Smart Innovation, Systems and Technologies 218

Victor Chang
Muthu Ramachandran
Víctor Méndez Muñoz *Editors*



Modern Industrial IoT, Big Data and Supply Chain

Proceedings of the IIoTBDSC 2020

KES International

Springer

Smart Innovation, Systems and Technologies

Volume 218

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago, DBLP.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/8767

Victor Chang · Muthu Ramachandran ·
Víctor Méndez Muñoz
Editors

Modern Industrial IoT, Big Data and Supply Chain

Proceedings of the IIoTBDSC 2020

Springer

Editors
Victor Chang
Artificial Intelligence and Information Systems Research Group, School of Computing, Engineering and Digital Technologies, Teesside University
Middlesbrough, UK

Víctor Méndez Muñoz
Faculty of Computer Science, Multimedia and Telecommunications
Universitat Oberta de Catalunya (UOC)
Barcelona, Spain

Muthu Ramachandran
School of Built Environment, Engineering, and Computing Software Engineering Technologies, and Emerging Practices (SETEP)
Leeds Beckett University
Leeds, UK

ISSN 2190-3018 ISSN 2190-3026 (electronic)
Smart Innovation, Systems and Technologies
ISBN 978-981-33-6140-9 ISBN 978-981-33-6141-6 (eBook)
https://doi.org/10.1007/978-981-33-6141-6

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

We are pleased to publish this proceedings of the First International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC). The novelty in the scope is attracting contributions to some crucial aspects of our interconnected world. People and devices are defining high-tech services with open problems to face in reliability, efficiency, speed and accuracy.

We have received 120 submissions. After a peer-review process and paper improvement, 46 final papers were selected.

The conference proceeding scope is to demonstrate that today's challenges can be faced with crossover knowledge fields. Industrial Internet of Things (IIoT) allows the network of IoT devices to communicate, analyze data and process information collaboratively. Big Data (BD) infrastructure problem domains require to deal with the 5V's: volume, velocity, variety and veracity. Supply chain requires modern advancement and support from IIoT and big data to make all services efficient, fast, accurate and reliable. These three can combine and work together to produce more significant impacts and contributions as follows.

The maturity of IoT technologies can grow and become part of our everyday lives. BD 5Vs is a must to provide affordable BD analytics using machine learning, AI, statistical and other advanced techniques, models and methods, which can create values for people and organizations adopting it. Suppliers can know the updates on their stocks and demands. Manufacturers and transport companies know the workloads, destination and resource distributions. Owners and customers are also stakeholders receiving benefits from these approaches. At the end of the day, the supply chain can reach a greater sustainable ecosystem with the help of IIoT and big data.

We are grateful for the opportunities to serve the community. We will organize IIoTBDSC 2021 again and will publicize it in due course.

We wish you a happy reading of our selected papers.

Yours sincerely,

Middlesbrough, UK — Prof. Victor Chang
Leeds, UK — Dr. Muthu Ramachandran
Barcelona, Spain — Prof. Víctor Méndez Muñoz

November 2020

Contents

About the Editors

Prof. Victor Chang is currently a Full Professor of Data Science and Information Systems at the School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK, since September 2019. He is a Research Group Leader for Artificial Intelligence and Information Systems Research Group at Teesside University. He was Senior Associate Professor, Director of Ph.D. (June 2016–May 2018) and Director of MRes (September 2017–February 2019) at International Business School Suzhou (IBSS), Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China, between June 2016 and August 2019. He was also a very active and contributing key member at Research Institute of Big Data Analytics (RIBDA), XJTLU. He was Honorary Associate Professor at the University of Liverpool. Previously, he was Senior Lecturer at Leeds Beckett University, UK, between September 2012 and May 2016. Within 4 years, he completed Ph.D. (CS, Southampton) and PGCert (Higher Education, Fellow, Greenwich) while working for several projects at the same time. Before becoming an academic, he has achieved 97% on average in 27 IT certifications. He won a European Award on Cloud Migration in 2011, IEEE Outstanding Service Award in 2015, best papers in 2012, 2015 and 2018, the 2016 European special award and Outstanding Young Scientist 2017. He is a visiting scholar/Ph.D. examiner at several universities, Editor-in-Chief of IJOCI and OJBD journals, Former Editor of FGCS, Associate Editor of TII and Information Fusion, Founding Chair of two international workshops and Founding Conference Chair of IoTBDS and COMPLEXIS since the year 2016. He is Founding Conference Chair for FEMIB since the year 2019. He published 3 books as sole authors and Editor of 2 books on cloud computing and related technologies. He gave 22 keynotes at international conferences. He is widely regarded as one of the most active and influential young scientists and experts in IoT/Data Science/Cloud/Security/AI/IS, as he has experiences to develop 10 different services for multiple disciplines.

Dr. Muthu Ramachandran is Principle Lecturer (Associate Professor) at the School of Built Environment, Engineering and Computing at Leeds Beckett University, UK. Muthu has extensive research experience coupled with a teaching background and experiences on software and systems engineering methods and lifecycle, software development, agile software engineering, project management skills, process improvement skills, Internet technology, mobile, networks and distributed computing, realtime and embedded systems, cloud computing, service-oriented architecture and IT systems development for the past 25 years. He was a research scientist at India Space Research Labs where he worked on real-time systems development projects. Muthu is an author of two books: *Software Components: Guidelines and Applications* (Nova Publishers, NY, USA, 2008) and *Software Security Engineering: Design and Applications* (Nova Publishers, NY, USA, 2011). He is also an edited co-author of a book, *Handbook of Research in Software Engineering* (IGI, 2010), and has edited the book KE for SDLC (2011). He has widely authored published journal articles, book chapters and conferences materials on various advanced topics in software engineering and education. He received his master's degree from the Indian Institute of Technology, Madras, and from Madurai Kamaraj University, Madurai, India. He is a member of various professional organizations and computer societies: IEEE, ACM, Fellow of BCS and Fellow of HEA. He was a speaker to the 5th international symposium on SOA Cloud 2012, London, COMPLEXIS 2016 and FEMIB 2019.

Prof. Víctor Méndez Muñoz is Consultant Professor at UOC, Spain. He has a background in several fields of big data, cloud computing and IoT, both in academia and in the industry. Currently, he is contributing to the IoT ecosystem for microbiological control at IUL, which is an innovative company in the biotechnology. We are building new models and architectures taking full advantage of the state-of-the-art opportunities in IoT microcomputers, microservices and edge computing, to ensure genuine progress. Furthermore, he fosters bridging the gap between theory and practice, as Collaborating Editor in scientific journals and conferences, training postgraduates at UOC and partaking in innovation programs with Universities.

Part I
Area 1: Industrial Internet of Things (IIOT) Fundamentals and Applications

Chapter 1
A Modelling Framework for CPS-Based Industry 4.0: Application to Manufacturing Systems

Zakaria Benzadri, Takieddine Bouheroum, Youcef Ouassim Cheloufi, Mohamed Nadir Hassani, and Faiza Belala

Abstract Recently, digital transformation, known as the Fourth Industrial Revolution (Industry 4.0), has become a promising technological framework used to integrate and extend manufacturing processes at the intra and inter-organizational levels of smart factories. Among the most important technologies in Industry 4.0, cyber-physical systems (CPS) are recent complex systems, subject to distributed control, cooperation, influence, cascading effects, and emerging behaviours. Nowadays, few research attempts and industrial companies are interested in integrating CPSs to study, design, and implement more intelligent manufacturing systems. The main objective of this paper is to bring the migration to CPS-based Industry 4.0 and its new features while proposing a cross-layers and generic architecture (I4.0-CPS) validated through a realistic case study of the manufacturing process of a grader in an Algerian company (ENMTP-Somatel-Liebherr).

1.1 Introduction

The industrial revolution is a concept that has fundamentally changed our society and economy. It marks the passage from a system of artisanal production based on coal, textiles, and steam engines (First Industrial Revolution—mechanization), through a

Z. Benzadri (✉) · T. Bouheroum · Y. O. Cheloufi · F. Belala
LIRE Laboratory, University of Constantine, 2-Abdelhamid Mehri, Constantine, Algeria
e-mail: zakaria.benzadri@univ-constantine2.dz

T. Bouheroum
e-mail: takieddine.bouheroum@univ-constantine2.dz

Y. O. Cheloufi
e-mail: youcef.cheloufi@univ-constantine2.dz

F. Belala
e-mail: faiza.belala@univ-constantine2.dz

M. N. Hassani
ENMTP-SOMATEL, Constantine, Algeria

V. Chang et al. (eds.), *Modern Industrial IoT, Big Data and Supply Chain*,
Smart Innovation, Systems and Technologies 218,
https://doi.org/10.1007/978-981-33-6141-6_1

system that takes advantage of the fields of electricity, oil and chemistry (Second Industrial Revolution—Electrification), and to a system whose dynamics come from electronics, computers, audio–visual, and nuclear power (Third Industrial Revolution—automation). Recently, digital transformation, known as the Fourth Industrial Revolution, is a promising technological framework used to integrate and extend manufacturing processes at the intra- and inter-organizational levels of smart factories [1]. This emerging digital transformation claims to be a fusion of the virtual world, computer design, management (operations, finance, and marketing) and real-world products, and objects. It refers to the use of digital technology to make manufacturing operations more agile, flexible, and customer-focused.

Today, the evolution of information and communication technologies has led to the birth of the new industrial revolution. As a result, several strategic initiatives have been launched around the world: "smart manufacturing" in the USA, "Internet+" in China (see [2]), "industry of the future" in France, and "Industry 4.0" in Germany; the term Industry 4.0 seems to be gaining international acceptance. Companies in the world including Algerian ones are facing the Industry 4.0 revolution which imposes new challenges and new modes of production, such as exploitation and massive management of data, interconnection of machines, dematerialization of communication and distribution channels, the restructuring of the company for flexible and personalized production, etc. In this view, we identify the following key challenges, requiring each factory to act for the transformation to Industry 4.0: (**C1**) standardization of rules and processes that permits us to identify a set of findings to move towards Industry 4.0; (**C2**) connectivity and change tracking of hundreds of heterogeneous services, machines, or factories distributed over different sites; (**C3**) transition from re-engineering software systems and manufacturing processes to modular and maintainable solutions.

1.1.1 Context and Problematic

The Industry 4.0 is defined as a collective term for technologies and concepts of company's value chain [1, 3–5]. Cyber-physical systems, Internet of things, big data, cloud computing, edge computing, and fog computing are among the core technologies for Industry 4.0. In this context, cyber-physical systems (CPSs) are recent complex systems, subject to distributed control, cooperation, influence, cascading effects, and emerging behaviours. They may be considered as sensors and actuators that monitor physical processes and create a virtual copy of the physical world in a plant [6]. In general, CPSs are a new class of engineered systems that offer close interaction between cyber and physical components; they are expected to play a major role in the design and development of future systems. However, nowadays, few research attempts and industrial companies are interested in integrating CPSs to study, design, and implement more intelligent manufacturing systems, in the context of Industry 4.0. CPSs can play a major role in smart manufacturing and production processes, thus offering significant advantages in terms of time, resources, and costs compared to conventional production systems. Indeed, current modelling approaches of this type of systems are based on classical notations and languages assuming that are monolithic, based on central control, global visibility, hierarchical structures, and coordinated activities.

Within the software engineering framework, model-based solutions may hold promise for the development cycle of CPS-based production systems. These solutions focus on a more abstract concern than conventional programming, taking into account the seamless integration of heterogeneous CPS components and their coordination. In this work, we identify, in an explicit manner, all the stages of designing a distributed and intelligent production system, depending on the above CPS and Industry 4.0 principles [7], from abstraction to architecture, and from model to realization (in response to the identified challenges: C1; C2; and C3).

Related Work

In the literature, few research works exist for modelling CPS; most are driven by a generic methodology without reference to an explicit design process, focusing instead on isolated stages such as simulation, development, or verification. On the other hand, the used model notations (BPMN, Rdps, UML, etc.) for this type of systems are often inappropriate, emphasizing only one aspect while neglecting others. Some architectures, deduced from published works, have been identified and summarized in the table below, and their authors, each according to its particular area of interest, take into consideration only significant layers of these architectures, which can possibly host the Industry 4.0 components, but not leading to a model abstracting the details of the corresponding production systems and facilitating their design.

Through Table 1.1, we notice that almost of the existing approaches represent good attempts to model their corresponding manufacturing systems. Their established architectures are structured on several hierarchical layers: the network, the plant, the system, the cell, and the station. We also note that only the RAMI4.0 reference architecture makes it possible to describe the production system according to several views. In this case, a complete three-dimensional description has been defined. Concerning the third column of Table 1.1, it indicates how to define the individual components of a production system; obviously the architecture 5C inherits the CPS principle to distinguish between the physical elements of a production system, its computation components, and its connection elements (network). So, more importance must therefore be given to their integration. The genericity criterion depicts that the last two works in Table 1.1 are intended for the description of particular systems developed in a very particular context. Examples of technologies and components are considered.

In this paper, we propose a multi-view, cross-layer, and generic architecture (called I4.0-CPS) to respond to our major concerns, namely the abstraction of static and behavioural aspects, the distribution and heterogeneity of its entities, and the intelligent handling of data. The proposed approach is illustrated through a realistic case study of the manufacturing process of a grader in an Algerian factory (ENMTP-Somatel-Liebherr).

Table 1.1 Related work traits

Architecture	Cross-layer	Multi-views	CPS-based	Generic	Underlying model
RAMI 4.0 [8]	+	+	−	+	−
Architecture 5C [9]	+	−	+	+	−
PERFoRM [10]	+	−	−	−	−
FC-CPMTS [11]	+	−	−	−	−

1.1.2 Paper Organization

The remainder of the paper is organized as follows. Section 1.2 is dedicated to the presentation of the proposed cross-layers CPS-based architecture for Industry 4.0 (I4.0-CPS). Section 1.3 describes the case study to be used throughout the paper to illustrate our approach. Finally, Sect. 1.4 concludes the paper by giving some perspectives.

1.2 A Cross-Layers Architecture for CPS-Based Industry 4.0

Knowing that the architectural reference model of Industry 4.0 (RAMI4.0) consists of a three-dimensional coordination system, it describes all the essential aspects of Industry 4.0, namely [8] a layer dimension, a life cycle dimension, and lastly a dimension specifying the hierarchy. In the same thought, we give in this section a multi-layered architecture based on CPS, which makes it possible to understand and model an intelligent factory while integrating new concepts related to the decentralization and intelligence of production processes in Industry 4.0. This lies in the way to make the computational components of Industry 4.0 collaborate with the physical world in a distributed way using data access and processing services available on the Internet. The proposed architecture, illustrated in Fig. 1.1, offers two different views to describe a production system in an intelligent factory: **(1)** ***the functional view*** consists of four levels (see Fig. 1.1) which permit to specify the physical (resources, products, machines, routers, sensors, etc.) or computational (servers, computers, web application, etc.) elements of a production system that are described according to their position in the different layers (or levels); **(2)** ***the network view*** consists of three levels: edge, fog, and cloud computing, associating and specifying the means intended for the corresponding functional level. In what follows, we describe each of the functional view levels by referring to the associated network view levels:

- **Manufacturing asset level (MA)**. In this level, different robots, machines, and applications are available for the manufacturing process. The machines can detect products and retrieve the process information they need. As a result, some corresponding functionalities are identified (summarized on the right in Fig. 1.1).

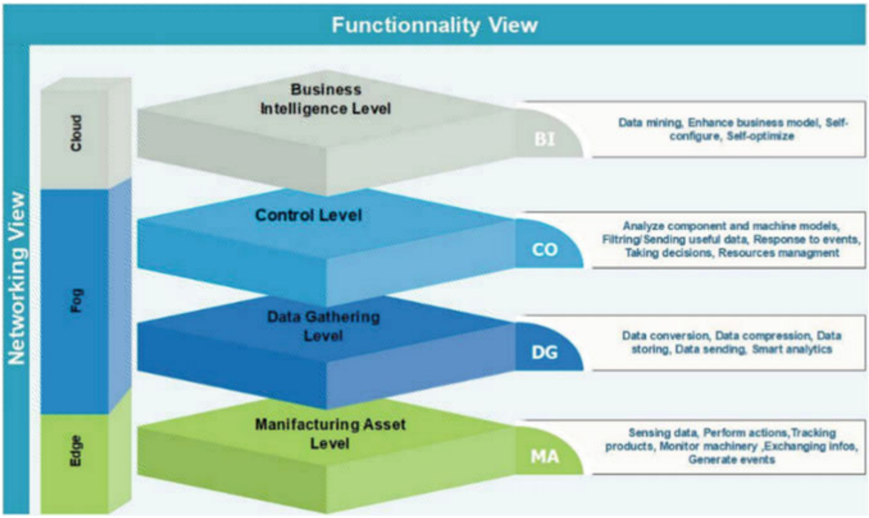

Fig. 1.1 Multi-layered architecture I4.0-CPS

- **Gathering data level (GD)**. Significant information must be inferred from the data collected. The IoT should offer promising transformational solutions for this, which can also be realized at the fog level.
- **Control level (CO)**. This level plays the pivotal role of the proposed architecture. Information is transmitted to it from every machine, process, or system connected to form the plant network. After collecting massive amounts of information, specific analyses must be used to extract useful information to better understand the state of individual components.
- **Business intelligence level (BI)**. This level acts as a supervisory control to make the machines self-configurable and self-adaptive.

1.3 Case Study: A Grader Manufacturing System

Algerian companies are facing a new industrial revolution that imposes new challenges, particularly with regard to new production and communication models. We contribute, through this work to bring a solution to a modelling problem posed by production systems in the context of Industry 4.0. ENMTP-Somatel-Liebherr, one of the Algerian national companies, is active in the production of hydraulic excavators, wheel loaders, self-propelled cranes, and graders. The proposed architecture dealing with the modelling of intelligent production systems is illustrated and projected on an example of the manufacture of the "SOMATEL grader 5410". This allows us to study in depth the manufacturing process of a grader in this company and to identify a set of findings to move this factory towards Industry 4.0, such as the appropriation

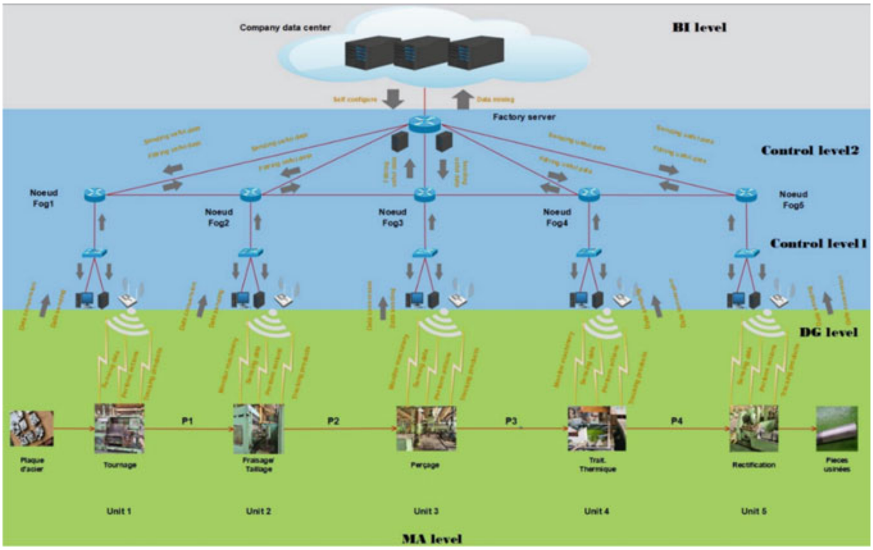

Fig. 1.2 Architecture instance of I4.0-CPS for the machining process

of new technologies, the control and the sharing of data across the network, organizational restructuring of the company, and the development of a digital plan. In Somatel-Liebherr, several micro-processes constitute the producing grader process: Debitage, Mecano welding, machining, pre-assembly, and final assembly. In this work, we are interested in the machining process, which begins with the blank steel wafer and ends with factory parts, having the desired shapes and characteristics. According to our formalization approach, Fig. 1.2 shows the application of the I4.0-CPS architecture for our case study. We deduce the following generic and cross-layers architecture for describing the spatial organization for our manufacturing process. We focus on the collaboration that occurs during the machining of parts used in, product manufacturing, synchronization, deployment, and decision-making. This collaboration is governed by a well-defined organization of the different elements constituting this I4.0-CPS architecture. Obviously, the physical and software components of the corresponding CPS are deeply intertwined, each operating at different spatial scales, represented by the different levels of the I4.0-CPS architecture: MA, GD, CO, and BI, and interacting with each other through fog or cloud computing, to describe the functioning of the intelligent machining process in SOMATEL as shown in Fig. 1.2. Two aspects are considered in this model:

Static aspect. We specify the physical elements (machines, resources, and manufactured products) at the MA level as well as some sensors useful to collect certain information. We have chosen to host each operation (turning, milling, drilling, heat treatment, and grinding) of the machining process in a separate unit. The next level in the architecture (GD level) contains the data conversion tools for exploitation. Level CO (control), distributed on two levels (CO_1 and CO_2) in our case study, contains the supervisory computers or servers, which collect the information from the lower levels and provide useful data or actions for other layers. We note that at level CO_1, the functionality associated with each operation in the machining process is performed by a separate fog node. CO_2 level is provided by two different servers: one for hosting the used applications and the other is reserved for data storage. At the BI level, we combine cloud servers to study, analyse, and make decisions in order to improve the functioning of the production process in question. Eventually, an optimization or reconfiguration can be considered.

Interaction aspect. The essential contribution of our approach is that it allows seamless integration between heterogeneous systems consisting of computing devices and/or distributed sensors and actuators. Sensors and actuators provide an interface between the physical and cyber worlds. The proposed I4.0-CPS architecture takes into account this aspect and pays particular attention to the interactions that may exist between the different components of this machining process in a well-defined organizational context, and they are schematized by arrows in Fig. 1.2. We thus distinguish between the inter-layer interactions linking the elements constituting the CPS, belonging to different levels of the I4.0-CPS architecture, and the intra-layer interactions governing the functional links between the elements of the same layer.

1.4 Conclusion and Perspectives

In this paper, we tackled the challenges in managing the life cycle development of production systems in smart manufacturing. We relied on CPSs to give an efficient framework for modelling this type of complex systems. Especially, we have defined a multi-level architecture I4.0-CPS dealing with structure and behaviour aspects.

Future research will aim at (1) the elaboration of the corresponding service-oriented architecture in order to describe typically the possible interactions between the physical and logical entities of CPS-based systems; (2) a transcription to a component-based model to take advantage of its abstraction and reuse aspects during the production system development life cycle: design, implementation, and deployment.

References

1. Qin, J., Liu, Y., Grosvenor, R.: A categorical framework of manufacturing for industry 4.0 and beyond. Procedia Cirp **52**, 173–178 (2016)
2. Chang, V., Xu, Y.K., Zhang, J., Xu, Q.: Research on intelligent manufacturing development approach for China's local valve industry. Smart Sustain. Built Environ. (2020)
3. Chang, V.I., Lin, W.: How Big Data transforms manufacturing industry: a review paper. Int. J. Strateg. Eng. (IJoSE) **2**(1), 39–51 (2019)
4. Benzadri, Z., Bouanaka, C., Belala, F.: Big-CAF: a bigraphical-generic cloud architecture framework. Int. J. Grid Util. Comput. **8**(3), 222–240 (2017)

5. Bouheroum, A., Benzadri, Z., Belala, F.: Towards a formal approach based on bigraphs for fog security: case of oil and gas refinery plant. In: 2019 7th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 64–71. IEEE (2019).
6. Khaitan, S.K., McCalley, J.D.: Design techniques and applications of cyberphysical systems: a survey. IEEE Syst. J. **9**(2), 350–365 (2014)
7. Xu, L.D., Xu, E.L., Li, L.: Industry 4.0: state of the art and future trends. Int. J. Prod. Res. **56**(8), 2941–2962 (2018).
8. Hankel, M., Rexroth, B.: The reference architectural model industry 4.0 (RAMI 4.0). ZVEI, April, 410. IBM. (2020). IBM Rational Software Architect Designer (2015). https://www.ibm.com/developerworks/downloads/r/architect/index.html.
9. Lee, J., Bagheri, B., Kao, H.-A.: A cyber-physical systems architecture for industry 4.0-based manufacturing systems. Manuf. Lett. **3**, 18–23 (2015)
10. Leitão, P., Barbosa, J., Pereira, A., Barata, J., Colombo, A.W.: Specification of the perform architecture for the seamless production system reconfiguration. In: Iecon 2016–42nd annual conference of the IEEE industrial electronics society, pp. 5729–5734 (2016).
11. Zhou, Z., Hu, J., Liu, Q., Lou, P., Yan, J., Li, W.: Fog computing-based cyber-physical machine tool system. IEEE Access **6**, 44580–44590 (2018)

Chapter 2
A Deep Classifier for Crowdsourcing User Requests

Feifei Niu, Chuanyi Li, and Bin Luo

Abstract Software feature requests are proposed by users online to ask for new features of software covering diverse aspects of software properties, such as usability, security, and performance. They are valuable sources of software requirements. Early detection of request categories enables them to be stored structurally and incorporated into requirements specifications. But manually analyzing and labeling the category of user requests is a labor-intensive and time-consuming task. In this paper, we propose a deep learning-based approach to automatically classify user requests, where both statistical and semantical text features are adopted. The main contributions of this work are comparing the effectiveness of different deep learning algorithms in feature request classification and exploring the proper way to apply deep learning to solve the requests classification issue. Three research questions are answered by experiments to illustrate the contributions. The experimental results derived from the dataset collected from Sourceforge.net show that the deep classifier works properly on utilizing different types of features, classifying user requests, and combining with active learning.

2.1 Introduction

As an early phase of software development, requirement classification is crucial. It can help to distinguish between types of requirements and help to detect early aspects in software development. For example, security is a significant aspect that counts for much [3] and usually has a high priority amount all requirements [2].

User requests are proposed by the crowd on open forums to ask for new features and are valuable resources of requirements. However, the number of user requests is growing rapidly and covers all aspects, making it a challenge to manage and utilize them. Besides, user requests are usually unstructured and colloquial, which turns out to be an obstacle to manually analyze and label their category.

F. Niu · C. Li (✉) · B. Luo
Software Institute, Nanjing University, Nanjing 210093, China
e-mail: lcy@nju.edu.cn

V. Chang et al. (eds.), *Modern Industrial IoT, Big Data and Supply Chain*,
Smart Innovation, Systems and Technologies 218,
https://doi.org/10.1007/978-981-33-6141-6_2

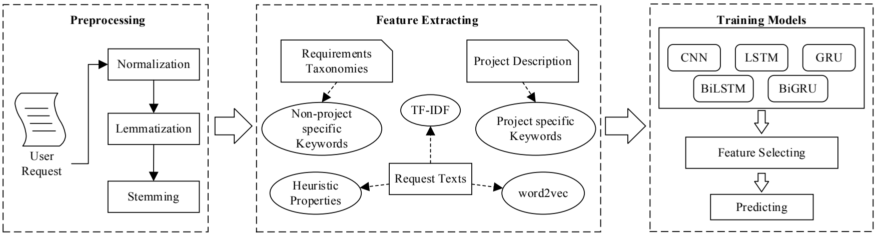

Fig. 2.1 Framework of the proposed deep learning based user feature requests classification approach

Recently, the rise of artificial intelligence has provided better solutions to many problems [16]. In this paper, we propose deep learning (DL) classifier to classify user requests automatically. We employ both semantic and statistical text features for training the classifiers including Word2vec [15], TF-IDF [12], keyword frequency (KF) [13], and heuristic properties (HP) [13]. The DL classifiers are convolutional neural network (CNN) [11], long short-term memory network (LSTM) [9], BiLSTM [17], gated recurrent units (GRU) [4], and BiGRU [17]. We proposed more HPs and evaluate the effectiveness of different HPs to select an optimized set of HPs for different projects. We carried out experiments on 3000 user requests collected from three projects on Sourceforge.net. The results are evaluated on accuracy, precision, recall, and *F*-measure. Most importantly, we designed two strategies based on cross-prediction and active learning to employ the classifier into actual scenarios. The framework of our research can be found in Fig. 2.1.

The remainder of this paper is organized as follows. Section 2.2 surveys requirement classification and related techniques. Section 2.3 introduces the approach. Experimental processes and results are illustrated in Sect. 2.4. Threats to validity are discussed in Sect. 2.5. Section 2.6 concludes the paper with a discussion of future work.

2.2 Background

2.2.1 Software Requirement Classification

Requirement classification has been studied for years. Cleland-Huang et al. [5] proposed an information retrieval-based approach to classify non-functional requirements (NFRs) with weighted indicator terms. Winkler and Vogelsang [19] have proposed an automatic classification of requirements employing convolutional neural networks. Kurtanovic and Maalej [10] automatically classify requirements into functional and non-functional. Baker et al. [1] leverage artificial neural networks and convolutional neural networks to classify non-functional requirements into five categories: maintainability, operability, performance, security, and usability. Mohamad et al. [14] identified security requirements from 3003 requirements with the random forest classifier. However, requirements are confidential assets for companies and are rarely available for domain experts [7]. The recent study for requirements engineering has transferred from requirements specifications to users' online reviews. Li et al. [13] defined taxonomy for user requests' classification and classified user requests into seven categories including security (SE), reliability (RE), performance (PE), life cycle (LI), usability (US), capability (CA), and system interface (SI). In this paper, we follow Li et al.'s [13] definitions of the seven categories.

2.2.2 Related Techniques

Recently, the blossoming of DL models allows us to extract semantic and syntactic information of texts automatically and obtain better and more comprehensive features [15]. Convolutional neural networks (CNNs) can retain word order information and learn sentence patterns composed of word sequences that span multiple words [8]. Recurrent neural networks (RNNs) have been widely used in processing sequential data and have achieved pretty good results. LSTM [9] and GRU [4] are the most commonly used RNNs. Besides, Schuster and Paliwal [17] proposed BiRNN models, which connect two opposite direction layers to the same output and can increase the input information to the network. In this paper, we evaluate these DL models, i.e., CNN, LSTM, BiLSTM, GRU, and BiGRU to classify user requests.

First, we employ feature extraction techniques to convert natural language texts into machine understandable representations including:

Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is commonly used to calculate the importance of a word in a document [12]. It increases proportionally to the number of times a word appears in a user request but is offset by the number of documents in the corpus that contains the word.

Keyword Frequency (KF). KF calculates the frequency of each type of keywords that appeared in each user request, where the keywords are predefined and categorized into the target seven classes, which are defined by Li et al. [13].

Heuristic Properties (HPs). Li et al. [13] have defined HPs as certain parts that strongly imply the possible category of the request.

Word2vec [15]. Word embedding maps words or phrases into vectors of numbers and is used to extract semantic information from texts.

2.3 Approach

The proposed approach, as depicted in Fig. 2.1, consists of three phases: preprocessing, extracting features, and training classifiers. Firstly, the user requests are preprocessed using common NLP techniques: normalization, lemmatization, and stemming. Then, both semantic and statistical features are extracted from feature

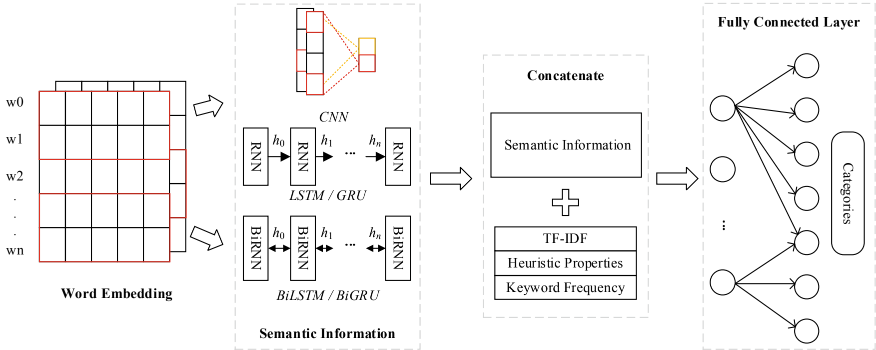

Fig. 2.2 Structure of CNN, LSTM, BiLSTM, GRU and BiGRU models

Table 2.1 Proportion of each type of user request in each project (1000 for each project)

Project	SE (%)	RE (%)	PE (%)	LI (%)	US (%)	CA (%)	SI (%)
KeePass	10.1	6.8	1.9	3.9	34.3	38.1	4.9
Mumble	2.4	6.3	3.5	12.3	33.2	33.7	8.6
WinMerge	0.9	7.4	2	5.7	41	40	3

texts. Semantic features are extracted via DL models. All features are concatenated as the input of the training model. After several fully connected layers, a softmax layer outputs a seven-dimensional vector that stands for the possibility of classifying to each category. The highest possibility is thought to be the final classification result. The model is shown in Fig. 2.2.

Our dataset consists of 3000 user requests of three projects on Sourceforge.net. The three projects are KeePass, Mumble, and WinMerge. The percentage of each category in each project is shown in Table 2.1.

To evaluate fairly, we use fivefold cross-validation experiments on each project. Each time, we use one fold for testing, one for validation, and one for training.

We use average accuracy, precision, recall, and $F1$ of fivefolds to evaluate the performance of the classifiers, where accuracy is the percentage of records classified correctly in the test set.

2.4 Experiment and Results

2.4.1 Experiment

Our experiments revolve around solving three research questions. The questions, research methods, and corresponding results are as follows:

RQ1 Are all the HPs helpful to identify the category of user requests? Is there an optimal set of HPs that is most useful for classification?

Method: To examine the effectiveness, we test on each HP to get accuracy variation with the number of added HPs changes. Firstly, we define an empty set S_{HP} and a set S that contains all eighteen HPs. S_{HP} is the set of HPs used as the input of models. For each experiment, choose one HP from S and add to S_{HP}, train the model with features extracted from S_{HP}, which is concatenated with Word2vec, TF-IDF, and KF. In this step, we test on all the eighteen HPs one by one and pick out the one that gains the highest result. The high accuracy indicates that this HP is more indicative. Move this HP from S to S_{HP}. Then, repeat this step, until S is empty. We can obtain the accuracy curve that depicts the accuracy varies with the added number of HPs increases. The algorithm that we employ is LSTM, which can learn long-term dependency information.

Results: We test on three projects, and the accuracy curves are shown in Fig. 2.3. The X-axis stands for the number of added HPs, and the Y-axis is the accuracy result. From the result, we can see that the accuracy curve increases as HPs added at first, but it reaches a peak and then begins to decline. The optimal HP set is obtained when the curve reaches pear. The accuracy improvement can be one percent with the optimal HP set compared with all HPs in three projects. The accuracy improvement of the optimal HP set is nearly 2 percent higher than models without HPs. However, the optimal HP set is different in three projects. In KeePass, the optimal HP set is 9, 13, 1, 5, 14, 4, 3, 7. In Mumble, the optimal HP set is 9, 13, 1, 12, 5, 7, 15, 8, 17, 3, and in Winmerge, the optimal HP set is 9, 1, 4, 10, 16, 8, 5, 17, 6. The optimal set is different between different projects. The reason may be that the semantic is different in different projects, so does the HPs' effectiveness. During the experiment process, we find that the ninth HP is most effective in both three projects.

RQ2 How well can we automatically classify user requests with DL classifiers? Do DL classifiers perform better than traditional machine learning classifiers?

Methods: Firstly, the word vector model is trained with the training set. Then, we build DL classifiers with word vectors as inputs. To inspect the accuracy improvement when combined with TF-IDF, KF, and HPs, we concatenate TF-IDF, KF, and HPs to the word vectors. We experiment on different DL classifiers to explore the distinction. To

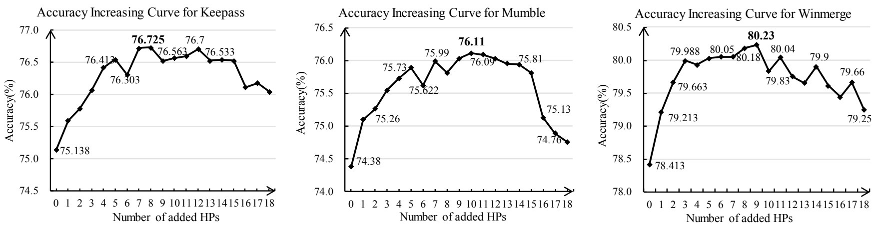

Fig. 2.3 Accuracy curve of projects with the number of input HPs increases

compare the performance of every algorithm, we apply one-way analysis of variance (one-way ANOVA). The hypothesis is that there is no significant difference between different DL classifiers.

Results: Table 2.2 presents the experimental results of different combinations of text features and models. The highest average accuracy of Word2vec is achieved by LSTM, which is 47.83, 39.93, and 52.20, respectively. The accuracy improvement of combining TF-IDF with Word2vec is 16–25%. When adding KF, the improvement is 8–11%. When the optimal set of HPs is added in, the improvement is 0.9–2.3%. The results show that TF-IDF, KF, HPs are all effective in improving accuracy results. When using all the statistical and semantic text features, the classifiers can obtain the best results. Our average results are higher than the previous results achieved by traditional machine learning algorithms by Li et al. [13] apart from the CNN classifier. One-way ANOVA test result is shown in Table 2.3. The results indicate that there is no significance between the five algorithms for KeePass. LSTM and BiLSTM gain the best result in Mumble, with CNN gaining the worst result. In Winmerge, there is no significance among LSTM, BiLSTM, and GRU, which are both better than BiGRU and CNN. In general, LSTM and BiLSTM perform better than GRU and BiGRU, and they are all better than CNN in all the three projects. The precision, recall, and $F1$-measure results of the LSTM classifier are shown in Table 2.4. $F1$-measure of all categories has been improved than that in Li et al. [13].

RQ3 In the actual scenarios, when users start a new project, all the user requests are unlabeled at first, and there is no available training set. How can we predict these user requests with DL algorithms?

Method: To label new user requests, there are two strategies. The first solution is to predict these requests with classifiers trained by other similar projects, which is cross-project prediction. We study the relationship between project similarity and cross-prediction accuracy. Firstly, we calculate project similarity through user requests with different algorithms like cosine similarity, Euclidean distance, Manhattan distance, and Jaccard index. Then, we train the classifier on one project to predict another project. The second solution is by means of active learning strategy. Active learning is a semi-supervised learning algorithm in the context of classification to reduce the efforts involved in labeling [6]. The chosen algorithm is LSTM and all the extracted features. The original dataset is split into training set and testing set with a rate of 0.2. The chosen sampling strategies are least confident (LC), small margin (SM), and random (R) [18].

Results: The similarity results and cross-project prediction results are shown in Table 2.5. All the similarity algorithms indicate that KeePass and Mumble gain the highest similarity result and KeePass against Mumble follows. The least similarity is Mumble and Winmerge. For predicting KeePass, Mumble, and Winmerge, Winmerge, Winmerge and KeePass gain higher results, respectively. The predicting results are proportional to similarity except for predicting Mumble. When predicting Mumble, Winmerge gains higher accuracy while KeePass gains higher similarity

Table 2.2 Accuracy of deep classifiers with different text features

Accuracy (%)	KeePass				Mumble				Winmerge			
	W	WT	WTK	WTKH	W	WT	WTK	WTKH	W	WT	WTK	WTKH
CNN	44.70	65.97	75.20	76.83	35.60	61.63	74.33	75.23	45.67	67.53	77.63	79.00
LSTM	**47.83**	65.43	75.18	77.00	39.93	63.20	**75.43**	76.53	**52.20**	68.53	**78.55**	**80.45**
BiLSTM	46.50	**66.95**	**75.55**	77.05	37.95	**63.30**	74.95	**76.58**	46.90	68.48	78.50	80.25
GRU	42.13	66.28	75.20	**77.13**	39.48	63.08	74.93	76.20	47.15	68.73	78.50	80.20
BiGRU	44.30	65.78	74.78	77.05	38.53	63.05	74.48	76.03	47.80	**68.98**	77.83	79.78

K—for keyword frequency, H—for heuristic property, T—for TF-IDF, and W—for Word2vec