Sensors and Materials, Vol. 31, No. 6 (2019) 1789–1801
MYU Tokyo

S & M 1899

Dynamic Data Driven-based Automatic Clustering and Semantic Annotation for Internet of Things Sensor Data

Szu-Yin Lin,[1*] Jun-Bin Li,[2] and Ching-Tzu Yu[3]

[1]Department of Computer Science and Information Engineering, National Ilan University, Yilan County, Taiwan
[2]Department of Statistics and Information Science, Fu Jen Catholic University, New Taipei City, Taiwan
[3]Institute of Information Management, National Chiao Tung University, Hsinchu City, Taiwan

(Received February 14, 2019; accepted April 1, 2019)

Keywords: clustering, semantic annotation, ontology, Internet of Things, sensor data

Faced with the advent of the era of smart Internet of Things (IoT), a large amount of sensor data and a large number of intelligent applications have been introduced into our lives. However, the dynamic and multimodal nature of data makes it challenging to transform them into machine-readable and machine-interpretable forms. In this study, a semantic annotation method is proposed to annotate sensor data through semantics. First, the method constructs an initial ontology based on the semantic sensor network (SSN) ontology for dynamic IoT sensor data. Second, through K-means clustering, new knowledge is extracted from input data, and the semantic information is used for updating the initial ontology. The updated ontology then forms the basis of semantic annotation. In this study, an experiment is performed to analyze the data collected from sensors every 10 s for a period of one month. From the results of simulation experiments, we found useful knowledge from new data. With more available knowledge, sensor data can be annotated with higher adequacy.

1. Introduction

The academic community and industry have become increasingly concerned about the Internet of Things (IoT) lately. IoT, which originated from the Electronic Product Code system proposed by MIT, is a global Internet infrastructure that connects physical and virtual objects through information extraction and communication capabilities. It symbolizes a technological revolution that enables the existing Internet to be interconnected with physical objects and devices (i.e., "things"), and virtual representations owing to technological advancements, including identification and contactless data exchange, distributed sensor networks, short-range wireless communications, and universal mobile accessibility. The current initiative on building IoT demands application and service platforms that can capture, communicate, store, access, and share data from the physical world. This will create new opportunities in a long list of

*Corresponding author: e-mail: szuyin@niu.edu.tw
https://doi.org/10.18494/SAM.2019.2333

ISSN 0914-4935

domains such as e-health, retail, green energy, manufacturing, smart cities, smart home, and also personalized end-user applications.[1] IoT also allows "people and things to be connected anytime, anyplace with anything and anyone ideally using any path, any network, and any service"[2] in a 6A connectivity paradigm.

Combining the context of IoT with semantic technologies, we can build integrated semantic systems to support semantic interoperability. The semantic technologies have been used in recent years as key solutions to provide formalized representations of real-world data.[3–6] It has also been widely applied to service-oriented technologies to provide interoperable interfaces, processes, and service descriptions. Reference 7 provides an overview of the recent deployments of applications of semantic technologies in various aspects of IoT. It emphasizes that the use of semantic technologies should take the dynamicity and constraints of the IoT domain into consideration. The semantic Web technologies include well-defined standard and description frameworks (e.g., RDF, OWL, and SPARQL) and a variety of open-source and commercial tools for creating, managing, querying, and accessing semantic data. However, these still do not eliminate the key roles of information analytics and intelligent methods, which can process and interpret the data and create meaningful abstractions. Semantic annotation is a process of attaching additional information to various concepts in a given text or any other content. Compared with traditional text annotations, semantic annotations are used by machines. There are many tools available.[8] It has been proven to be a useful method to describe related sources of information in order to improve data aggregation and data filtering.[2] Semantic annotation can support more effective mechanisms to be designed to create context-aware applications and to integrate IoT data. Through this technology, real world resources and sensor data can be connected to the existing semantic Web.[9]

Semantic technologies facilitate data conceptualization and abstract representation. The conceptualized data can be more easily interpreted by machines and thus can be interlinked with necessary resources on the Web. However, IoT sensing data are not the same as traditional general data. The main difference between IoT sensing data and general data is the way they are generated. IoT sensing data are a collection of outputs of physical devices that perceive and detect some type of input from the physical environment. It could be a collection of large amounts of real-time, continuous, and distributed data. The sensor data related to different events and occurrences can be analyzed and summarized into knowledge, which helps in decision or action making. If device interconnection and data integration/processing can be achieved, the idea of context awareness enables applications, machines, and human users to better understand their surroundings. Moreover, the data collected by different sensors and devices are usually multimodal and diverse in nature. Owing to the huge amount of sensor data, it is difficult to determine their relationships. For the reasons mentioned above, we will discuss how to structure, annotate, and share sensor data, and how to make sense and transform them into actionable knowledge and intelligence, so that machines can determine their class and property much more easily, which is the goal pursued in this study.

The semantic sensor network (SSN) ontology is built by the W3C Semantic Sensor Network Incubator group. This group shows one of the most significant efforts in the development of an information model for sensory data and provides a high-level schema to describe sensor devices,

their operation and management, observation and measurement data, and process-related attributes of sensors. The SSN ontology is a domain-independent ontology that describes sensors and observations for use in a sensor network and sensor web applications, by merging sensor-, observation-, and system-focused views.[10] Previous studies mainly use combinations of the SSN ontology with other domain ontologies to explore the relationships among sensor data, and such a method is used to drive semantic annotation.[11–14] We think that the method can be improved. The method in related research usually builds an ontology and uses the ontology to classify sensor data. Once the collected sensor data do not belong to any attribute of the ontology, classification will become difficult. In this study, we will introduce a bottom-up method to build the ontology and revise the top-down structure related to the research studies mentioned. In our proposed study, a data mining method is employed to analyze all sensor data to determine their relationships, and the retrieved relationships are used to update the attribute and field of the ontology mentioned in related research. The relationships among data are used to drive the data ontology and considered the foundation of semantic annotation.

2. Semantic Annotation Method for Dynamic IoT Sensor Data

In this study, we present a data-driven semantic annotation method for IoT sensor data. The main concept of this method is to construct an ontology based on the given sensor data. In the literature, most methods construct the ontology directly, and then semantic annotation is performed on the basis of the ontology. However, if the ontology is not well defined beforehand, difficulties may arise in semantic annotation. Therefore, if the ontology can be built with the data collected by the sensor itself, we can improve the precision of semantic annotation. According to the reason mentioned above, we propose a bottom-up method to build an ontology in this study and use this ontology for semantic annotation. In this section, we introduce the flow of the proposed method (see Fig. 1).

- Data Preprocessing: In the preliminary step, an ontology is built on the basis of the SSN ontology.

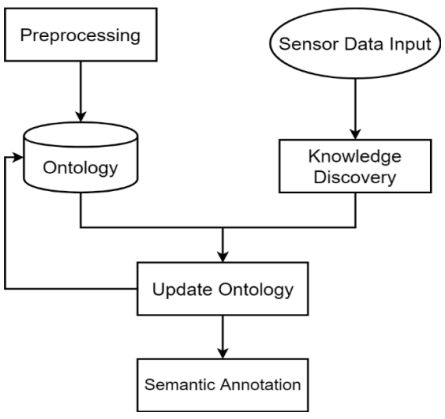


Fig. 1. Flowchart of the proposed method.

- Knowledge Extraction: In this step, a data mining method is adopted to cluster historical and real-time sensor data in order to extract knowledge.
- Ontology Update: According to the knowledge extracted from the last step, the original ontology is modified to a new ontology.
- Semantic Annotation: After the previous steps are completed, semantic annotation of the sensor data is performed.

The phases of the proposed method are described in detail below (see Fig. 2).

2.1 Phase 1

At this phase, an initial ontology is firstly built as the basis for the following steps. In this study, we extend the model in Ref. 15 and use the SSN and DUL ontologies as our basic ontologies. The SSN ontology has been used with some domain ontologies to develop various smart thing ontologies, such as the smart product ontology.[7] The SSN can also be used along with other ontologies, such as SWEET Ontology, Quantity Kinds, and Units Ontology.

- The DUL (DOLCE-UltraLite) ontology provides a set of upper-level concepts that can be the basis for easier interoperability among many middle- and low-level ontologies. It was chosen

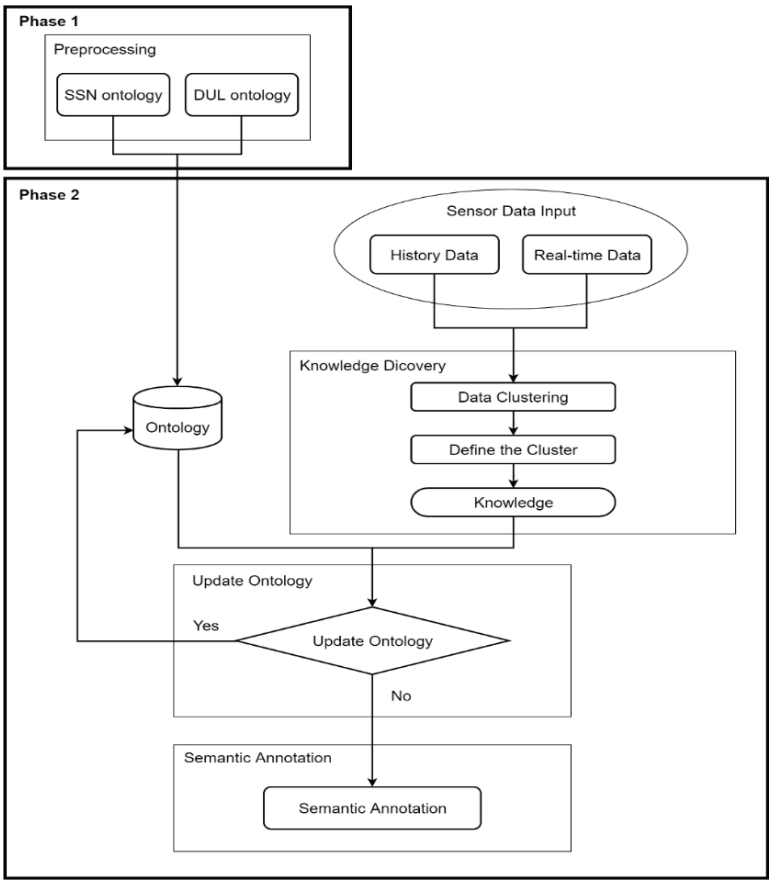


Fig. 2. Two-phase automatic clustering and semantic annotation method.

as the upper ontology because it is more lightweight and has an ontological framework and basis, for example, with qualities, regions, and object categories that are consistent with the modeling. Table 1 shows the contents of DUL ontology.

- The full SSN ontology consists of 41 concepts and 39 object properties, and directly inherits from 11 DUL concepts and 14 DUL object properties. In this research, we choose part of the SSN ontology instead of the full SSN ontology for a higher efficiency. We also choose the parts related to the sensor data we collected (see Table 2).

2.2 Phase 2

In this phase, a data mining method is employed to extract knowledge from historical and real-time sensor data. There are many alternative data mining techniques in the literature, and some of them are used within the same architecture.[16] In this section, K-means clustering is chosen as the method for knowledge discovery. The following paragraph describes the flow of this step and introduces the chosen clustering method (see Fig. 3).

- Knowledge Discovery: In this phase, knowledge discovery describes the process of automatically searching large volumes of data for patterns. It can be considered as the extraction of knowledge from the data. This discovery process is developed out of the data mining domain and is very closely related to terms in both methodology and terminology. The challenge in extracting knowledge from the data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, and optimization to

Table 1
Contents of DUL ontology.

Section	Module	Classes	Properties
DUL	DUL	DesignedArtifact, Event, InformationObject, Method, Object, PhysicalObject, Processes, Quality, Region, Situation	describes, hasLocation, hasPart, hasParticipant, hasQuality, hasRegion, includesEvent, includesObject, isDescribedBy, isLocationOf, isObjectIncludedIn, isParticipantIn, isQualityOf, isRegionFor, isSettingFor, satisfies

Table 2
Contents of SSN ontology used in the research.

Section	Module	Classes	Properties
Skeleton	Skeleton	FeatureOfInterest, Observation, Property, Sensing, Sensor, SensorInput, SensorOutput, Stimulus	detects, featureOfInterest, forProperty, hasProperty, implementedBy, implements, isPropertyOf, isProxyFor, observationResult, observedBy, observedProperty, ofFeature, sensingMethodUsed
Model	Process	Input, Output, Process	hasInput, hasOutput, isProducedBy
Sensor	Measuring	SensingDevice, SensorDataSheet	observes
Observation	Observation		madeObservation, observationResultTime, observationSamplingTime, qualityOfObservation
Base	Data	ObservationValue	hasValue

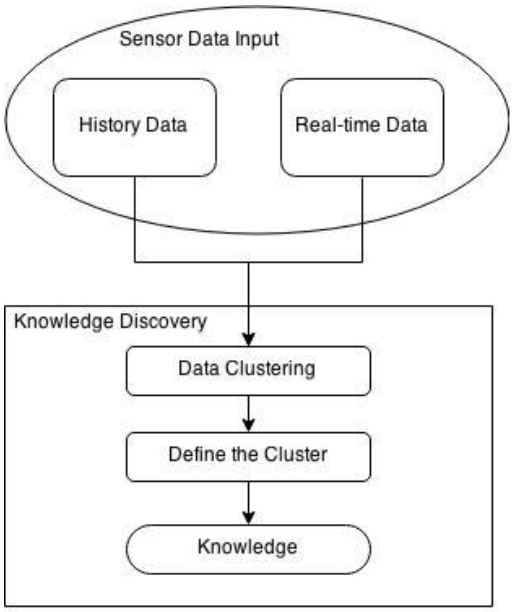


Fig. 3. Knowledge discovery.

deliver advanced business intelligence and web discovery solutions.[17] In our proposed method, clustering analysis has been applied to the process of knowledge discovery.

- Data Clustering: Clustering analysis is the task of grouping a set of objects in the same cluster that are more similar to each other than to those in other clusters. Cluster analysis is not an automatic task, but an iterative process of knowledge discovery. In this research, we take K-means as the clustering methodology. K-means clustering aims to partition n observations into k clusters, in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Figure 4 shows the flow of K-means clustering. The K-means clustering method has the following steps:[18]
  0. Given k.
  1. Partition objects into k nonempty subsets.
  2. Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., mean point, of the cluster).
  3. Partition a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized (where c_i is the centroid of cluster C_i).
  4. Assign each object to the cluster with the nearest seed point.
  5. Go back to step 2 and stop when the assignment does not change.
- Define cluster: After the clustering step, we can recognize that there are several clusters without any knowledge. We have to define the cluster by human activity. The methodology to define knowledge is that by setting thresholds for different attributes for data after clustering, we can define the clusters based on the thresholds we defined. After defining the clusters, the clusters become a whole new knowledge. Owing to the volatile nature of sensor data, presently this step can be only performed manually by domain experts.

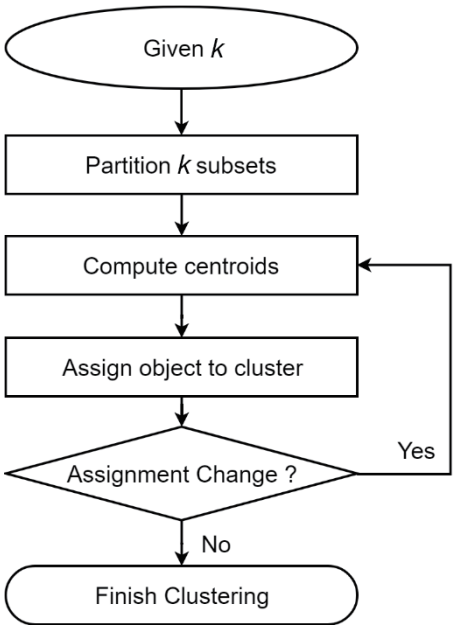


Fig. 4. K-means clustering.

- Updating Ontology: After defining the knowledge, we have to build the knowledge back into our base ontology. The knowledge should be defined as the class for the ontology, because it is a new concept for the ontology. For the second time or later, we use the ontology updated from the last time as the base ontology and perform the steps above to update the ontology.
- Semantic Annotation: In this step, we use the ontology built by the above steps to annotate the semantics in the raw data. There are several tools for semantic annotation.

The advantage of the semantic annotation method is that we can define the sensor data with more knowledge than by other methods. In related references, a new ontology is built first without modifying the original ontology. Once the amount of data becomes huge, the interpretation of the data will become difficult. However, in this study, the proposed mechanism is semiautomatic, because after clustering the data, there are just clusters without knowledge. The knowledge of the clusters is defined by people. Once there is a knowledge base for the data, maybe we can transform the semiautomatic method into an automatic approach. We think that this can be considered in the future.

3. Simulation and Results

3.1 Simulation environment

Protégé is an ontology editing and knowledge management system developed by Stanford University and is an open source software. Owing to its excellent design and numerous plug-ins, Protégé has become one of the most widely used ontology editors. It supports users in building both simple and complex ontology-based applications. In addition to Protégé, RStudio

is used in this study to perform the data mining tasks for its powerful and productive user interface of RStudio IDE. Moreover, handy R packages help us to draw diagrams to show the results of clustering. Information Abstraction Toolkit is also used for data preprocessing.

3.2 Case definition

The data set we used was the sensor data obtained from a test bed from the Communication Systems Research Centre at the University of Surrey. This data set was collected from sensor nodes employed in an office every 10 s during a one-month period, bringing the total number of data entries to 274960. The data include temperature (see Fig. 5), light level, and noise level. We choose 400 samples for each dataset by averaging the original dataset.

After the initial base ontology is built, the relationship between classes and properties is shown in Fig. 6. It is generated by the Protégé plug-in named Ontograf.

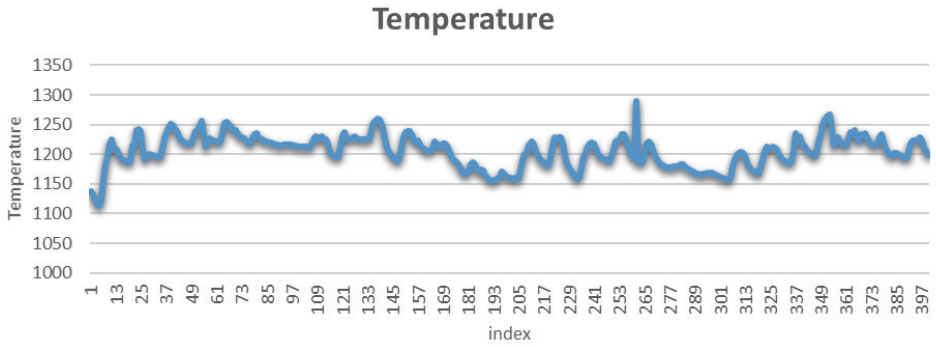


Fig. 5. (Color online) Temperature case.

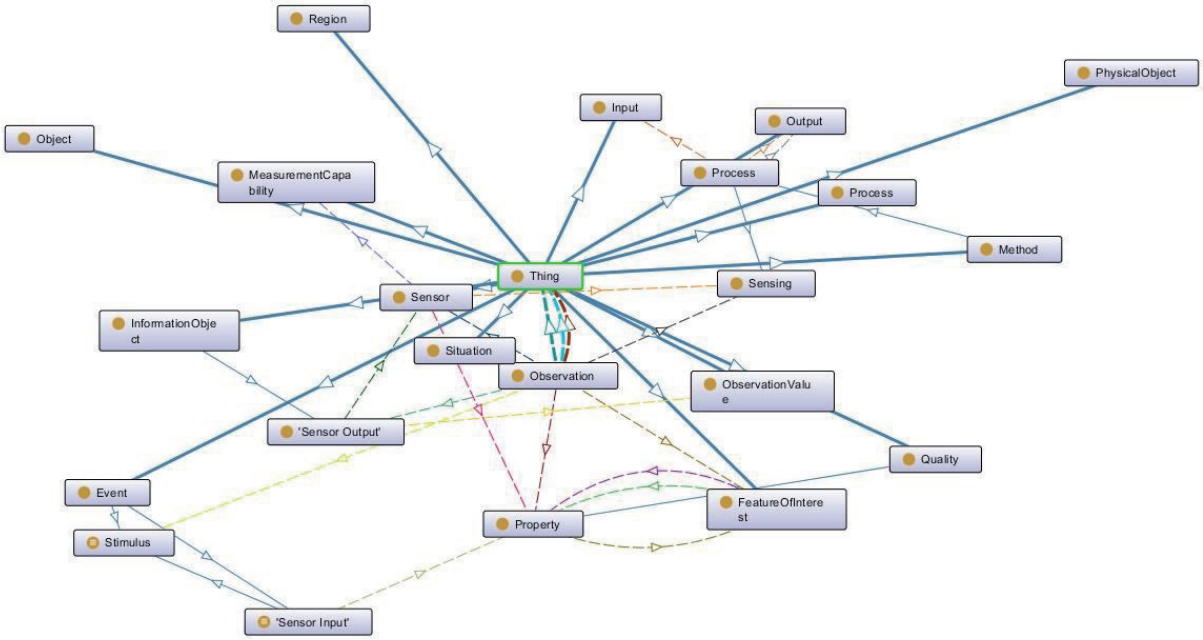


Fig. 6. (Color online) Base ontology.

3.3 Simulation results

In this case, we take the temperature sensor as the target. Our data set is divided into four separate weeks to distinguish the variance of the data due to time. The descriptions below show individual clustering results, concepts, and attributes of data for each week.

- Week 1: Figure 7 shows the clustering result for the temperature data of week 1. We separate the data into four clusters, and the four clusters with concepts and attributes are defined in Table 3.
- Week 2: Figure 8 shows the clustering result for the temperature data of week 2. We separate the data into four clusters, and the four clusters with concepts and attributes are defined in Table 4.
- Week 3: Figure 9 shows the clustering result for the temperature data of week 3. We separate the data into four clusters, and the four clusters with concepts and attributes are defined in Table 5.

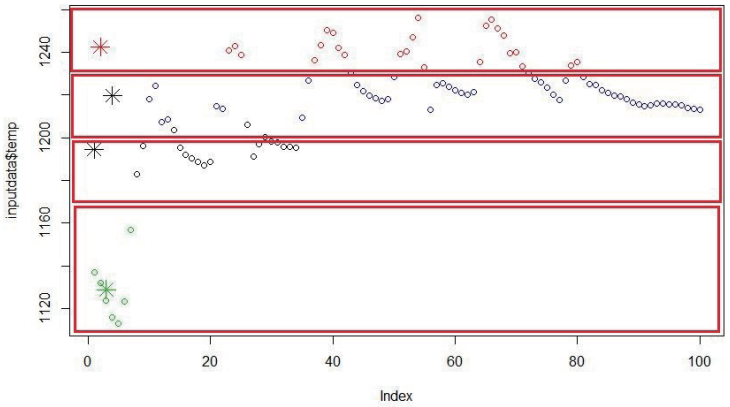


Fig. 7. (Color online) Clustering result for week 1.

Table 3
Concepts and attributes for week 1.

Concepts	Attributes
Cold	temp < 1170
Cool	1170 ≤ temp < 1207
Warm	1207 ≤ temp < 1231
Hot	temp ≥ 1231

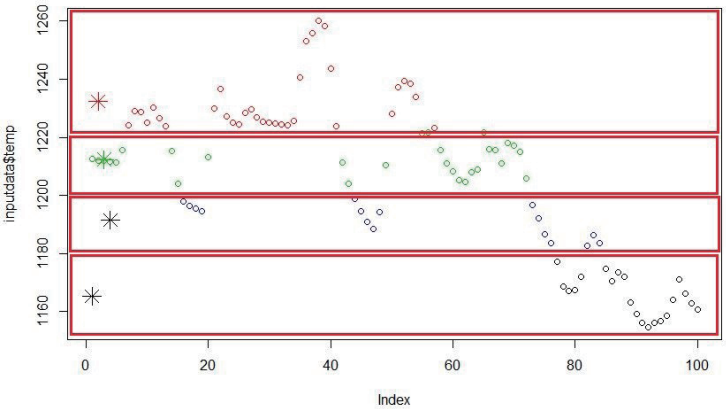


Fig. 8. (Color online) Clustering result for week 2.

Table 4
Concepts and attributes for week 2.

Concepts	Attributes
Cold	temp < 1185
Cool	1185 ≤ temp < 1213
Warm	1213 ≤ temp < 1235
Hot	temp ≥ 1235

- Week 4: Figure 10 shows the clustering result for the temperature data of week 4. We separate the data into four clusters, and the four clusters with concepts and attributes are defined in Table 6.

On the basis of the above results, our base ontology can be updated with these four newly retrieved concepts, as shown in Fig. 11. Compared with the initial base ontology in Fig. 6, a new branch called 'Temperature' is built, and four new nodes 'Cool', 'Cold', 'Hot', and 'Warm' are generated, by the automatic clustering results in the updated ontology of Fig. 11.

The results show that in the cases we implemented, we can cluster the data set in a simple manner. For this reason, we can easily discover knowledge from the data set and update the knowledge to our ontology. This method can be utilized in many application areas, such as smart home or smart environment. On the basis of our results, the quantity of knowledge will vary according to the number of clusters partitioned. Thus, the new ontology is expected to be different.

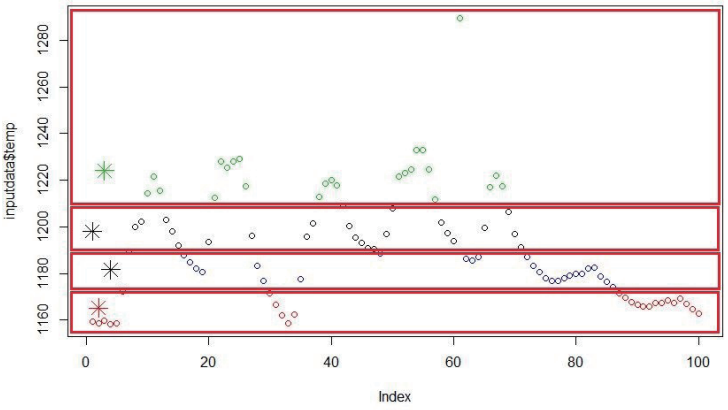


Fig. 9. (Color online) Clustering result for week 3.

Table 5
Concepts and attributes for week 3.

Concepts	Attributes
Cold	temp < 1173
Cool	1173 ≤ temp < 1190
Warm	1190 ≤ temp < 1210
Hot	temp ≥ 1210

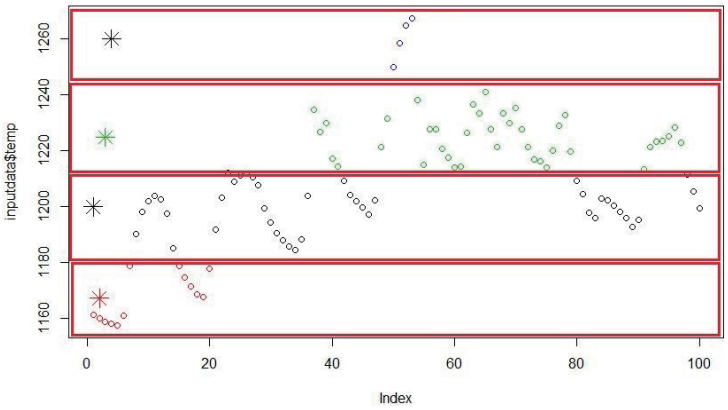


Fig. 10. (Color online) Clustering result for week 4.

Table 6
Concepts and attributes for week 4.

Concepts	Attributes
Cold	temp < 1181
Cool	1181 ≤ temp < 1213
Warm	1213 ≤ temp < 1245
Hot	temp ≥ 1245

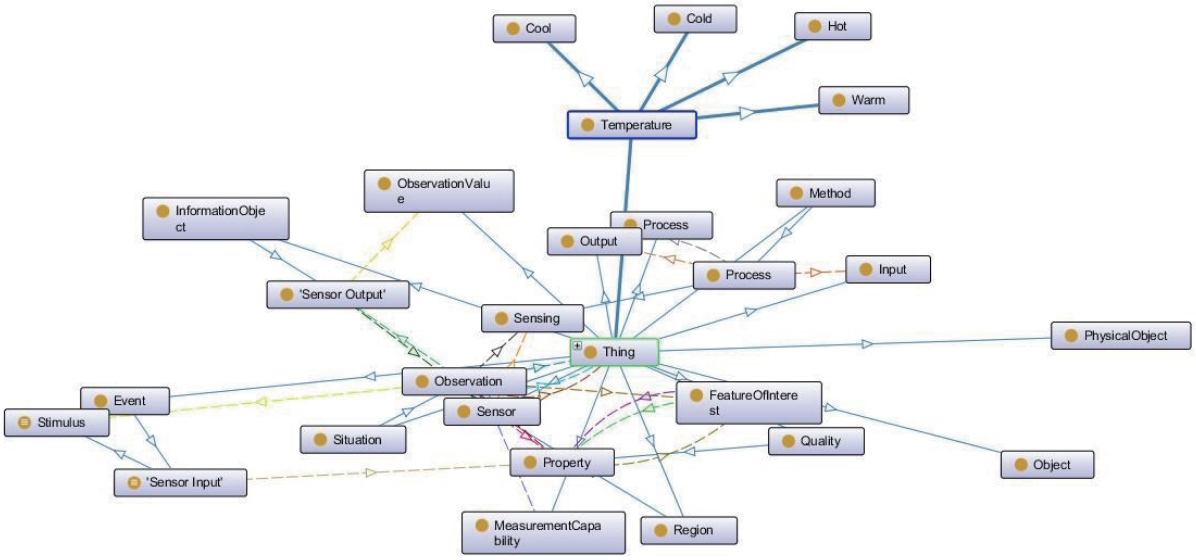


Fig. 11. (Color online) Updated ontology.

4. Conclusions

The amount of sensor data is getting larger these days. Determining how to annotate the data and make them machine-interpretable is an important issue. Ontology construction and semantic annotation are used to resolve this issue. In our work, an initial ontology based on the SSN and DUL ontologies is firstly built. Then, historical sensor data are used to extract knowledge. We adopt the K-means clustering method to obtain and group the data set. After the data are clustered, we define the groups with new knowledge. This newly extracted knowledge modifies our initial base ontology into an updated new ontology. Finally, we use the latest and most updated ontology to perform semantic annotation. From the results presented in Sect. 3, we can determine whether the result for the data set we collected is distinct. However, once the result of clustering is not obvious, it is difficult to define the knowledge of clusters, such that building a new ontology will be difficult.

The following are some of the issues worth paying attention to for further discussion: the automation of knowledge definition and the ontology generation without an initial ontology. The phase of knowledge definition in our proposed method is performed manually. To make the phases automatic, if the relationships among the data have been found first, the rule base can be built. After clustering the data, it may be easier to give the clusters some knowledge. The proposed method has built a base ontology first. However, how to build an ontology from the beginning without a base can be discussed in the future. Building a data ontology from the data collected instead of providing an ontology can much better approximate the real situation, such that the accuracy of the semantic annotation of concern can be improved.

Acknowledgments

This research work was supported by the Ministry of Science and Technology, Taiwan under the grant 107-2410-H-197-004.

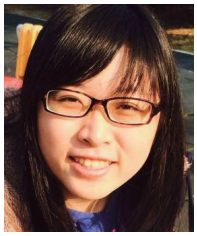
References

1 E. Borgia: Comp. Commun. **54** (2014) 1. https://doi.org/10.1016/j.comcom.2014.09.008
2 B. Manate, V. I. Munteanu, and T. F. Fortis: 2014 Eighth Int. Conf. Complex, Intelligent and Software Intensive Systems (CISIS, 2014) 582. https://doi.org/10.1109/CISIS.2014.84
3 M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. Le Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor: J. Web Semant. **17** (2012) 25. https://doi.org/10.1016/j.websem.2012.05.003
4 M. Compton, C. A. Henson, L. Lefort, H. Neuhaus, and A. P. Sheth: Proc. 2009 2nd Int. Workshop on Semantic Sensor Networks **522** (2009) 17.
5 K. Janowicz and M. Compton: Proc. 2010 3rd Int. Conf. Semantic Sensor Networks **668** (2010) 64.
6 E. Maleki, F. Belkadi, B. J. v. d. Zwaag, and A. Bernard: IFAC-PapersOnLine **50** (2017) 13059. https://doi.org/10.1016/j.ifacol.2017.08.2005
7 P. Barnaghi, W. Wang, C. Henson, and K. Taylor: Int. J. Semant. Web Inf. Syst. **8** (2012) 1. https://doi.org/10.4018/jswis.2012010101
8 P. Oliveira and J. Rocha: 2013 IEEE Symp. Computational Intelligence and Data Mining (CIDM, 2013) 301. https://doi.org/10.1109/CIDM.2013.6597251
9 W. Wei and P. Barnaghi: Semantic Annotation and Reasoning for Sensor Data: Smart Sensing and Context (Springer, Berlin, 2009) p. 66.
10 K. Kotis and A. J. S. W. J. Katasonov: Int. J. Distributed Systems and Technologies **4** (2013) 47. https://doi.org/10.4018/jdst.2013070104
11 J. Jeong, T. S. Yoon, and J. B. Park: Expert Syst. Appl. **105** (2018) 1. https://doi.org/10.1016/j.eswa.2018.03.051
12 M. Rida, A. Makhoul, H. Harb, D. Laiymani, and M. Barhamgi: Ad Hoc Networks **84** (2019) 158. https://doi.org/10.1016/j.adhoc.2018.09.012
13 C. Angsuchotmetee, R. Chbeir, and Y. Cardinale: Future Generation Comput. Syst. (2018). https://doi.org/10.1016/j.future.2018.01.044
14 B. Yang and M. Mareboyana: J. Network Comput. Appl. **35** (2012) 577. https://doi.org/10.1016/j.jnca.2011.05.008
15 C.-T. Yu, Y.-H. Zou, H.-Y. Li, and S.-Y. Lin: 2018 1st Int. Cognitive Cities Conf. (IC3, 2018) 188. https://doi.org/10.1109/IC3.2018.00-30
16 H. Banaee, M. Ahmed, and A. Loutfi: Sensors **13** (2013) 17472. https://doi.org/10.3390/s131217472
17 U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth: AI Mag. **17** (1996) 37. https://doi.org/10.1609/aimag.v17i3.1230
18 J. Han, J. Pei, and M. Kamber: Data Mining: Concepts and Techniques (Elsevier, MA, USA, 2011) p. 740.

About the Authors

Szu-Yin Lin received his M.S. and Ph.D. degrees in information management from National Chiao-Tung University, Taiwan, in 2012. He was a visiting scholar at Coventry University, UK, in 2012. He was an assistant professor from 2013 to 2017 and an associate professor from 2017 to 2019 in the Department of Information Management, Chung Yuan Christian University, Taiwan. Since 2019, he has been an associate professor in the Department of Computer Science and Information Engineering, National Ilan University, Taiwan. His research interests include the areas of intelligent data analysis, applied artificial intelligence, deep learning, management information systems, service-oriented computing, and multiagent distributed computing. (szuyin@niu.edu.tw)

Jung-Bin Li received his M.Sc. and Ph.D. degrees in information management from London School of Economics and Political Science, UK in 1996, and National Chiao-Tung University, Taiwan in 2006, respectively. He is currently an assistant professor at the Department of Statistics and Information Science in Fu Jen Catholic University in Taiwan. His research interests include the areas of data classification, big data analytics, and artificial intelligence in financial engineering. (071635@mail.fju.edu.tw)



Ching-Tzu Yu received her B.S. and M.S. degrees in computer science and information management from National Chiao-Tung University, Taiwan, in 2013 and 2015, respectively. Since 2015, she has been an engineer at Zyxel Communications Corp. (handsomeme0706@gmail.com)