases on the model fitting. The purpose of FM algorithm is to return fitted sensing data collected by the sensor network to the Sink node.

And a part of pseudo-code for FM algorithm is in following passage:

```plaintext
Input: monitoring data sequence \(L(x) = y_0, y_1, \cdots, y_N\), tolerance \(\varepsilon\)
Output: Model \(M\)

1 for \(x_i\) from \(a\) to \(b\), \(i\) from 0 to \(N\) do
2   for each \(s(x) \in L(x)\), \(s(x_i) = y_i, i = 0, 1, 2, \cdots, N\) do
3     if there is a fitting model \(M\), such as \(\max_{a \leq x_i \leq b} |L(x) - s(x)| \leq \varepsilon\) then
4       if there exists \(\text{Min} \sum [L(x) - s(x)]^2\) returns the model \(M(x)\)
5       return
6     end
7   end
8 end
9 end
```

For a given sensing data sequence \(L\), our goal is to find the best-fitting model \(M\). This algorithm for finding the best-fit model \(M\) is defined as the FM. The main idea of this algorithm is fitting in a large number of monitoring values stored in the sensor nodes. So there is the existence of a model \(M\), which can be fitted all monitoring values within a user-specified error tolerance range of this sequence. Thus the FM algorithm achieves the purpose of maximizing the compression of data.

### 3.3 BFM Algorithm

We know that the data fitting is divided into polynomial interpolation and spline interpolation in \([9]\). Spline interpolation has a better effect than polynomial interpolation in fitting the data collected by us. In the second section of the paper we introduce the theoretical knowledge of the Spline function and Cubic B-spline. As we know that cubic B-spline fitting can compress large amounts of data.

There are some disadvantages: the energy and bandwidth of WSNs are limited, which is difficult to adapt to the needs of a large number of data transmission. According to some of the characteristics of the sensor network itself, how to reduce energy consumption is a vital issue. In order to solve the problem of energy consumption, we present a BFM algorithm based on cubic B-spline interpolation and FM algorithm as we mentioned above.

The following is a specific description about this algorithm:

Set maximum data compression ratio is \(t\); Model function is \(M(x)\). Elements which are not fitted in the sequence \(L\) are \(f\).

Then \(t(M) = \frac{s(x) - \sum_{i=1}^{t} s(x_i)}{s(x)}\), where \(t(M)\) is transfer ratio, \(M\) is the best model.
Input: monitoring data sequence $L(x): y_0, y_1, \cdots, y_N$, tolerance $\varepsilon$, the number $f$
Output: Model $M(x)$
1 for $x_i$ from $a$ to $b$, $i$ from $0$ to $N$ do
2 for each $s(x_i) \in L(x)$, $s(x_i) = y_i$, $i = 0, 1, 2 \cdots, N$ do
3 Construct $t(M) = \frac{s(x)}{s(x) + \sum_{i=1}^{f} L(x)}$
4 if $t(M) = 1$, such that for $\max_{x \in X; i \in B} |L(x) - s(x)| \leq \varepsilon$ then
5 return model function $M(x)$
6 end
7 end
8 end
9 end

Fig. 1. Algorithm flowchart

The flowchart of the BFM algorithm shows in Fig. 1. **Step 1**: The data are collected by sensor network for monitoring area, and then each node in WSNs is not immediately transferred back to the sink node. **Step 2**: The data is stored in the local memory of node. **Step 3**: When the amount of monitoring data cached in the node local memory reach a given threshold, based the theory of Cubic B-spline, we compression of data which is evolved from fitting process. **Step 4**: The best fitting model $M$ can be found and then the model parameters are transferred back to the Sink node. The data fitting process is carried out by each sensor node, which is exclusively independently. So the BFM algorithm is getting rid of the dependence for the spatial correlation of nodes.
4. Experiments and Analysis

4.1 Experiment Illustration

In Fig. 2, we take the part of values of sin function as an observation sequence, respectively polynomial and B-spline for fitting approximation. The blue curve represents B-spline fitting, while the red curve represents the polynomial interpolation fitting. Clearly, we can see the interval in red curve, but hardly see the interval in blue curve. The blue curve has a higher approximation than the red curve. In conclusion, the B-spline interpolation fitting is better.

Based on the above viewpoint, this paper approves a method of spline and proposes a new data processing algorithm called BFM (a fitting model based on cubic B-spline).

In this paper, the information can be collected and monitored within short distances. Therefore, the topology of simple star network is applied. Each node of star topology is linked to the central node by point-to-point communication lines. Star topology has the following advantages: simple structure, easy to implement, manage and monitor. In Fig. 3, all terminal nodes are connected to the aggregation node. And the data is transferred to the computer by the aggregation node, converted into the required data through the serial port communication software.
4.3 Experimental Set

In this paper, the experimental dataset is collected by Zigbex wireless sensor module of HANBACK electronic (HBE-USN_E Ubi-Zigbex). The dataset contains 30-day monitoring data, which are derived from 10 sensor nodes. These data are samples of temperature, humidity and light. However, monitoring data collected by the sensor module can't be used directly. To solve the problem of data conversion and storage in the experiment, we developed a serial communication software.

The serial communication software parameter settings: the connection port is COM3; the baud rate is 57600; the number of data bits is 8; the number of stop bit is 1; the data cache location is D:\COMDATA; the data is saved as the form of decimal.

Fig. 4 shows the change trend of monitoring value of humidity continuously periodically with time. And there is similar trend of change which we collected by data in other applications [10]. In order to reduce data redundancy and the energy consumption of the data transmission, we use the BFM algorithm to maximize the fitting of monitoring data for such trend.

4.4 Experimental Analysis

Currently, when network nodes have no the space-time correlations, there is a traditional detector data acquisition strategy of University of California, Berkeley, the Tiny DB [11], transmitting all monitoring data to Sink node. Our
strategy is just returning a best-fitting model $M$. In figure 5, we compare the method of BFM algorithm with Tiny DB model. It can be seen in Fig. 5 that no matter indoors and outdoors, our methods both achieved a comparatively small data transfer ratio, saving a large amount of energy of data transmission at the same time. Although the BFM algorithm will still lose a small amount of accuracy, but it won’t impede majority of applications. We develop the BFM algorithm by C++, and use some monitoring datasets as the input of the BFM algorithm. When the nodes of sensors are in the conditions which have no correlation of spatial and temporal, monitoring data are saved relatively independent in the local memory and fitted in each node. The maximum of data condense is 5%, while the minimum is about 38%.

5. Conclusion

In view of the characteristics of WSNs that the energy is limited and hard to recharge, we propose a new data processing algorithm, called BFM algorithm which is based on the cubic B-spline. The BFM algorithm can achieve the purpose of data compression by transmitting a fitting model and its parameters, instead of the monitored values collected by nodes. Though experimental stimulate, the BFM algorithm can reduce the amount of data during transmission and thus extend the network lifetime. Also, The BFM algorithm can guarantee reliable confidence within a certain error tolerance range. Furthermore, the BFM algorithm is independent on the spatial and temporal correlation of the data. Therefore, when the monitoring values of data have no or low spatial and temporal correlation, the BFM algorithm also have the ability of work normally. For the several reasons in the paper, the future prospects of the BFM algorithm are promising.

Acknowledgments

We would like to thank teachers and peers who guide the study of data processing based on model fitting for wireless sensor networks.

References

[1] Peng, Y., Kou, G., Shi, Y., and Chen, Z. A Descriptive Framework for the Field of Data Mining and Knowledge Discovery, International Journal of Information Technology & Decision Making, Vol. 7, Issue: 4, Page 639 – 682, 2008.
[2] Kimura N, Latifi S. A survey on data compression in wireless sensor networks. In: Proc. of the Int’l Conf. on Information Technology: Coding and Computing, Vol.2. Piscataway: IEEE Press, 2005. 8–13.
[3] Huamin Chen, Jian Li, Mohapatra P. RACE: Time Series Compression with Rate Adaptivity and Error Bound for Sensor Networks[C]//Proc. of IEEE International Conference on Mobile Ad-hoc and Sensor Systems. [S. L.]: IEEE Press, 2004.
[4] Q Cao, T Abdelzaher, T He, et al. Towards optimal sleep scheduling in sensor networks for rare event detection[C]. Information Processing in Sensor Networks, Los Angeles, USA, 2005.
[5] Siwang Zhou, Yaping Lin, and Songtao Ye. A storage wavelet progressive data compression algorithm in sensor networks [J]. Computer Research and Development, 2009, 46(12): 2085-2092.
[6] Zhiyong He, Chenfeng Long, and Tongsen Wang. Data compression algorithms of hierarchical cluster model for sensor networks [J]. Computer Engineering, 2009, 35(13): 105-108.
[7] Spline functions, B-spline, Cubic spline interpolation[EB/OL]. http://www.docin.com/p-174800058.html#documentinfo.
[8] Shiqian Dai. Approximation algorithm of adaptive cubic spline interpolation [D]. Dalian: Dalian University of Technology, 2008: 1-45.
[9] Qingyang Li, W. Y. Numerical analysis [M] Beijing: Tsinghua University Press, 2008.
[10] G Tolle, Sonoma redwoods data[OL], http://www.cs.berkeley.edu/~get/sonoma, 2005.
[11] S. Madden, M. J. Franklin, J. M. Hellerstein, et al. The Design of an Acquisitional Query Processor for Sensor Networks[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. San Diego, USA, 2003: 491-502.
[12] Judd, Kenneth L. (1998). Numerical Methods in Economics. MIT Press. p. 225. ISBN 978-0-262-10071-7;
[13] Carl de Boor (1978). A Practical Guide to Splines. Springer-Verlag. pp. 113–114.