MuMoCo: a Framework for Improving Data Quality with Multiple Modalities Cooperation in Wireless Sensor Networks

Guo-Ying WANG[1] and Shen-Ming QU[2]

[1] Information Engineering College, Zhejiang A&F University
Hangzhou, 311300, China

[2] Computing Center，Henan University
Kaifeng, 475004, China

Abstract

High quality sensor data stream is crucial to wireless sensor networks applications. However raw data streams in wireless sensor networks tend to be not reliable. Therefore, improving sensor data quality is an important issue for all kinds of wireless sensor networks applications. In this paper, we proposed an integrated framework, *MuMoCo*, which is based on such a fact: the factors leading to outlier or data missing such as events, insufficient power, or malicious nodes have similar influence on each modality of data created by a node at the same time. Not considering the correlation among different modalities of data may probably lead to a contradictory: different conclusions of data verification according to different modalities of data. Taking advantage of multiple modalities cooperation of sensor data to avoid such contradictory and improve data quality, the *MuMoCo* framework includes data source quality assessment, data authenticity verification and recovery, and data conversion with quality assurance. Using the *MuMoCo* framework, we can obtain following benefits: more efficient filtering of data source based on data source quality assessment; more accurate data authenti-city verification and data recover; and more comprehensive utilization of sensor data.

Keywords: *wireless sensor networks, data quality improvement, multiple modalities cooperation, data clean, outlier detection, data recovery.*

1. Introduction

Wireless sensor networks are composed of a group of self-organized sensors, and the sensor nodes cooperate to sense, collect and process the information of objects in the coverage area of network and propagate the information back to the observer. One of the important characteristics of the wireless sensor networks is data-centric. Furthermore, the sensor data volume is very huge. The applications of wireless sensor networks need extract required valuable information from the sensor data sea. Therefore the data quality of wireless sensor networks directly affects the correctness and accuracy of information extracted and final conclusion drawn. High quality sensor data stream is crucial to the successful of wireless sensor networks applications.

However raw data streams created in wireless sensor networks tend to be not reliable, containing incomplete, inaccurate, incorrect, inconsistent and other types of noise. The probability of creating erroneous data will sharply increase when the battery power of a sensor node is insufficient [1]. On the other hand, a sensor node can be affected by the environment and produce erroneous data, especially when it was deployed in harsh environments. These internal or external factors can lead to unreliable sensor data, which may contain some "dirty" data called outlier. Furthermore, sensor data may also be lost due to some reasons such as the node failure or network congestion. The unreliable data will affect the quality of the raw data and the final fusion conclusions. Events occurred in real world such as forest fires and chemical leaks can not be determined using inaccurate and incomplete data [2], so it is very important to ensure the reliability and accuracy of sensor data before making a determined decision.

Therefore, improving the quality of sensor data is an important fundamental issue that all kinds of wireless sensor networks applications must face. For example, in the applications using localization and tracking in wireless sensor network, high quality sensor data can improve the accuracy of positioning and tracking; in the event monitoring oriented applications such as forest fire and chemical leak alarm, high quality sensor data can improve the accuracy of monitoring events alarm, reduce the false positive rate and false negative rate.

Compared with the data in a common database or data warehouse, sensor data in wireless sensor networks have significant features of spatial and temporal correlations, which mean that there exist some relations between the data of neighbor nodes at the same time and between the data of a node at different time. Such correlations of

sensor data can be used to improve data quality.

However, many researches mainly consider the spatial and temporal correlations of a single modality data, without taking into account the correlation between different modalities of data, such as temperature, humidity, light and so on.

In this paper we study methods for improving sensor data quality for wireless sensor networks applications by means of the cooperation of multiple modalities of sensor data, mainly taking into account the correlations between data of different modalities and relations among the sensor data quality of different modalities in the same sensor node, as well as spatial and temporal correlations of a single modality data.

An integrated framework, *MuMoCo*, is proposed in this paper. The *MuMoCo* framework is based on such a fact: the factors leading to "dirty" data or data missing such as events, insufficient power, or malicious nodes have similar influence on each modality of data created by a node at the same time. Not considering the correlation among different modalities of data may probably lead to a contradictory: different verification conclusions of the data from a node at the same time are drawn using different modalities of data. Multiple modalities cooperation of sensor data can easily avoid such contradictory and improve the quality of sensor data.

The *MuMoCo* framework includes: data source quality assessment; data authenticity verification and recovery; and data conversion with quality assurance. Data source quality assessment is used to filter faulty sensor nodes. Data authenticity verification can identify "dirty" sensor data in data stream and data recovery can be used to solve the problem of missing or erroneous sensor data. Data conversion with quality assurance can help the quantity and form of data meet the requirement of specific practical applications without quality discount.

2. Related Work

Data quality problems exist not only in wireless sensor networks, but have existed since the beginning of the existence of data. Data are increasing sharply in information age nowadays, so the quality of data is a problem which can't be ignored [3]. Initial studies on data quality management problem in the fields of database and data warehouse focused on the definitions and parameters of data quality [4][5].

2.1 Data quality in wireless sensor networks

In wireless sensor networks, the data created by each node exist as data stream style [6]. Therefore, compared with ordinary data quality, the quality of the data in the data stream processing need also consider data rate adaptation problems such as data up-sampling and down-sampling [7].

Sha proposed a consistency-driven data quality management framework for wireless sensor networks, in which data quality is integrated into the energy-efficient design of a sensor system. They defined a consistency model including the time consistency and digital consistency, and the model also considered the requirement of application and data dynamic property of sensing area [8].

Yates [9] discussed the tradeoff problem of data quality and query cost, proposed and evaluated several methods of query and cache data in the sensing regional server, studied strategies of assessing cache hit rate using calculation and the approximate value of the sensor data for some application requirements. The strategies can improve concurrency quality and save cost. The reason for the win-win is that system delay is very important, the benefits of query cost and use of the approximate data exceed the negative impacts of the data quality due to the approximate value. On the contrary, there is a linear trade-off between the query cost and data quality when data quality was data accuracy driven.

Environment monitoring is one of the most important applications of wireless sensor network. The success of these applications depends on the quality of data collected. Careful analysis of the collected sensor data is very important, which not only can help us recognize the characteristics of the monitored area, but also can reveal the limitations and opportunities that should be considered in the future design of the sensor networks systems.

Through analyzing sensor data from a real-world water monitoring applications and examining the similarity, abnormality and failure mode of data, researchers reached following conclusions: (1) information similarities such as similar patterns and numbers similarity is common, which provides a good chance to trade-off between energy efficiency and data quality; (2) analysis of correlations of space and modalities provide a method to assess the consistency of data and discover data conflicts, which mean sensor failure or event; (3) external severe environmental conditions may be the most important factor of conflict failure. As the research found that the

main type of failure is communica-tion failure due to the lack of the synchronization [10].

2.2 Data quality improvement for wireless sensor networks

The Detection of outlier value (i.e., the "dirty" data) in the sensor data is a necessary step to reduce the impact of the noise data. According to the physical characteristics of the data and the range of data value, abnormal values can be identified from a single data itself [11] or a multiple data together [12], and can also be identified using the relevance between sensor data and the history data of the same node (temporal correlation) and neighbors (spatial correlation) [13] [14].

The methods of outlier value detection mainly include statistics based, the nearest neighbor based, cluster based, classification based, spectral decomposition based and so on. Wu et al. proposed a method to identify abnormal sensors and event boundaries using a local technology based on statistical Gaussian model [15]. Preetha proposed a non-parametric algorithm for the identification of outliers through the enhancement of an association classification approach using FP-Growth, which can calculate the minimum support and minimum confidence automatically [16].Bettencourt et al proposed a local abnormal value discovery technology, which can be used to identify errors and events found in the ecological field of application of wireless sensor networks [17]. Zhang's research ensure high data quality with online outlier discovery technology, and propose abnormal values detection technical based on support vector machine of a quarter-spherical. In order to reduce the false alarm rate and increase the detection rate of abnormal values, spatial and temporal correlations are used to detect outlier collaboratively [18]. Zhang also proposed a statistics-based outlier detection methodology using time-series analysis and geostatistics, taking advanta-ge of the spatial and temporal correlations [19].

The estimation of expected value of sensor data is another basic problem about data quality improvement, which can solve data missing or erroneous data rectification. Using a layered non-supervised fuzzy ART neural network to express the prototypes of the data set and express the input mode of missing data based on network, missing sensor data in wireless sensor networks can be estimated [20]. Using temporal and spatial correlations, sensor nodes can be dynamically divided into clusters, and nodes in the same cluster have similar monitoring time series. The time correlation can also be explored for energy saving. Some other researchers proposed a general framework to address several important issues, including how to divide the sensor into clusters, how to dynamic maintenance of clusters according to environmental changes, and how to schedule sensor in cluster, how to explore temporal correlation, how to store data at the sink node [21]. Petrosino et al. proposed a neuro-fuzzy recession method to clean sensor data, using the well known ANFIS model to reduce the uncertainty of the data, in order to obtain a more accurate sensor data estimates [22].

In the research of Hermans et al., the quality of sensor data can be assessed. Data fusion can be done based on the estimated quality. The system consists of local and distributed heuristics to assess the quality of the data, focusing on the accuracy and consistency of the data. During the data fusion, the values of sensor data can be inferred by the multiple sensor nodes using Dempster-Shafer evidence theory. The quality assessment and data fusion are carried out in the network, and therefore do not depend on a strong sink node [23].

3. *MuMoCo* Data Quality Improvement Framework

The *MuMoCo* framework, shown in Figure 1, mainly include following three aspects: (1) data source quality assessment based on data quality feedback; (2) data authenticity verification and recovery with multiple modalities cooperation; and (3) data conversion with quality assurance, including quantity-quantity and quantity-quality conversion.

3.1 Data source quality assessment based on data quality feedback

Footnotes Sensor data are created by sensor nodes, so high quality data source can provide high quality sensor data. Data source quality can be estimated according to the quality of data the sensor node provided. The quality of a single data source can be regarded as good if it provides high quality sensor data, or be regarded as poor if it provides low quality sensor data. As to the combined quality of multiple data sources, it can be assessed by the evaluation of whether the sensor data they created are beneficial to the improvement of data quality. The quality of data source is an important basis for filtering data source.

To assess the quality of sensor nodes, following aspects need to consider.

Data quality feedback: The higher the tendency of data from a node determined to be "dirty" is, the lower quality of the node is considered. According to different weight of each modality sensor data in specific application, different score-deduction is assigned to each "dirty" data and missing data for each modality, respectively, and continuous several "non-dirty" data may leads to a score-plus.

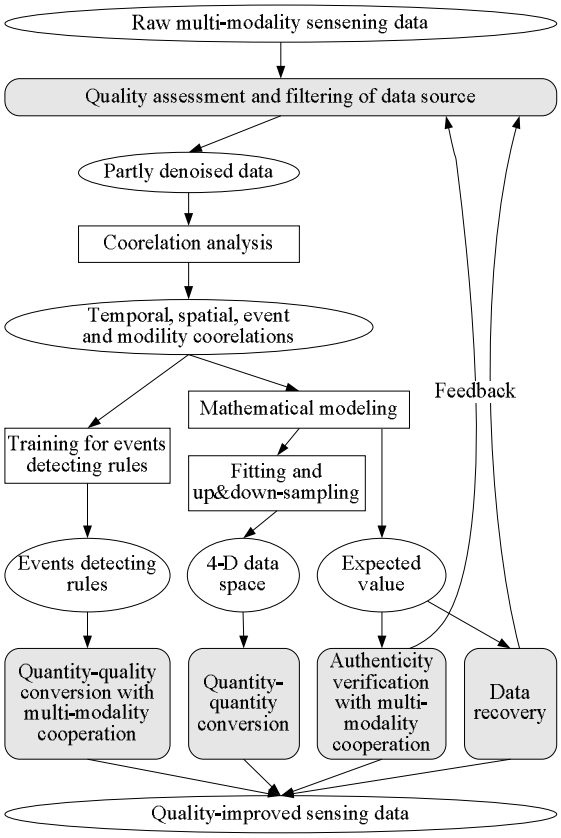

Fig. 1 Proposed *MuMoCo* data quality improvement framework for wireless sensor networks.

Fox example, after integrating the assessment results of different modalities of data using weighted collaborative algorithm, if the percents of "dirty" data from a particular data source at given time and space sliding window reaches a certain threshold, the quality of the data source is considered very poor and should be filtered out, so as not to affect the accuracy of correlation analysis.

Data authenticity verifiability: In wireless sensor networks, there may be some potential unsafe problems, which may lead to some "dirty" content in sensor data include. Whether these "dirty" data can be identified according to sensor data provided by multiple data sources is also a factor of quality assessment of data source.

Smaller spatial distance of sensor data and higher sampling frequency can lead to more accurate result of data authenticity verification according to the spatial and temporal correlations of sensor data. So different score-pluses or score-deductions are assigned to different average distance of nodes in wireless sensor networks, and are assigned to different sampling frequencies. According to the difference among the temporal and spatial correlations of different modalities of data, specific score weights are assigned to each data modality.

Data recoverability: Data missing caused by node failure or network transmission packet loss in wireless sensor networks is inevitable. So the possibility of recovering the missing data or "dirty" data using known information and the recovery quality are important aspects for the quality assessment of multiple data sources. The way of assessing data source quality by data recoverability is similar to the way by data authenticity verifiability, in addition to the evaluation of the quality of the recovered data.

According to aforementioned several factors, we proposed the model for multiple data sources quality assessment, which is based on the different impacts of various factors on the quality of data sources. The model is expressed as following:

$$Q = DQ*W_{dq} + CR*W_{cr} + CV*W_{cv} \quad (1)$$

$$DQ = f_{dq}(a, Score_d, Score_m, Score_n) \quad (2)$$

$$CR = f_{cr}(a, d, h, p, f, m, s) \quad (3)$$

$$CV = f_{cv}(a, d, h, f, m, s) \quad (4)$$

In these equations,

- *DQ,* the data quality, represents the degree of data source meet the requirement of application, which is determined by some parameters such as specific application type *a* (localization, events alarm and so forth.), score of "dirty" data $Score_d$, score of missing data $Score_m$, score of normal data $Score_n$ and so on.
- *CR,* the capability to recover "dirty" data, represents the assessment of data recoverability, which is determined by following parameters: specific application type *a*, average horizontal distance of nodes *d*, average height of nodes *h,* the precision of sensor data from nodes *p*, the frequency of node's sensing action *f*, data modality *m*, the stage of the network *s* (initial, normal or aging stage) and so on.
- *CV*, the capability to verify data authenticity, represents the assessment of data authenticity verifiability, which is a function of specific application type *a*, average horizontal distance of nodes *d*, average height of nodes *h*, the frequency of node's sensing action *f*, data modality *m*, the stage of the network *s* and so on.
- W_{dq}, W_{cr}, and W_{cv} represent the weights assigned to the

three factors *DQ, CR* and *CV* respectively, according to the influence of each factor to data quality in specific application.

3.2 Data authenticity verification and recovery with multiple modalities cooperation

Lack of electricity power or harsh environments can make sensor nodes create "dirty" sensor data, and node failure or network congestion may lead to the missing of some sensor data. There are close relations between data created by a sensor node and its physical location and acquisition time, which mean the spatial correlation and temporal correlation respectively. Different modalities of sensor data may show different spatial or temporal correlations. The quality of sensor data depends on the status of the sensor node, so the quality of different modalities of sensor data created by a node at the same time is nearly equal. For example, the nodes with insufficient power or malicious nodes may provide untrue data, no matter which modality the data is. Therefore, when a node provides multiple modalities of sensor data, there are close relations among the quality of different modalities of data. Using spatial and temporal correlation of sensor data, taking into account the relations among the quality of different modalities of data, we can verify the authenticity of the data and recover missing sensor data.

According to the characteristics of different modalities of sensor data, data authenticity verification and data recovery can be done based on correlations among sensor data. Following factors should be considered.

Temporal correlation: Sensor data created by a node in different time have certain relations, according to which, data authenticity verification and data recovery can be done using other sensor data from the same node.

Spatial correlation: Based on spatial position relations of sensor nodes, there exists certain relevance among sensor data from different nodes at the same time, according to which, data authenticity verification and data recovery can be done using sensor data from the adjacent nodes.

Event correlation: Some events or environmental status change such as shade of trees, position of water or heat also have influence on sensor data from nodes, and therefore the regulation of the impact of the common event factors on each modality of data should be considered, so as to verify data authenticity and restore data more accurately.

Modality correlation: The three correlations mentioned above are all based on the regulations of effect of time, positions and events on each single modality of data, without considering the relation of multiple modalities of sensor data. In fact, the factors leading to "dirty" data or data missing such as events, insufficient power, or malicious nodes have similar influence on each modality of data created by the same node at the same time.

Therefore, multiple modalities cooperation are particularly considered for data authenticity verification and data recovery in this paper, that is, events detection and expected value estimation of sensor data are done using the cooperation of multiple modalities of sensor data. For example, each modality of data from a node at the same time may be determined as not "dirty" according score of the same modality of data, but they may be considered as "dirty" if adding scores of all modalities of data up together with assigned weights. On the other hand, if a modality of data is determined strongly as "dirty", other modalities of data from the same node at the same time may be considered as "dirty" no matter how much their scores are.

As to different data modalities (temperature, humidity, CO_2, etc.) and different events (shade of trees, fire, water, etc), according to the cooperation of time, space, event and modality, the model of data recovery with quality assurance is expressed as:

$$D_{recover} = f_{recover}(M_n, W_n, D_t, D_s, E) \quad (5)$$

In this equation,

- $D_{recover}$ means data recovered for the case of data missing.
- M_n means different modalities of data.
- W_n means the weights according to the effects of the authenticity determining result of different modalities of data on the data recovered.
- D_t means the sensor data created by the same node at former time.
- D_s means the sensor data from adjacent nodes.
- E means events and environmental information.
- $f_{recover}$ means a function which recover sensor data from M_n, W_n, D_t, D_s, and E.

Compared with traditional data recovering technologies, data from wireless sensor networks are much easier to recover for the strong correlations among data.

Using the same method of data recovery, the expected value of data can be obtained. If a value of sensor data is far from its expected value, the data should be considered as "dirty" data, otherwise, the data can be determined as not "dirty". As a result, we can verify the authenticity of sensor data created by sensor nodes during the networks working, as following expression:

$$V = f_{\text{verify}}(D_{node}, D_{expect}) \quad (6)$$

$$D_{expect} \approx D_{recover} \quad (7)$$

In above equations,

- V means the verifying result of sensor data authenticity.
- D_{node} means the raw sensor data created by a sensor node.
- D_{expect} means the expected value of sensor data.
- f_{verify} means a function which compare senor data D_{node} with the expected values D_{expect} to verify the authenticity of sensor data.

The reason of the relation of D_{expect} and $D_{recover}$ is that the methods of getting these two values are basically identical and slightly different. Data recovery needs more accuracy and is used for data missing which occurred occasionally, so the computing complexity can be a little high. While data authenticity verification is to compare original sensor value with expected value and need not very high data accuracy, and is a regular operation to almost all sensor data, so less amount of correlative sensor data should be used in order to reduce the computing complexity.

3.3 Data conversion with quality assurance

Sometimes it is necessary to do in-depth analysis on sensor data, so as to get some non surface information, and to draw some non-intuitive conclusions. Sensor data are usually the direct monitoring result of sensor nodes, and are expressed with some quantitative values. But to extract conclusion information from a huge number of sensor data in a practical application, only quantitative values are not sufficient, and the conclusion may not directly use a certain modality data. Therefore comprehensive analysis on the sensor data is required before drawing a qualitative or quantitative conclusion. The models for the quantity-quantity, quantity-quality conversion of sensor data with multiple modalities are studied here.

(a) Quantity-quantity conversion based on multi-dimensio-nal model fitting

When the intervals of raw data created by sensor nodes are not identical with the intervals of data that the actual application wants to use, quantity conversion of the sensor data in the aspect of time is needed. When the locations or number of sensor nodes are not the same with the positions or number of sensor data needed by a practical application, quantity conversion on the sensor data in spatial aspect should be done. It is more likely that the quantity conversions in both temporal and spatial aspects are all needed, which means the up-sampling or down-sampling of sensor data in the dimensions of time and space.

According to different modalities of sensor data, fitting sensor data in temporal and spatial dimensions, a four-dimension (3d of space and time dimension) continuous space of sensor data are constructed using interpolating methods. Then the fitting data of given time and give position can be easily acquired from the four-dimension data space. The model of quantity-quantity conversion can be expressed as

$$DO_n = f_{quantity\text{-}quantity}(DI_m, \langle P, T\rangle_n) \quad (8)$$

In above equation,

- DO_n means the output n sensor data after conversion.
- $f_{quantity\text{-}quantity}$ means the fitting four-dimension space of sensor data.
- DI_m means the input m sensor data for the conversion.
- $\langle P, T\rangle_n$ means the given positions of n points in the four-dimension space of sensor data, in which P means the position in real world 3d space and T means the point of time.

(b) Quantity-quality conversion based on event recognition rules

Sensor data are usually presented as numerical quantitative form. If a practical application needs a qualitative conclusion such as forest fires, chemical leaks or other events, analysis and conversion of one or more modalities data from one or more sensor nodes according to the characteristics of the various modalities data are needed to obtain the qualitative conclusions of event determination results.

The conversion from quantitative sensor data to qualitative expression relies on the distribution characteristics of all modalities of data and the qualitative identifying rules of sensor data. According to the distribution characteristics of each modality of data during events, the rules to recognize events can be obtained through training. Such rules can be used to perform intelligent deduce based on sensor data. The model of quantity-quality conversion can be expressed as:

$$RS = f_{train}(D, E_n) \quad (9)$$

$$O = f_{quantity\text{-}quality}(DI_m, RS) \quad (10)$$

In the equations above,

- RS means the rule set to recognize events occurred and environmental status change.
- f_{train} means training procedure to find rules.
- D means a great deal of sensor data in the training set.
- E_n means the events and environmental status to train to recognize.
- DI_m means the inputted quantitative sensor data to do events reorganization based on the rule set RS.

- *O* means the qualitative conclusion outputted.

3.4 The usage of *MuMoCo* framework

To put *MuMoCo* framework into practice, some steps described here can be followed.

(1) Deploy wireless sensor networks application platform, and collect some sensor data with multiple modalities as the training set.

(2) Combining the characteristics of various modalities of data, find the temporal correlation, spatial correlation, events correlation of each modality data, and correlation among different modalities of data.

(3) According to the change of the temporal, spatial correlation of different modalities of data in case of different events, train the training set for events recognition rules, and then analyze the regulations of the variation of sensor data when specific event occurred. Integrating the rules of different modalities of sensor data for the same event, decide the parameters of the quantity-quality conversion model of sensor data.

(4) Model the correlations of time, space and modality, and then design the algorithms for expected value of sensor data. Using the algorithms, construct the models for quantity-quantity conversion, data authenticity verification and data recovery.

(5) According to the assessment of authenticity verifiability and recoverability of sensor data and quality of results of authenticity verification and recovery as feedback, evaluate the quality of data source. The evaluating result can be used for the filtering of data source. Fox example, the quality of the data source can be considered very poor and should be filtered out, so as not to affect the accuracy of correlation analysis, if the percents of "dirty" data from a particular data source at given time and space sliding window reaches a certain threshold.

(6) After sensor node quality assessment and sensor node filtering, go to step (2) again to optimize each kind of correlation.

Using such a loop structure with feedback, the training set of data is examined again and again, and the results of events decision rules, quantity-quantity conversion, data authenticity verification, data recovery and data source quality assessment are more and more accurate. A certain number of times later, the precision reach a certain degree, then the feedback does not need to carry out any more, and the rules and parameters of each model can be used for later data quality improvement.

4. Conclusion and Future Work

Improving sensor data quality for wireless sensor networks has important significance in the practical applications of various fields. In this paper, we proposed a framework of data quality improvement for wireless sensor networks, *MuMoCo*, emphasizing multiple modalities cooperation of sensor data. Unlike using single modality data for data quality improvement, *MuMoCo* can avoid the potential contradiction of data quality assessment conclusions using the correlation of single modality of data.

Using *MuMoCo* framework in wireless sensor networks, we can obtain following benefits: more efficient filtering of data source based on data source quality assessment; more accurate data authenticity verification and data recover; and more comprehensive usage of sensor data, which means the quantity-quantity and quantity-quality conversion of sensor data, and can improve the data quality for wireless sensor networks applications finally.

We proposed the structure of the *MuMoCo* framework in this paper. However, we have not described its detail implementation. In practice, each module of the framework can be implemented in centric or distributed way. The centric method is easy to implement but is too energy-consumed for sensor nodes, and the distributed method is just the opposite. The implementation and evaluation of *MuMoCo* framework are the following work to do.

Acknowledgments

This work is supported by Zhejiang Provincial Natural Science Foundation of China (LY12F02016), and Henan provincial Natural Science Foundation of China (112300410009; 112300410125).

References

[1] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, & D. Gunopulos, "Online outlier detection in sensor data using non-parametric models", in Proceeding of the 32nd International Conference on Very Large Data Bases, 2006, pp. 187-198.
[2] F. Martincic & L. Schwiebert, "Distributed event detection in sensor networks", in Proceedings of the International Conference on Systems and Networks Communication, 2006, pp. 43-48.
[3] T. C. Redman, "Data quality for the information age".

Publisher: Artech House, 1996.
[4] D. M. Strong, Y. W. Lee, & R. Y. Wang, "Data quality in context", Communications of the ACM, 1997, Vol. 40, No. 5, pp. 103–110.
[5] Y. Wand & R. Y. Wang, "Anchoring data quality dimensions in ontological foundations", Communications of the ACM, 1996, Vol. 39, No. 11, pp. 86–95.
[6] J. Gama & M. M. Gaber, "Learning from data streams processing techniques in sensor network", Publisher: Springer Berlin Heidelberg, 2007.
[7] A. Klein, H. Do, & W. Lehner, "Representing data quality for streaming and static data", in Proceedings of the International Workshop on Ambient Intelligence, Media, and Sensing (AIMS), 2007, pp. 3-10.
[8] K. Sha & W. Shi, "Consistency-driven data quality management of networked sensor systems", Journal of Parallel and Distribted Computing, 2008, Vol. 68, No. 9, pp. 1207-1221.
[9] D. J. Yates, E. M. Nahum, J. F. Kurose, & P. Shenoy, "Data quality and query cost in pervasive sensing systems", Pervasive and Mobile Computing, 2008, Vol. 4, No. 6, pp. 851-870.
[10] K. Sha, G. Zhan, S. Al-Omari, T. Calappi, W. Shi, & C. Miller, "Data quality and failures characterization of sensing data in environmental applications", CollaborateCom, 2008, No. 10, pp. 679-695.
[11] P. N. Tan, M. Steinback, & V. Kumar, "Introduction to data mining". Publisher: Addison Wesley, 2006.
[12] P. Sun, "Outlier detection in high dimensional, spatial and sequential data sets", Doctoral dissertation, University of Sydney, Sydney, 2006..
[13] D. Janakiram, V. Adi Mallikarjuna Reddy, A.V.U. Phani Kumar, "Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks", ComSware, 2006, pp. 1-6.
[14] S.R. Jeffery, G. Alonso, M. J. Franklin, W. Hong, and J. Widom, "Declarative Support for Sensor Data Cleaning", International Conference on Pervasive Computing, 2006, pp. 83-100.
[15] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng. "Localized Outlying and Boundary Data Detection in Sensor Networks", IEEE Trans. Knowl. Data Eng., 2007, Vol. 19, No. 8, pp. 1145-1157.
[16] S.Preetha, V.Radha, "Enhanced Outlier Detection Method Using Association Rule Mining Technique", International Journal of Computer Applications, 2012, Vol. 42, No. 7, pp. 1-6.
[17] L.A. Bettencourt, A. Hagberg, and L. Larkey, "Separating the Wheat from the Chaff: Practical Anomaly Detection Schemes in Ecological Applications of Distributed Sensor Networks", in Proc. IEEE International Conference on Distributed Computing in Sensor Systems, 2007, pp. 223-239.
[18] Y. Zhang, N. Meratnia, and J.M. Paul Havinga, "Ensuring high sensor data quality through use of online outlier detection techniques". International Journal of Sensor Networks, 2010, Vol. 7, No. 3, pp. 141-151.
[19] Y. Zhang, N.A.S. Hamm, N. Meratnia, A. Stein, M. van de Voort, "Statistics-based outlier detection for wireless sensor networks", International Journal of Geographical Information Science, 2012, No.1, pp. 1–20.
[20] Y. Y. Li, L.E. Parker, "Classification with missing data in a wireless sensor network", in Proceedings of IEEE Southeastcon, 2008, pp. 533 - 538.
[21] C. Liu, K. Wu, J. Pei, "An Energy-Efficient Data Collection Framework for Wireless Sensor Networks by Exploiting Spatiotemporal Correlation", IEEE Transactions on Parallel and Distributed Systems. 2007, Vol. 18, No. 7, pp. 1010-1023.
[22] A. Petrosino, A. Staiano, "A Neuro-fuzzy Approach for Sensor Network Data Cleaning", in Proceedings of KES'07, 2007, pp. 140-147.
[23] F. Hermans, N. Dziengel, Jochen H. Schiller, "Quality Estimation based Data Fusion in Wireless Sensor Networks", in Porceedings of MASS'09, 2009, pp. 1068-1070.

First Author Guo-Ying WANG received the B.S. degree in computer science from Beijing JiaoTong University, Beijing, China, in 1999 and M.S. degree in computer science from Guangxi University, Nanning, China, in 2004. In the same year, He joined the faculty of the Computer Science Department, Information Engineering College, Zhejiang A&F University, where he is currently a lecturer. His research interests include computer networks, peer-to-peer networks, wireless ad-hoc and sensor networks, and mobile networks.

Second Author Shen-Ming QU received the B.Eng. degree from Hebei University and the M.S. degree from Henan University. Currently, he is a lecturer in Computing Center, Henan University. His research interests include Computer Network, information retrieval and computer vision.