論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	A study on abnormal-node detection and root cause analysis for dependable IoT sensor networks
著者(和文)	ベルジャブ ニスリーン
Author(English)	Nesrine Berjab
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第12298号, 授与年月日:2022年12月31日, 学位の種別:課程博士, 審査員:横田 治夫,宮崎 純,DEFAGO XAVIER,渡部 卓雄,吉瀬 謙二
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第12298号, Conferred date:2022/12/31, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

Doctoral Program

論 文 要 約

THESIS OUTLINE

系・コース： Department of, Graduate major in	情報工学 系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of （ 学術 ）
学生氏名： Student's Name	Nesrine Berjab	指導教員（主）： Academic Supervisor(main)	横田 治夫

Thesis Outline

Since the emergence of the Fourth Industrial Revolution, there has been a growing trend in the use of elements of the Internet of Things (IoT). As a result, IoT environments such as smart homes, smart factories, and smart cities are becoming increasingly popular and have permeated various areas of our lives. In this context, the use of sensors on IoT devices ensures a seamless connection between the devices and the physical world. Indeed, modern IoT devices are computer-like devices that come with a wide range of heterogeneous sensors connected through a dynamic and distributed wireless sensor network (WSN). In this context, the primary requirements of such devices are that they monitor their environmental conditions, report sensor data, and perform appropriate actions in response to the surrounding circumstances.

However, the accuracy of such decision-making depends upon the reliability and trustworthiness of the collected sensor data. Unfortunately, these devices are resource constrained and susceptible to many abnormalities, leading to unreliability and sometimes to a significant failure of parts of the entire system. Indeed, the stream data acquired by heterogeneous IoT sensors are seldom perfect. Most collected data streams include missing or abnormal values caused by failure, malfunction, or environmental events. Furthermore, IoT security issues have also been gaining increasing attention in recent years due to rising threats and attacks against its devices. Specifically, false data injection attacks (FDIAs) aim to tamper with the sensed data reflecting the normal system operation state. Such unreliable data affect real-time monitoring and compromise the quality of data analysis. Therefore, reliable methods for recovering the missing data, detecting the abnormal ones, and analyzing the root cause are critical to improving the reliability of IoT.

This doctoral dissertation thus presents solutions to these challenges to ensure a dependable IoT sensor network. First, it presents methods to recover missing sensor data and detect abnormal nodes jointly rather than independently. Then, it proposes a sensor state trust model to assess the sensor's trustworthiness. Finally, to analyze the root cause of abnormalities, this thesis proposes a semantic-based domain ontology to integrate the heterogeneous data and retrieve information from the constructed knowledge graph. The dissertation considers the observed spatiotemporal (ST) and multivariate-attribute (MVA) correlations of heterogeneous sensor data to achieve higher estimation accuracy and detection performance. With the four proposed methods, this thesis presents a set of solutions to overcome the highlighted challenges and improve reliability in the IoT.

First, the dissertation introduces FuzHD++, a new distributed method to recover missing sensor data and detect abnormal nodes jointly rather than independently. Both elements, data recovery and abnormal node detection, rely on the observed temporal correlation of sensor data to effectively achieve reliable recovery estimation and detection performance. In the data recovery process, the system adopts a matrix profile to extract the top-k repeated patterns from different sensor nodes. Furthermore, it utilizes the k-nearest neighbor estimator to recover the missing data based on the extracted pattern information of multiple neighbor nodes. During the abnormal node detection process, the system adopts a refined fuzzy rule-based detection method. The refined fuzzy rule-based inference system integrates the expert rules and the rules obtained from sensor data analysis to treat ambiguity in the decision-making process. I validate the performance of FuzHD++ by comparing it with existing methods using two real-world datasets. The results show that the proposed missing sensor data recovery method TkRP improved the root mean square error (RMSE) by more than 0.25 compared to the most accurate existing methods. Furthermore, FuzHD++ shows an improvement of 0.15 for detecting abnormal sensor readings by achieving an accuracy of at least 0.89 and 0.92 with Intel Lab data and Yokota Lab data, respectively.

Then, the dissertation proposes Multi-criteria Xcorr, a new centralized cross-correlation method based on spatial correlation, to conduct a more in-depth analysis and extract the sensor relationships. The cross-correlation is extracted in both space and time by conducting shape-based logical subclustering and two-phase analysis methods. In the first analysis phase, the system uses a variable-size sliding window and a median absolute deviation (MAD) measure. If the collected sensor data streams output a certain percentage of anomalous points, the MAD will flag these points as anomalous measurements, and all the sensor data and the window size will be passed to the second analysis phase. The latter performs both tumbling-window and sliding-window analyses to extract multicriteria cross-correlation measures. Finally, all the extracted sensor time-series features will be fed to the shape-based clustering to generate a sensor similarity-like graph. The latter reflects the similarity degree of the sensor with the other nodes. The nodes with a low similarity degree below the threshold will be identified as abnormal nodes. The experiments demonstrate that the proposed method achieves an accuracy of at least 0.90 and 0.93 with Intel Lab data and Yokota Lab data, respectively.

The dissertation also proposes T-LHDM (a trust model based on low and high decision-making processes), a sensor state trust model based on a Markov chain to represent sensor state transitions in normal operating conditions. To demonstrate the model, I first observed sensed data in a hierarchical decision-making process by intentionally exploiting the sensor data's temporal and spatial correlations. Then, the consistency between the precomputed set of sensor state transitions and the observed ones was used to detect corruptions and assess the sensor's trustworthiness. The experiments using two real-world datasets demonstrated that the proposed model achieves an accuracy of at least 0.95 with Intel Lab data. I also stress the significance of

considering the sensor's ST correlations to assess the trust of the sensed data, which has been neglected in previous studies. To this end, I conducted a comparative evaluation against a contemporary temporal correlation-based method and three well-known correlation tests. The results showed that T-LHDM outperformed all the competing methods.

Finally, the dissertation also proposes A2S2, an Anomaly Analysis on Semantic Sensor ontology, designed to model and describe the collected sensor data and its associated multivariate-attributes heterogeneous information. The objective is to integrate the heterogeneous sensor information and background knowledge as an interlinked RDF Knowledge Graph. The idea behind capturing both is that analyzing the heterogeneous sensor data and its associated information at a particular period and location will lead to insights into understanding the underlying root cause of an observed anomaly. The approach is solely based on semantics techniques such as SPARQL queries. Starting from the extended SOSA/SSN description, I first derive the physical process model from capturing the correlations among the heterogeneous sensors. This allows me to derive diagnosis rules and, finally, to obtain the identification results. The experimental results show that the identification accuracy exceeds 0.94 in the fire scenario, 0.83 in the FDIAs scenario, and 0.79 in the malfunction errors scenario.

With the four proposed methods, this thesis presents a set of solutions to overcome the highlighted challenges and improve reliability in the IoT. The raw sensor data were enhanced with spatial, temporal, and multivariate attributes correlations to enable high-level semantic reasoning. The works in this dissertation contribute directly to designing a dependable IoT sensor network that supports reliable methods for recovering the missing data, detecting anomalies, and analyzing the root cause. Specifically, the proposed methods are applicable to handle different anomalies, including FDIAs, which have been neglected in previous studies.