International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:10, 2017

Ontology-Based Backpropagation Neural Network Classification and Reasoning Strategy for NoSQL and SQL Databases

Hao-Hsiang Ku, Ching-Ho Chi

***Abstract*—**Big data applications have become an imperative for many fields. Many researchers have been devoted into increasing correct rates and reducing time complexities. Hence, the study designs and proposes an Ontology-based backpropagation neural network classification and reasoning strategy for NoSQL big data applications, which is called ON4NoSQL. ON4NoSQL is responsible for enhancing the performances of classifications in NoSQL and SQL databases to build up mass behavior models. Mass behavior models are made by MapReduce techniques and Hadoop distributed file system based on Hadoop service platform. The reference engine of ON4NoSQL is the ontology-based backpropagation neural network classification and reasoning strategy. Simulation results indicate that ON4NoSQL can efficiently achieve to construct a high performance environment for data storing, searching, and retrieving.

***Keywords*—**Hadoop, NoSQL, ontology, backpropagation neural network, and high distributed file system.

I. Introduction

OVER the past decade years, there is massive data growth in all areas. With the explosive growth of data, distributed databases are widely used in various applications, including e-commerce, social networking, recommendation system, location-based service, and so on [9]. Big data have the characteristics of large scale, many kinds, fast generation, high value but low density. Big data application is the use of data analysis methods, from the big data mining effective information, to provide users with auxiliary decision-making, to realize the process of large data value [4]. With the development of information technology and the increase of data, a well-designed classification and reasoning strategy for NoSQL and SQL database should consider problems of integrity, interoperability, adaptivity and modularization.

1) Integrity. Data are with different types and formats. Hence, different appliances need to be controlled and displayed after translating and normalizing. To tackle this problem, NoSQL is with key-value format to manipulate data, which can accommodate a large number of data format.
2) Interoperability. It is difficult for an application handling and controlling heterogeneous applications. To tackle this problem, a web-based platform can accommodate with heterogeneous data format using XML and HTML 5 standards.
3) Adaptivity. Most commercial products are not easily to increase some specific functions. However, a well-designed classification and reasoning strategy should observe and handle multi-events in the same time. Hence, many kinds of thresholds should be designed and defined for these events. To tackle this problem, it is an important impact factor to define multi-triggers, which can automatically adapt ontology and trigger multiple appliances for different domains.
4) Modularization. Different algorithms are combined into a service. It is usually a fix model. Users are different to set another mode to handle this service. To tackle this problem, an Ontology-lite is extracted from ON4NoSQL, which can record different preferences and can automatically set parameters for a new environment.

For the aforementioned four technical issues, this study designs and proposes ON4NoSQL, which can overcome problems of integrity, interoperability, adaptivity and modularization and can provide an efficient classification strategy in NoSQL and SQL databases.

The rest of this paper is organized as follows. Section II describes related works. Section III describes the working flow of ON4NoSQL. Section IV describes the simulations of O4NoSQL. Finally, Section V gives conclusion remarks.

II. Related Work

Recently years, researchers are devoted into increasing speed of data access. Regarding the aforementioned issue, these enhancement methods can be roughly divided into three classes, which are described as follows.

1) To classify data by specific characteristics: Classification is defined as the data which is classified by specific characteristics. Agarwa et al. proposed an ontology-based system that creates a knowledge graph of the extracted information to make it machine understandable [1]. Authors used a knowledge graph to generate a Bayesian network and reason out the status of users. Ali et al. proposed a merged ontology and support vector machine (SVM)-based information extraction and recommendation system [2]. It is highly productive when analyzing retrieved information, and provides accurate recommendations. Ramesh et al. aimed to incorporate semantics knowledge in all the phases of Web Usage

Hao-Hsiang Ku is with the Department of Computer Science and Information Engineering, Hwa Hsia University of Technology, No.111, Gongzhuan Rd., Zhonghe Dist., New Taipei City 235, Taiwan, R.O.C. (corresponding author, e-mail: kuhh@go.hwh.edu.tw).

Ching-Ho Chi is with the Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan (e-mail: s105065519@m105.nthu.edu.tw).

mining process. CloSpan, a state-of-the-art algorithm for Sequential Pattern mining, is applied over the Semantic space to generate frequent Sequential Patterns [8]. It can promise a significant improvement on the quality of the recommendations.

2) To classify data by unknown characteristics: Clustering is defined as the data which are classified by unknown characteristics. Kwon et al. proposed a clustering scheme that classifies human behavior into 11 different categories including active and inactive activities in daily life [6]. Authors show that using proposed scheme can be classified with 99.24% accuracy. Kravchenko et al. indicated that integrating knowledge from different sources may be based on the ontology requirements for the development of which will be a pre-formed sheet [7]. Seo et al. proposed a new co-occurrence factor algorithm to compute suitable clusters. It focusses on the classification of semantic words using a user's hashtag data and co-occurrence hashtag information [9]. Xu et al. proposed a hybrid index for multi-dimensional query in HBase which is based on behavior-based collective classification mechanism [13].

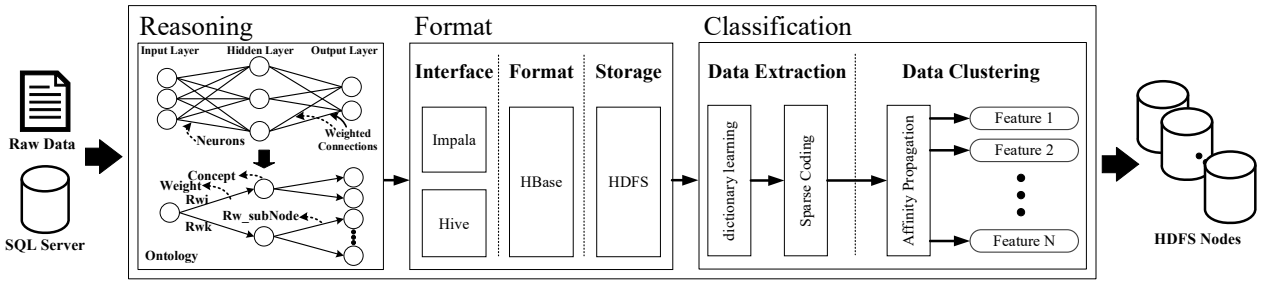

Fig. 1 The working flow of Ontology-based backpropagation neural network classification and reasoning strategy

3) To enhance existing access model: To enhance existing access model is defined as re-modify or re-set functions or parameters to achieve higher capability. A. Azqueta-Alzúaz et al. developed a tool for parallel massive data loading over HBase, the key-value data store of the Hadoop ecosystem [3]. It can enhance the massive data load based on key-value solutions based on range-partitioned data stores on distributed database systems over HBase. Guo et al. present the pre-partitioning and hash according to the data characteristics. The cluster is divided into several regions, then through Rowkey hash mapping, data are stored evenly to each partition [5]. Tang et al. proposed a hybrid index for multi-dimensional query in HBase, behavior-based collective classification method, to improve the classification performance in sparsely labeled networks [10]. It can classify data into suitable classes by hybrid index. Wei et al. proposed a method to improve multiple dimension data management performance in HBase, which is an optimized storage model and index scheme to provide efficient query over big multiple dimension data and multiple query patterns [12].

According to the aforementioned issues, it is difficult to classify data by unknown characteristics. Hence, this study designs and proposes an Ontology-based Backpropagation Neural Network Classification and Reasoning Strategy for NoSQL and SQL Database, ON4NoSQL. Furthermore, ON4NoSQL considers problems of integrity, interoperability, adaptivity and modularization.

III. Ontology-Based Backpropagation Neural Network Classification and Reasoning Strategy

This section describes the working flow of Ontology-based backpropagation neural network classification and reasoning strategy for NoSQL big data applications, which is called ON4NoSQL. ON4NoSQL is responsible for enhancing the performances of classifications, characteristics and evaluations in NoSQL and SQL databases to build up mass behavior models and ontology-lite, which is illustrated as Fig. 1.

Mass behavior models are constructed by MapReduce techniques and Hadoop distributed file system, which provides a new user to quickly initial his/her services by others' passed experiences. Data will deliver to multiple layer manipulation unit and classify these data into NoSQL database, Hbase. The core of Mass behavior model is sparse coding based machine learning mechanism. The detail working flow is described as follows.

A. Multiple Layer Manipulation Unit

Multiple layer manipulation unit is responsible for manipulating different data types, including event data, XML files, SQL-Lite database and SQL database, etc. It is not only with multiple formats but also with heterogeneous types. Hence, multiple layer manipulation unit is with three steps, which are data classification, data normalization and file store. The detail working flow of multiple layer manipulation unit is illustrated as Fig. 2.

When data run into data classification phase, data will be classified different data types. After classification phase, data will be managed by Hive and Impala. It is convenient to control data by query planner, query coordinator, and query exec engine. The detail integration of SQL database and NoSQL database is illustrated as Fig. 3. The manipulated data will be stored into Hbase. File store is based on Hadoop and Hbase, which is a Key-value-based NoSQL database. Finally, data will be divided into many blocks and stored into many DataNode.

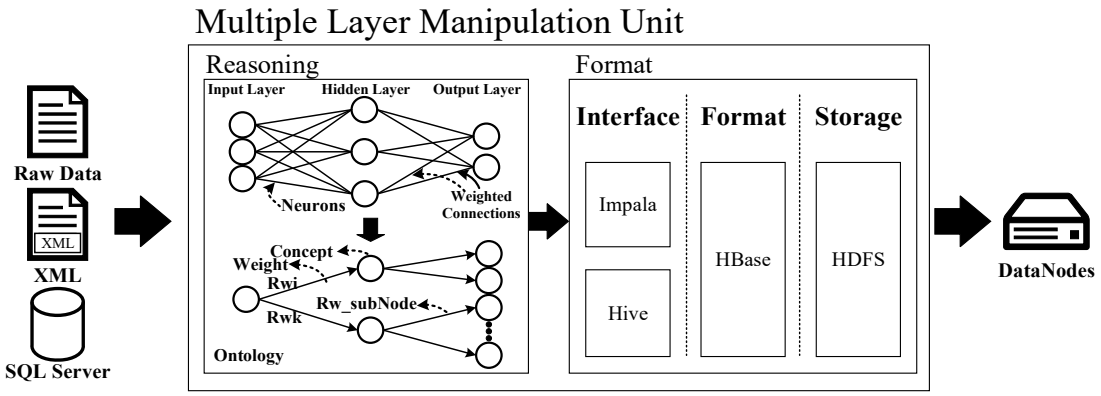

Fig. 2 Multiple layer manipulation unit

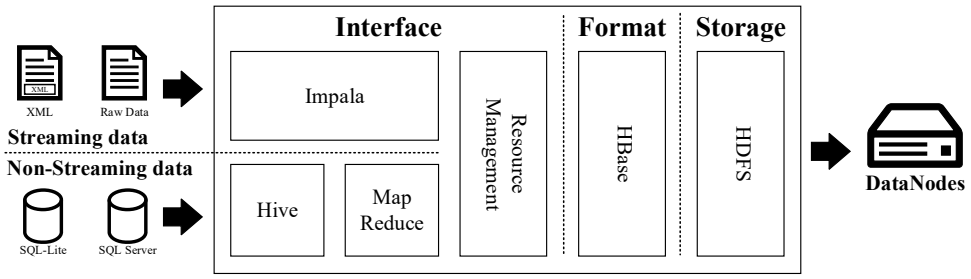

Fig. 3 Integration of SQL and NoSQL database

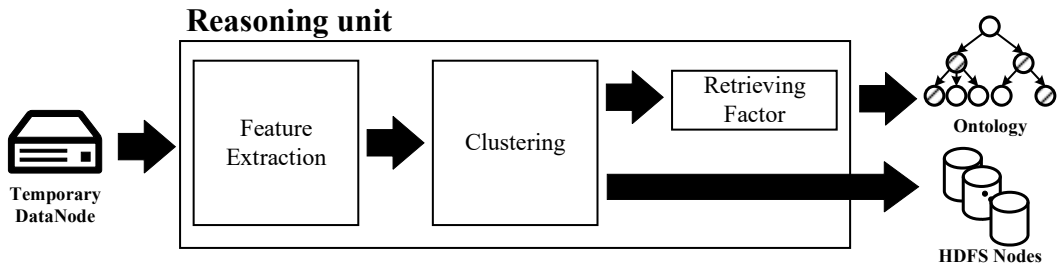

Fig. 4 Reasoning unit of ON4NoSQL

B. Reasoning Unit

Reasoning unit is responsible for reason out important factors. Hence, MapReduce can do parallel computing of the mass behavior model. Three steps are illustrated as Fig. 4, which are feature extraction, clustering, and retrieving factors.

1. Feature Extraction

The core of feature extraction is based on dictionary learning and sparse coding techniques. Dictionary learning is to learn a set of representative elements from the input data. The dictionary elements may be found by minimizing the average representation error, together with L1 regularization on the weights to enable sparsity, as (1) and (2)

$$\min_{D\in C,\alpha_i\in R^{K\times 1}} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\| X_i - D\alpha_i \|_2^2 + \lambda\|\alpha_i\|_1 \quad (1)$$

$$C \triangleq \left\{D \in R^{M\times K} s.t. \left\|d_j\right\|_2^2 = 1, \forall j = 1, \dots\dots, K\right\} \quad (2)$$

Xi is a one-dimensional matrix, λ is a weight, which is greater than or equal to 0. α_i is the sparse code of X_i, D is dictionary, which includes K columns, and d_j is j column vector.

In sparse coding phase, the trained dictionaries are computed by Least Absolute Shrinkage and Selection Operator (LASSO), which reason out the sparse code α_i of X_i. The sparse vector is computed as (3).

$$\min_{\alpha_i\in R^{K\times 1}} \|X_i - D\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \quad (3)$$

After feature extraction phase, many feature parameters are produced, which can increase the accuracy of data clustering.

2. Clustering

The core of clustering phase is based on Affinity Propagation

(AP). Affinity Propagation is a clustering algorithm based on the concept of message passing between data points. AP selects the exemplar, the central point of a cluster, to build up the similarities, as (4)

$$s(i,j)=-||X_i-X_j||^2 \quad (4)$$

AP is with Responsibility and Availability to check the reliability. AP accumulates evidence "responsibility" R(i,k) from data point i for how well-suited point k is to serve as the exemplar for point i, and accumulates evidence "availability" A(i,k) from candidate exemplar point k for how appropriate it would be for point i to choose point k as its exemplar, as (5)-(7) [11].

$$r(i,j) = s(i,j) - \max_{k\neq j}(a(k,i) + s(i,k)) \quad (5)$$

$$a(j,i) = \min(0, \sum_{k\neq j} \max\{0, r(k,j)\} + \sum_{k\neq j,i} \max\{0, r(k,j)\}) \quad (6)$$

$$c_i^* = \arg\max_j r(i,j) + a(j,i) \quad (7)$$

3. Retrieving Factor

After clustering phase, retrieving factor phase is to retrieve, select and construct factors by relationship weights. The construct strategy is illustrated as Figs. 5 and 6. Fig. 5 illustrates the selecting strategy of important factors by backpropagation neural network and ontology. Fig. 6 illustrates the adjustment strategy of ontology-lite. The Rwi and Rwk are the related weight between nodes i and k, respectively. The Rwp and Rwc are the related weight between nodes p and c, respectively. The detail formula is shown as (8)-(10):

$$Rwi_{input-hidden} = \sum_{j=1}^{r} w_{ij} \quad (8)$$

$$Rwk_{input-hidden} = \sum_{j=1}^{r} w_{kj} \quad (9)$$

$$Rw_{subNode} = \frac{Rw}{\text{sum}(the\ parent's\ subNode)} \quad (10)$$

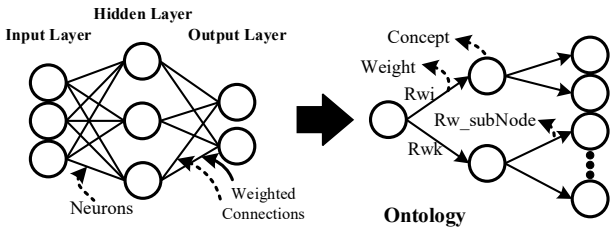

Fig. 5 Retrieve, select and construct factors by backpropagation neural network and ontology

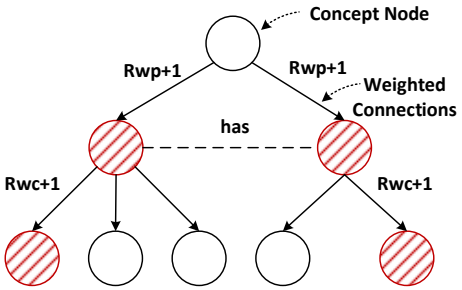

Fig. 6 Adjustment strategy of ontology-lite

TABLE I
ON4NOSQL BASED ON MYSQL DATABASE

MySQL		10K	20K	30K	40K	50K	100K	300K	500K	1000K
Original	Throughput(ops/sec)	1024.49	859.25	808.56	817.08	902.45	958.24	657.25	564.9	557.69
	Avg. Latency(us)	940.08	1136.27	1215.54	1205.53	1093.11	1034.3	1514.85	1763.9	1788.32
ON4NoSQL	Throughput(ops/sec)	1100.26	953.49	1012.73	1102.32	1203.23	1282.32	1023.38	1002.03	1019.83
	Avg. Latency(us)	878.91	929.31	1009.73	982.23	873.39	869.68	982.93	991.03	987.38

TABLE II
ON4NOSQL BASED ON HBASE DATABASE

Hbase		10K	20K	30K	40K	50K	100K	300K	500K	1000K
Original	Throughput(ops/sec)	661.68	873.21	974.88	1274.17	1365.22	1247.16	1128.09	1214.95	1180.8
	Avg. Latency(us)	1399.73	1082.15	983.62	752.42	784.53	706.85	877.47	816.83	841.64
ON4NoSQL	Throughput(ops/sec)	782.03	953.43	1102.93	1428.39	1637.48	1934.19	1988.61	1969.23	2003.42
	Avg. Latency(us)	1281.39	1045.32	923.94	693.27	572.49	463.42	409.17	429.42	398.21

IV. SIMULATION

This section presents the performance between relation database and big data database. The testing results are listed as Tables I and II, and illustrated as Figs. 7 and 8. This simulation results indicate that the ON4NoSQL can efficiently increase throughputs and reduce the latency.

V. CONCLUSION

The study designs and proposes an Ontology-based backpropagation neural network classification and reasoning strategy for NoSQL big data applications. ON4NoSQL is responsible for enhancing the performances of classifications in NoSQL and SQL databases to build up mass behavior models. ON4NoSQL overcomes problems of integrity, interoperability, adaptivity and modularization.

1) Integrity. ON4NoSQL combines SQL and NoSQL database, which use with key-value format to manipulate data, which can accommodate a large number of data format.

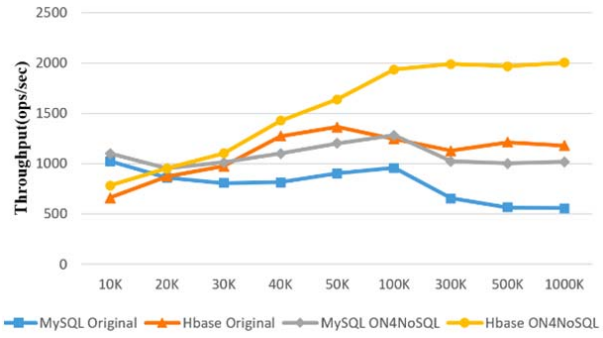

Fig. 7 Throughput of ON4NoSQL

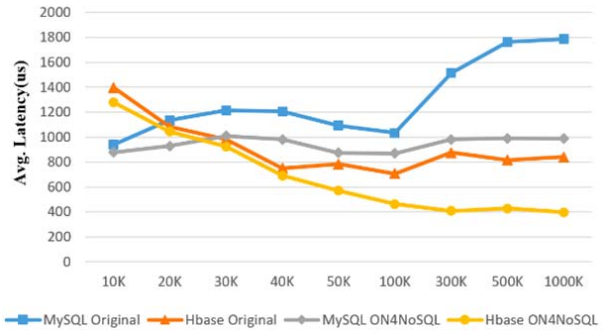

Fig. 8 Average latency of ON4NoSQL

2) Interoperability. ON4NoSQL is a web-based platform which can accommodate with heterogeneous data format using XML and HTML 5 standards.
3) Adaptivity. ON4NoSQL can define and set up multi-triggers, which can automatically adapt ontology and trigger multiple appliances for different domains.
4) Modularization. ON4NoSQL can record different preferences and can automatically set parameters for a new environment.

ACKNOWLEDGMENT

The Ministry of Science and Technology of the R.O.C. under the grant MOST 105-2221-E-146-009 supports this research.

REFERENCES

[1] P. Agarwa, R. Verma, A. Mallik, "Ontology based disease diagnosis system with probabilistic inference", India International Conference on Information Processing (IICIP), pages: 1 – 5, 2016.
[2] F. Ali, D. Kwak, P. Khan, S. H. A. Ei-Sappagh, S. M. R. Islam, D. Park, K. Kwak, "Merged Ontology and SVM-Based Information Extraction and Recommendation System for Social Robots", IEEE Access, Vol.5, pages: 12364 – 12379, 2017.
[3] A. Azqueta-Alzúaz, M. Patiño-Martinez, I. Brondino, R. Jimenez-Peris, "Massive Data Load on Distributed Database Systems over HBase", 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), pages: 776 - 779, 2017.
[4] X.X. Dou, X.P. Wang, Application of Big Data Analysis Method in Supply Chain, Advances in Networks, Vol. 4, No. 1, pp. 1-5, 2016.
[5] J. Guo, X. Wu, "Research on optimization of community mass data storage based on HBase", Third International Conference on Cyberspace Technology (CCT 2015), pages: 1 – 4, 2015.
[6] M. Kwon, M. Ju, S. Choi, "Classification of various daily behaviors using deep learning and smart watch", 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), pages: 735 – 740, 2017.
[7] Y. Kravchenko, E. Kuliev, I. Kursitys, "Information's semantic search, classification, structuring and integration objectives in the knowledge management context problems", 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), pages: 1 – 5, 2017.
[8] C. Ramesh, K. V. C. Rao, A. Govardhan, "Ontology based web usage mining model", 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), pages: 356 – 362, 2017.
[9] S. Seo, J. Kim, L. Choi, "Semantic hashtag relation classification using co-occurrence word information", 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), pages: 860 – 862, 2017.
[10] X. Tang, B. Han, H. Chen, "A hybrid index for multi-dimensional query in HBase", 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS), pages: 332 - 336, 2016.
[11] K. Wang, J. Zhang, D. Li, X. Zhang, T. Guo, "Adaptive Affinity Propagation Clustering", Acta Automatica Sinica, Vol 33, No. 22, pages:1242-1246, 2007
[12] Z. Wei, J. M. Qu, L. Liu, C. Q. Zhu, W. J. Yin, "MDDM: A Method to Improve Multiple Dimension Data Management Performance in HBase", 2015 IEEE 17th International Conference on High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICESS), pages: 102 - 109, 2015.
[13] J. Xu, L. LI, X. Lu, S. Hu, B. G, W. Xiao, L. Yao, "Behavior-Based Collective Classification in Sparsely Labeled Networks", IEEE Access, Vol.5, pages: 12512 - 12525, 2017.

Hao-Hsiang Ku received the B.S. degree from the Department of Management Information Systems, Chung-Hua University, Hsinchu City, Taiwan, in June 2001, the M.S. degree from the Department of Management Information Systems, National Pingtung University of Science and Technology, Pingtung City, Taiwan, in June 2003, and the Ph.D. degree from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan City, Taiwan, in June 2009. He is currently an Associate Professor of Computer Science and Information Engineering, Hwa Hsia University of Technology, New Taipei City, Taiwan. His research interests include medical information systems, wireless sensor networks, and embedded multimedia applications.

Ching-Ho Chi received the B.S. degree from the Department of Information and Telecommunications Engineering, Ming Chuan University, Taoyuan, Taiwan, in June 2016. He is currently a master student of Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan. His research interests include Linux Techniques, wireless sensor networks, and embedded multimedia applications.