International Journal of Software Engineering and Its Applications
Vol. 9, No. 1 (2015), pp. 27-34
http://dx.doi.org/10.14257/ijseia.2015.9.1.03

A Novel Way of Weighting in the Risk Factor Management

Donghyok Suh and Kunsoo Oh*

Department of Architectural Engineering, Namseoul University
{ absuh, [1]ohkunsoo} @nsu.ac.kr

Abstract

A method of distinguishing several events included in sensor data was proposed, when various events were sensed simultaneously to the data several sensors sensed and reported and individual event data were mixed in an environment in which a data stream was sequentially obtained. This study had each sensor distinguish each event through a fast analysis of the data sensed by each sensor in this condition. For this purpose, clustering was made with sensor data, and first, internal variance of event cluster classified by each sensor was calculated, and how this changed with the passage of time was checked. Next, the cluster of each sensor sensing the same event was compared. Through this process, the sensed event could be more clearly distinguished. This study suggested the measure to divide a small quantity of sensors when it senses the several events. This measure can be used as application for the small mobile vehicle or robot to sense the peripheral situation.

Keywords: *Weighting, Clustering, Time slot, Context inference*

1. Introduction

In a ubiquitous sensor network or the Internet of things environment, sensors sense and report meaningful events. Sometimes, various sensors sense and report a single event. In this case, duplication should be removed before it reaches the gateway. Sometimes, a single sensor senses and reports several events. Numerous situational factors act in the periphery of the sensor. In this case, it is preferentially necessary to classify and identify numerous events included in the value measured by the single sensor. In the meantime, sometimes, two or more sensors sense each of different numerous events. For example, a mobile vehicle has a limitation that many sensors cannot be mounted. It would be a better situation in which one sensor is mounted in a mobile vehicle by class, but be that as it may, often numerically enough sensors can be furnished. Various kinds of sensors can be used, and it can be installed variously though it is not diverse. In this case, a small number of no single sensor are installed, and it should consider the activity of other mobile vehicle which acts in the mobile vehicle peripheral and it should detect the peripheral situation. Thus, the study on division and identifying of the various events in the sensor data was reported by sensing when the sensor cannot sense with various events as purposes.

If so, when the various events are sensed in the several sensors, the measure to distinguish the several events including the sensor data should be suggested. This study can be conducted in the data stream environment. The sensors sense and report the event by the passage of time at a certain interval. In the condition that the sensor reports consistently, the measure to analyze rapidly is needed. The event included in the sensor data should be identified and analyzed, so the meaning should be classified. This study aims to contribute to the efficient context inference by distinguishing each event through a rapid analysis of this condition. To achieve this goal, clustering about sensor data should

* Corresponding Author

ISSN: 1738-9984 IJSEIA

be conducted, and the result values of clustering based on the sensor values should be compared and the measure to optimize is suggested. By through this process, the sensed event can be more clearly distinguished. By comparing the results of analysis for each time slot, the process of the situation can be inferred.

This study is composed as follows. In Chapter 2, the related researches were determined, and in Chapter 3, multiple sensors and the multiple events and division measure which were suggested in this study were described. In Chapter 4, the experiment and evaluation were conducted and the conclusion was shown in Chapter 1.

2. Related Works

The clustering can be defined as the process that ‘the point which has similar characteristic with the given data points in the multidimensional space should be clustered in the same cluster, and the point which has each different characteristic should be clustered in the other cluster [Lee, Seok-ryong, Lim, Dong-hyeok, Jeong, Jin-wan]. The clustering can be classified with five types largely, first, Partitioning Methods are corresponding to k-means, k-medoids, CLARANS techniques. Second, the Hierarchical Methods are corresponding to BIRCH, CURE, CHAMELEON. Third, the Density-based Methods are corresponding to DBSCAN, DENCLUE, OPTICS. Fourth, the grid-based Methods are corresponding to STRING, WaveCluster, CLIQUE. Fifth, the Model-Based Clustering Method is corresponding to EM(Expectation-Maximization) [3, 6, 7, 8].

CLARANS is based on the randomized search, and it is the clustering technique which was designed to increase efficiency by reducing the search space using two users input arguments. BIRCH is the multi-phase clustering technique to generate the hierarchy data structure which is called CF-tree by scanning the database, and any clustering can be used for clustering the terminal node of CF-tree. This technique is the first technique to effectively process the peripheral point in the database areas. DBSCAN tried to minimize the domain knowledge required for deciding the input arguments, and the cluster of any shapes can be generated by distributing the data point. The basic idea should include the determined minimum number of points in a given radius of each point in the cluster. Therefore, this technique needs only two input arguments(number of radii and minimal points)[3],[4],[5]. CLIQUE is the technique to identify the dense cluster automatically in the subspace of the given higher-order data space. In other words, there may be the cluster in the subspace though there is no cluster in the given space, and it must be the appropriate technique. As the input arguments, the global density limit value for the grid size and cluster dividing the space is needed. The ideas about the clustering in the part space are extended as the concept of the clustering, and it is the technique to select the specific parts dimensions closely associated with the data point and to find the cluster of this dimensions [1, 2].

CURE generates the cluster having the diverse size as not globular shape. This technique shows the multiple points to scatter and distribute each cluster. The cluster which has not the globular shape can represent its shape better by using the several points. This algorithm will be closed when the number of generated cluster reaches the given value with input arguments.

3. Weighting Method in the Risk Recognition

The sensors sense the information about attached object and peripheral situation. If there are more constant levels of change amount in the sensor mote, it can be transmitted to the host.

One sensor can sense the several each different events. If the elderly walking around construction site want to sense the upcoming motorcycle to him or her or the children

frolicking, and the three kinds of events are included in the sensing data by the sensors. But if the several events are sensed by 2 or more sensors, the measure to divide the several events included in the multiple sensor data should be required. The number of sensor that can be mounted are not extremely limited, but it should consider the environment which cannot furnish the mass of the sensor. This study suggested the measure to divide a small quantity of sensors when it senses the several events. This measure can be used as application for the small mobile vehicle or robot to sense the peripheral situation. The algorithm which is suggested by this study is as follows.

The thing that sequentially obtained data is put in the stream form can be assumed

$$DS =< D_1, D_2, D_3, \cdots, D_t >,$$

and the time to flow in can be assumed

$$TS =< T_1, T_2, T_3, \cdots, T_t >.$$

Where the $D_1, D_2, D_3, \cdots, D_t$ mean set of sensing values which were reported by the sensor mote.

According to the existing method, the element $D_i =< \alpha_1, \alpha_2, \alpha_3, \cdots, \alpha_n >$, in D_i, and n sensors were processed by using average of obtained value from the sensor.

This study suggested the measure to process the sensor n as the single parameter as follows. By considering the internal and external variance between α_k and α_k center which were generated in T_i about the obtained data values with the continuing data stream form, it can be processed as the single parameter by deciding the weighted value based on the accuracy and precision.

3.1. Definition of Internal Variance

When $\overline{\alpha_k} = 1/n(\alpha_k) \sum_{t \in T_i} \alpha_k, \ (\alpha_k \in D_i)$

internal variance $V(\alpha_k) = 1/n(\alpha_k) \sum_{t \in T_i} \left\| \overline{\alpha_k} - \alpha_{k_t} \right\|^2$.

The above internal variance value can be defined as measuring a type of sensor in time-slot internal. The internal variance means the variance of obtained data from the sensor. In the sensor peripheral environment, the elements such as the bad elements of impact and sensor parts, and the error of value sent from sensor are the main factors to increase the variance. Therefore, the variance value due to error is connected to the precision of sensor, and it should be used as the main factors for the weighted value.

3.2 Definition of External Variance

When , $\bar{\alpha} = 1/n(D_i) \sum_{\alpha_k \in D_i} \alpha_k$,

External variance $V(D_i) = 1/n(D_i) \sum_{\alpha_k \in D_i} \| \overline{\alpha_k} - \alpha_k \|^2$.

This study assumed data which was sensed by 2 or more sensors. It conducts the arithmetic operation to calculate the external variance from the center values of clusters which were formed by each sensor. The external variance value means the variance between the clusters which was formed by the obtained data values from each different sensor. In the condition to consider the internal variance, and each sensors can send each different value which has regular error in the same situation. The obtained each different value form each different cluster. But the obtained each different value is by sensing consequentially same situation and event. Therefore, considering the variance in the classified cluster by each different sensor, how the sensed clusters are far from each other

can be known. Being far away means classifying the same event with different cluster. So the external variance is related to the accuracy. In other words, it means that the accuracy will decrease as the external variance increased.

In the experiment of this study, the sound sensors are set 2, so above external variance value can be considered as the distance between the events. About the clustering of data which is reported by each sensor, the cluster means the event. So the above formula can be simplified by calculating the distance from average between the same events of each different sensor.

$$V(D_i) = (\|\bar{\alpha} - \alpha_1\|^2 + \|\bar{\alpha} - \alpha_2\|^2)/2$$

This formula can simplify 2 sensors, and it means the average of calculating distance from total average of cluster about each cluster of each sensor. This value can be considered as special case of $n(D_i)=2$ from the definition of the above external variance mathematically.

So it can define the weighted value as the following formula.

$$w_k = \frac{(\sum_{\alpha_k \in D_i} \mathrm{V}(\alpha_\mathrm{k})) - \mathrm{V}(\alpha_\mathrm{k})}{\sum_{\alpha_k \in D_i} \mathrm{V}(\alpha_\mathrm{k})} \times \frac{V(D_i) - \|\bar{\alpha} - \alpha_2\|^2}{V(D_i)}$$

The formula is described as follows:

The internal variance and external variance value are reflected in this weighted value, the maximum weighted value can be 1 when the external variance and internal variance are 0. As this value is closer to 1, the accuracy and precision of data will be increased. The weighted value shows the minimal value when the external variance or internal variance is same with the whole variance, and it is shown as 0. If one of the external variance and internal variance increase extremely, the reliability of corresponding data will decrease, so the weighted value can be low close to 0.

So as the accuracy and precision of data collected by the sensor are high, the weighted value will increase.

According to these results, the data which is processed by the single parameter is as follows.

$$d_i = \sum_{j=1}^{k} \left[(w_j \alpha_j) \Big/ \left(\sum_{w_k \in W} w_k \right) \right]$$

The value which was divided as the sum of the raw data after calculating the inner product of weighted value and single data can be applied to all data. This processing value has the following properties.

The sum of $\sum d_i$=1 (for all i), and all processing values is 1.

($0 \le d_i \le 1$) and all processing values have the value in the closed domain between 0 and 1.

It normalized high significance as about 1 and low significance as about 0 by multiplying the data and weighted value. If the decided weighted value above is 1, it can reflex the raw data, so the processing value can be 1. In other words, the very ideal case that there is no internal variance and external variance of all data can be only 1. So the sensor data which is close to ideal accuracy and precision is obtained, the processing value above can be close to 1. In addition, when the weighted value is 0, the raw data can be processed as 0, so the processing value also becomes 0. This situation occurs when the variance of all data increase extremely, and in this case, it can be judged that the value sent from the sensor is not significant.

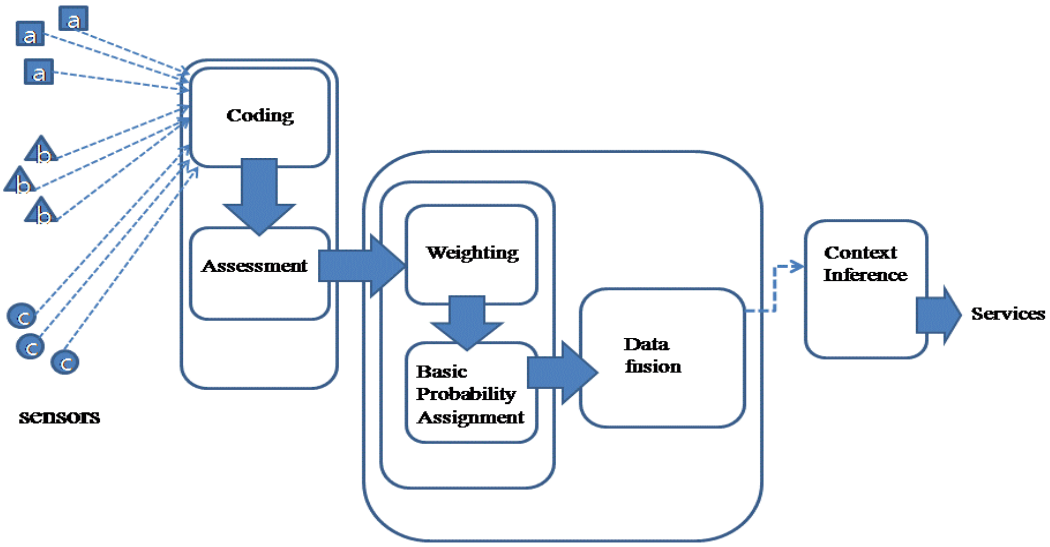

Figure 1. Context Inference and Weighting Method

4. An Experiment and Evaluation

We Details of the experiment: it sensed the sound data including three events through two sound sensors, and it is reported as the host. Each sensor sensed same three events and the information of three events are included in the data stream reported by sensor. The sensing time interval was 0.1 second. The time interval about data stream reported sequentially was 30 seconds. 300 raw data is included in each time interval, and 3 events are mixed with this raw data. The experiment conducted clustering based on this data, and the division measure suggested by this study was applied.

As a result of an experiment, the following Figure 1 is a graph of the entire raw data sound sensors 1 and 2 sensed reported.

Two sensors of A, B sense the sound, it was reported that can sense 10 times per 1 second. Three kinds of events were included in this sensing data. Thus, this study aims to divide the events mixed with the data lump.

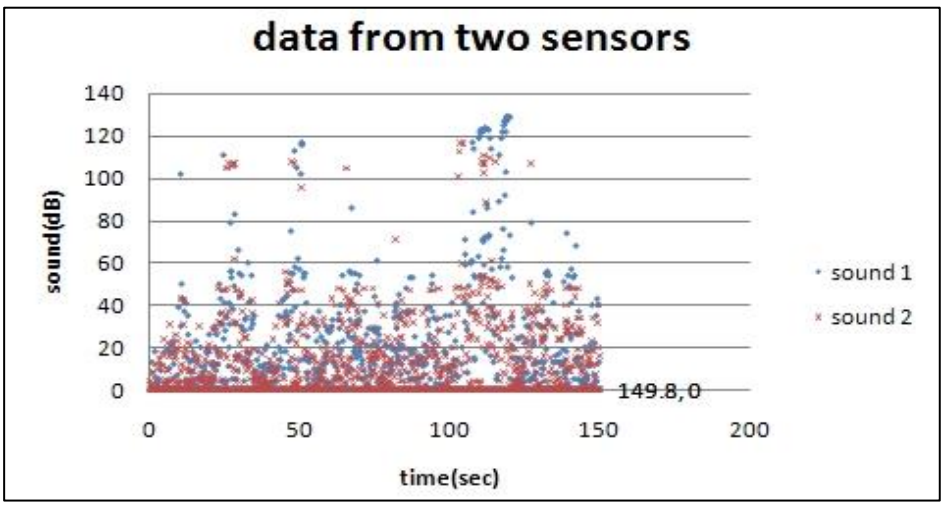

Figure 2. Data from Two Sensors, Sound Sensor 1 and 2

Figures 2 below shows the distribution of the obtained data by each time interval.

The sequentially obtained data were divided at a certain time interval. In this experiment, the data streams obtained per 30 seconds interval were divided.

The time slot 1 was divided from 0.1 second to 30 seconds and the time slot 2 was divided from 30.1 second to 60 seconds in Figure 2, so the time slot interval 5 was divided from 120.1 second to 150 seconds.

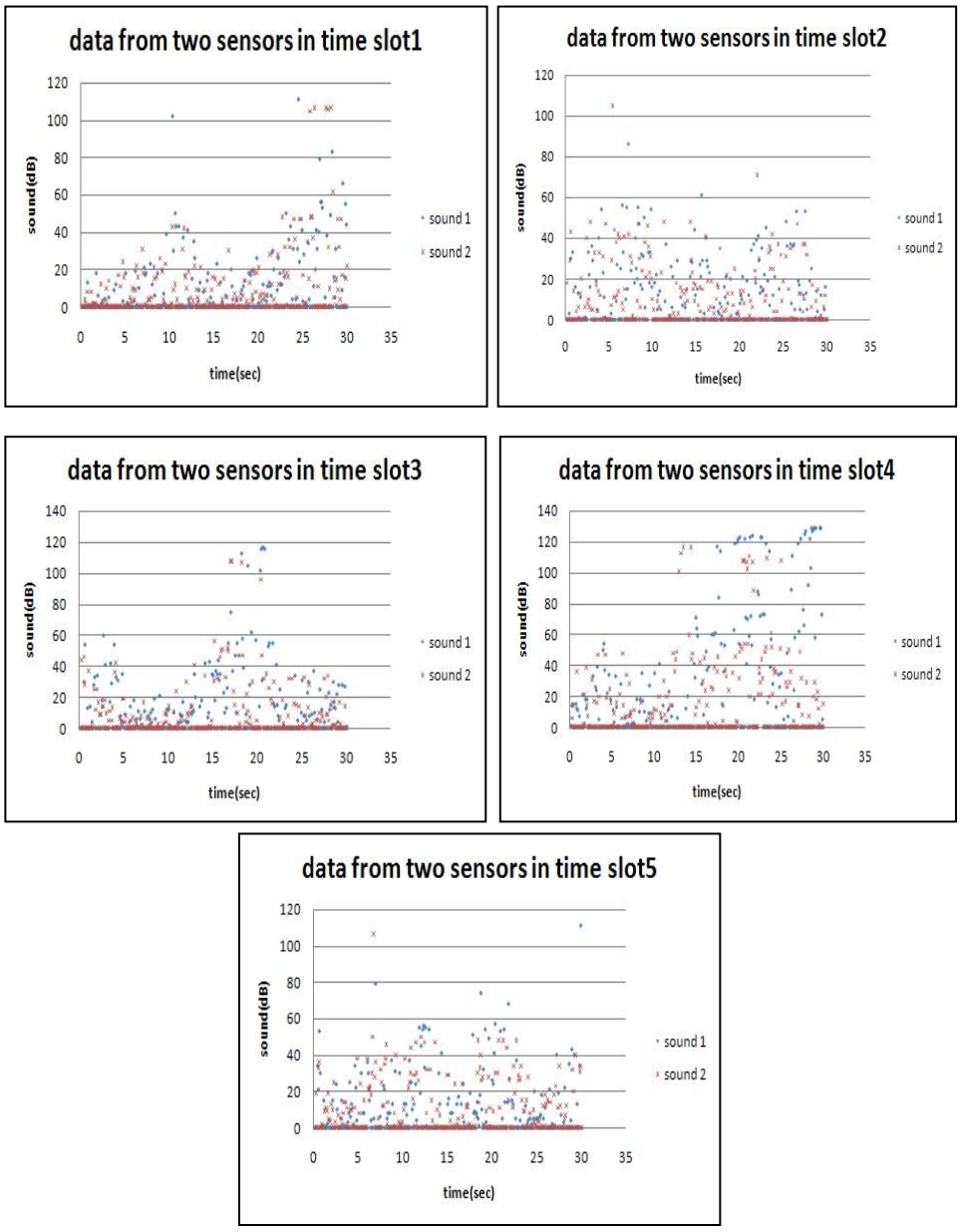

Figure 2. Sensing Data by Sensor in Each Time Interval

Thus, in the result of dividing the obtained data by the measure which was suggested by this study, it was divided as following Figure 3.

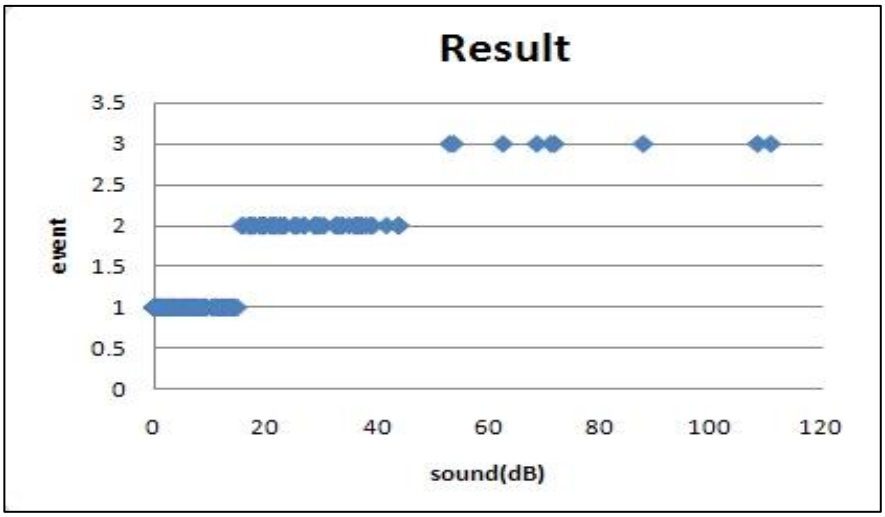

Figure 3. Result of Classification of Three Events Mixed in Sensors A and B

What was proven through the above experiment was that when several sensors included several events in the data stream, this could be classified into each event. Like this, classification by each event can help data fusion processing using this, and based on the event information included in the sensor, situations could be inferred, so it has enough value of use.

5. Conclusion

A method of distinguishing several events included in sensor data was proposed, when various events were sensed simultaneously to the data several sensors sensed and reported and individual event data were mixed in an environment in which a data stream was sequentially obtained. This study had each sensor distinguish each event through a fast analysis of the data sensed by each sensor in this condition. For this purpose, clustering was made with sensor data, and first, internal variance of event cluster classified by each sensor was calculated, and how this changed with the passage of time was checked. Next, the cluster of each sensor sensing the same event was compared. Through this process, the sensed event could be more clearly distinguished.

Acknowledgements

Funding for this paper was provided by Namseoul University.

References

[1] T. Kanungo, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, "An Efficient k-Means Clustering Algorithm, Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, **(2002)** July.
[2] D. H. Suh and S. S. Yoon, "Weighting Method Based on Event Frequency for Multi-sensor Data Fusion in Wireless Sensor Network for the People with Disability", Journal of Assistive Technology, vol. 5, no. 1, **(2011)**, pp. 37-47.
[3] S. Guha, R. Rastogi and K. Shim, "CURE: A Efficient Clustering Algorithm for Large Databases", Proceedings of A CMSIGMOD, **(1998)**, pp. 73-84.
[4] R. Ng and J. Han, "Efficient and Effective Clustering Method for Spatial Data Mining", Proceeding s of the 20th VLDB Conference, **(1994)**, pp. 144- 155.
[5] C. C. Aggarwal, C. Procopiuc, J. L. Wolf, P. S. Yu and J. S. Park, "Fast Algorithms for Projected Clustering", Proceedings of the A CMSIGMOD International Conference on Management of Data, **(1999)** June, pp. 61-72.

[6] T. Kanungo, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, "An Efficient k-Means Clustering Algorithm", Analysis and Implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, **(2002)** July.
[7] M. Dipti and T. Patel, "K-means based data stream clustering algorithm extended with no. of cluster estimation method", International Journal of Advance Engineering and Research Development (IJAERD), vol. 1, no. 6, **(2014)** June.
[8] M. Khalilian, N. Mustapha, N. Suliman and A. Mamat, "A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets", Proceedings of the International Multi-Conference of Engineers and Computer Scientists, vol. 1, **(2010)**.
[9] V. N. Hamed, Z. Kamran and N. Nasser, "Context-Aware Middleware Achitecture for Smart Home Environment", International Journal of Smart Home, vol. 7, no. 1, **(2013)** January, pp. 77-86.
[10] W. Mo and Z. Ding, "A Novel Template Weighted Match Degree Algorithm for Optical Character Recognition", International Journal of Smart Home, vol. 7, no. 3, **(2013)** May, pp. 261-270.
[11] H. Komari Alaei, S. Iman Pishbin and K. Salahshoor, "A New PCA Cluster-Based Granulated Algorithm Using Rough Set Theory for Process Monitoring", International Journal of Database Theory and Application, vol. 4, no. 4, **(2011)** December, pp. 1-12.
[12] P. Thi Bach Hue, S. Wohlgemuth and I. Echizen, Nguyen, "An Experimental Evaluation for a New Column – Level Access Control Mechanism for Electronic Health Record Systems", International Journal of u - and e - Service, Science and Technology, vol. 4, no. 4, **(2011)** December, pp. 1-14.
[13] S. Goyal and G. Kumar Goyal, "Radial Basis (Exact Fit) and Linear Layer (Design) ANN Models for Shelf Life Prediction of Processed Cheese", International Journal of u - and e - Service, Science and Technology, vol. 5, no. 1, **(2012)** March, pp. 63-70.
[14] Y. Gu, L. Sun and J. Wang, "Comparative Analysis of Single and Mixed Spatial Interpolation Methods for Variability Prediction of Temperature Prediction", International Journal of Hybrid Information Technology, vol. 6, no. 1, **(2013)** January, pp. 67-76.
[15] S. Ji, L. Huang and J. Wang, "A Distributed and Energy-efficient Clustering Method for Hierarchical Wireless Sensor Networks", International Journal of Future Generation Communication and Networking, vol. 6, no. 2, **(2013)** April, pp. 83-92.

Authors

Donghyok Suh, he is a professor at Namseoul University. He received the M.S. degrees in computer engineering from Hoseo University in 2005 and the Ph.D. in computer science from Chungbuk National University in 2012. His research interests included in stream data processing and data fusion in wireless sensor network.

Kun soo Oh, he is a professor at Namseoul University. He received the B.S. degree in Architecture from Hongik University in 1980, and the M.S. and Ph.D. in Architecture from Hongik University in 1994. His research interests are included in information technology architecture and data fusion in wireless sensor network.

