International Journal of Software and Informatics, ISSN 1673-7288
http://www.ijsi.org, ijsi@iscas.ac.cn, +86-10-62661048
IJSI, 2023, 13(1): 1–4, doi: 10.21655/ijsi.1673-7288.00293

Preface

Preface to the Special Issue on Theories and Technologies for Big Data Governance

Xiaoyong Du (杜小勇)[1,2], Xiaochun Yang (杨晓春)[3], Yongxin Tong (童咏昕)[4]

[1](Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872, China)
[2](School of Information, Renmin University of China, Beijing 100872, China)
[3](School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)
[4](School of Computer Science and Engineering, Beihang University, Beijing 100191, China)
Corresponding author: Xiaoyong Du, duyong@ruc.edu.cn; Xiaochun Yang, yangxc@mail.neu.edu.cn; Yongxin Tong, yxtong@buaa.edu.cn

Citation Du XY, Yang XC, TongYX. Preface to the special issue on theories and technologies for big data governance, *International Journal of Software and Informatics*, 2023, 13(1): 1–4. http://www.ijsi.org/1673-7288/293.htm

In the era of digital economy, data has become an important factor of production, and big data technology has become the key technological engine for the development of data element markets. However, the phenomenon of "valuing collection, scale, and utilization while undervaluing management, quality, and security" is prevailing during the utilization of big data in recent years. Scientific and effective big data governance is useful for improving the level of big data management and decision-making, generating high-quality data, reducing cost, enhancing security control, and lowering risks. Based on the supporting role of data governance to big data technology, this special issue aims to explore the central challenges faced by big data governance and focuses on the management of various links in data's full life-cycle by big data governance, i.e., the quality issues on the metadata and master data appearing in the data integration and storage links such as data acquisition, cleaning, and conversion and the policy and standard issues such as big data capitalization and standardization, as well as data sharing, security, and application; meanwhile, it discusses the framework system of big data governance. The special issue pays attention to the existing data theories, explores new theories, new technologies, and new methods for big data governance by means of blockchain, federated learning, knowledge graph, data pricing, data analysis, etc., and studies the latest application achievements of big data governance in various fields. Specifically, it includes four aspects: (1) big data quality management technology; (2) federated computing technology for big data; (3) big data governance technology in complex and dynamic environments; and (4) applied technologies for big data governance.

In recent years, significant progresses in the research on basic theories and key technologies of databases and novel data governance theories and methods for big data have been made in China. The representative research results are summarized as follows.

(1) Big data quality management technology. Inferior data has become one of the major obstacles hindering the flow of current data elements. At present, how to effectively

Received 2023-01-18; Accepted 2023-01-20; IJSI published online 2023-03-30

manage and improve the quality of heterogeneous multi-source big data is the core issue in the field of big data governance. Aiming at the challenge of information silos in big data of government affairs, the team of researchers from Renmin University of China proposed a model based on the prediction phase and the error-correction phase, established the semantic alignment between attributes and standard metadata, and achieved the standardized governance of silo metadata. Focusing on the management of the metadata of Internet of Things (IoT), the team of researchers from Tsinghua University proposed a physical metadata management solution in Apache IoTDB for accelerating aggregate queries, which improved the metadata quality and optimized the system efficiency by integrating both synchronous and asynchronous computational strategies.

(2) Federated computing technology for big data. In recent years, federated computing technology with the goal that "original data is bound to its domain, and data is available but invisible" has provided a new perspective for studying cross-domain big data governance. Facing the issue of multi-source cross-domain data error detection, the team of researchers from Zhejiang University proposed a cross-source data error detection method based on federated learning, namely FeLeDetect, which uses cross-source data information to improve error-detection accuracy without privacy leakage. Focusing on trusted federated learning technology, the team of researchers from Chengdu University of Information Science and Technology designed an Efficient Decentralized Federated Learning (EDFL) framework by integrating blockchain technology into federated learning, which employs a consensus mechanism based on proof-of-contribution, a role-adaptive incentive algorithm, and a blockchain partition storage strategy to reduce the storage overhead and improve the learning efficiency. In terms of the optimization of connecting and querying data federation, the team of researchers from Beihang University proposed a θ-join algorithm based on secure Multi-Party Computation (MPC) and designed a series of optimization strategies by integrating multiple privacy computing technologies, which significantly improves the query efficiency of large-scale data federation.

(3) Big data governance technology in complex and dynamic environments. Big data governance in complex and dynamic scenarios is the central challenge to the full life-cycle management of big data; how big data governance technology can adapt to the high-frequency and dynamic changes in the data itself is one of the research hotspots today. Facing the challenges to dynamic metadata management in scenarios oriented towards big data sharing, transmission, and update, the team of researchers from National University of Defense Technology proposed a dynamic sketch for big data governance, which could simultaneously achieve the space overhead increasing linearly with the dataset cardinality and the constant time overhead of data processing and analysis and effectively support demanding multiple big data processing and analysis tasks. Focusing on dynamic data protection in an open big data environment, the team of researchers from ZTE Corporation carried out research from various aspects such as data availability, processing efficiency, and system scalability, designed a dynamic data protection system, BDMasker, and proposed a query rewriting technology based on query dependency model, which thus enables a whole process of dynamic desensitization with no impact on the business.

(4) Applied technologies for big data governance. As big data governance technology emerged in recent years, a large number of new data governance applications have appeared in the fields of public security, public health, E-government, financial telecommunications, and social media. Facing the challenge of data governance in heterogeneous information networks, the team of researchers from Northeastern

University proposed a maximum core mining problem in heterogeneous information networks based on attribute fairness, designed the Adv-FkPcore algorithm to circumvent the challenge of high computational complexity judged by subgraphs in the mining phase, and optimized the traversal efficiency of the algorithm for heterogeneous information networks by means of the Traversal Method with vertex Sign (TMS). Aiming at the issue of privacy leakage in surveillance video applications, the team of researchers from Xi'an Jiaotong University proposed a safe and fast video retrieval model for massive surveillance videos. According to the characteristics of high computing power in clouds and low computing power in surveillance cameras, the team carried out customized knowledge distillation for the model by designing a tolerance training strategy, deployed the distilled lightweight model inside a surveillance camera, and simultaneously used a local encryption algorithm to encrypt the sensitive part of the images, thereby protecting privacy with extremely low resource consumption.

This special issue contains five representative papers on big data governance, and their contents are briefed as follows.

In *Construction and Optimization Co-occurrence-attribute-interaction Model for Column Semantic Recognition*, a model based on the prediction phase and the error-correction phase is proposed for coping with the issue of difficult interconnection among the metadata semantics of various silo government systems in government data governance. In the prediction phase, a Co-occurrence-Attribute-Interaction (CAI) model is proposed, while in the error-correction phase, the predicted result of the model is optimized by combining the co-occurrence between semantic labels and introducing the error correction mechanism.

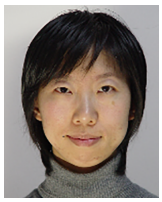
In *Cross-source Data Error Detection Approach Based on Federated Learning*, a federated learning-based cross-source data error detection method, FeLeDetect, is proposed, which used cross-source data information to improve error detection accuracy without privacy leakage. In order to reduce the communication overhead and the cost of manual labeling in federated training, a series of optimization methods are designed, which greatly improved the error detection rate in both local and centralized scenarios.

In *Jump Filter: A Dynamic Sketch for Big Data Governance*, a dynamic sketch for big data governance is proposed, which could achieve the space overhead increasing linearly with the dataset cardinality and the constant time overhead of data processing and analysis simultaneously and effectively support demanding multiple big data processing and analysis tasks.

In *BDMasker: Dynamic Data Protection System for Open Big Data Environment*, a dynamic data protection system for open big data environment, BDMasker, is proposed, which achieves a whole process of dynamic desensitization with no impact on the business by the precise query analysis and the query rewriting technology based on query dependency model.

In *Secure Multi-party θ-join Algorithms Toward Data Federation*, a secure multi-party θ-join algorithm for data federation is proposed, and a series of optimization strategies are designed by combining privacy computing technologies such as secure MPC, without disclosing the respective original data, which thus significantly reduces the security computation cost needed for join queries and greatly improves the query efficiency.

Xiaoyong Du, Ph.D., professor of Renmin University of China, doctoral supervisor, director of the Key Laboratory of Data Engineering and Knowledge Engineering, Ministry of Education, and chairman of CCF Big Data Committee. He is mainly engaged in research on intelligent information retrieval, high-performance databases, and unstructured data management. He has presided over more than 20 national key research and development projects and is the winner of multiple national and provincial science and technology awards.



Xiaochun Yang, Ph.D., professor of Northeastern University, doctoral supervisor. She has presided over 20 projects from the National Science Fund for Distinguished Young Scholars, the National Natural Science Foundation of China, and the National Key Basic Research Program of China (973 Program). She has published more than 100 papers and won three best paper awards at international conferences and four best paper awards at national conferences. She has authorized and publicized 23 Chinese patents, including one US patent. She is the winner of 21 provincial and ministerial awards. She is mainly engaged in research on big data management and knowledge engineering, database theory and systems, intelligent system recommendation, data quality management, and data privacy protection.

Yongxin Tong, Ph.D., professor of Beihang University, doctoral supervisor. He is mainly engaged in research on big data, databases, federated learning, privacy computing, and swarm intelligence. He has presided over more than 10 projects from the National Natural Science Foundation of China, the National Science Fund for Distinguished Young Scholars, and key programs and won multiple science and technology awards such as the first prize in the natural science of Chinese Institute of Electronics (CIE).