International Journal of Software and Informatics, ISSN 1673-7288
http://www.ijsi.org, ijsi@iscas.ac.cn, +86-10-62661048
IJSI, 2021, 11(1): 1−4, doi: 10.21655/ijsi.1673-7288.00244

Preface

Preface to the Special Issue on Data Management and Analysis Technique Supporting AI

Lei Chen (陈雷) [1], Hongzhi Wang (王宏志) [2], Yongxin Tong (童咏昕) [3], Hong Gao (高宏) [2]

[1] (Department of Computer Science and Engineering, the Hong Kong University of Science and Technology, Hong Kong 999077, China)

[2] (Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China)

[3] (School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Corresponding author: Hongzhi Wang, wangzh@hit.edu.cn.

Citation Chen L, Wang HZ, Tong YX, Gao H. Preface to the Special Issue on Data Management and Analysis Technique Supporting AI. *International Journal of Software and Informatics*, 2021, 11(1): 1−4. http://www.ijsi.org/1673-7288/00244.htm.

In recent years, the data management and analysis technique supporting Artificial Intelligence (AI) has become one of the hot issues in the field of big data and AI. Using and developing theory and technology of data management and analysis provide a basic support for improving the efficiency and effectiveness of the life cycle of AI systems and will surely further promote the development of AI technology based on big data and its wider application. In particular, AI technology represented by machine learning extracts knowledge by modeling data, and one training process includes multiple sub-processes such as data selection, feature extraction, algorithm selection, hyper-parameter tuning and effect evaluation. After the effect evaluation is obtained at the end of the training, it is usually necessary to manually analyze model effect to mine the relationship of model effect with data, features and algorithms, and the training sub-processes are adjusted and iterated for multiple rounds based on data analysis and artificial experience. Apparently, machine learning tasks are much more complicated than query and analysis tasks of database systems. Due to the large number of training sub-processes and iteration adjustments of machine learning and many sub-processes requiring manual participation, the training process is still task-oriented, and the training sub-process is customized and optimized according to the features of the task. This approach has a high cost of labor participation and cannot reuse resources such as data, features and models between multiple tasks. Therefore, there are problems of high cost, low efficiency and high energy consumption. Then how to reduce the management cost in AI computing processes such as machine learning to improve its intelligent computing efficiency has become a core challenge in this field.

In contrast, data management technology, especially the development history of database management systems, has formed a methodology for efficient data management, including data model as the core, hierarchical structure to achieve independence between data and application, task-oriented descriptive language and query optimization technology to improve task execution efficiency. From the perspective of data management, each sub-process of machine learning

Received 2021-03-25; IJSI published online 2021-03-31

involves different types of data read-write, conversion and calculation and has significant data management and analysis requirements. Applying key database technologies and system construction experience to the AI computing processes such as machine learning will help develop a systematic management plan for the AI field, thereby improving overall efficiency.

Therefore, this special issue focuses on the optimization and supporting role of database technology for AI in the process of the integration of data management and AI, including two aspects: (1) the optimization of the theoretical technology of traditional data management and analysis for the data and computing process of AI; (2) the promoting role of the design concept of traditional data management system in developing the general and easy-to-use AI platform. In particular, in recent years, many scholars in China have tried to use and develop existing database theory and technology to study how to build new data management and analysis methods to support AI. Representative research results can be summarized as follows:

(1) Data management and analysis provide support for AI data. The effectiveness of machine learning relies heavily on a large amount of training data. How to efficiently filter, organize, store and read large-scale and high-quality data for model training, evaluation and service of machine learning is an important issue to be solved in the development of AI today. For example, the research team of Tsinghua University proposed a time series data cleaning and repairing method with regard to intelligent time series classification and used dynamic programming to solve the problem of data anomalies with the optimal repairing path, there by verifying the feasibility and effectiveness of this method. Especially, it can improve the quality of AI results. The research team from Northeastern University focused on knowledge tracking, aiming to track changes in the knowledge level of students in real time according to their historical learning behavior, and predicted the future learning performance of students. It proposed a deep knowledge tracking model, LFKT, which integrates learning and forgetting behavior. This model adopts an efficient deep neural network, uses the students' answers to the questions as indirect feedback of the degree of knowledge mastery in the knowledge tracking process and builds a knowledge tracking model that integrates learning and forgetting behavior.

(2) Data management and analysis support AI algorithm optimization. Existing research pays more attention to the effect of AI algorithms, but less to the optimization of operation efficiency. Combined with indexed access of optimized AI data and other technologies, an optimization theory system regarding the efficiency of time and space of AI algorithms can be built to further improve the learning efficiency and service response speed of AI technology. For example, the research team from Peking University designed an efficient multi-tree index structure and a random sampling strategy for graph nodes to improve the execution efficiency of graph classification algorithms and solve the problem of embedding expression in temporal graph. In addition, the research team from Tsinghua University researched the optimization of AI models in the database. They optimized AI model inference by designing a "pre-screening + verification" framework, analyzed and explored the optimization technology of multiple machine learning models such as decision trees and improved the usability of decision tree training and inference operations by extending the SQL language. In addition, facing the problem of data silos, the research team from Beihang University studied the problem of federated learning-to-rank and designed a cross-silo federated learning strategy. The privacy protection technology based on the sketch structure and the federated semi-supervised learning method are designed to improve the effectiveness of the school algorithm.

(3) Data management and analysis support the AI model management. A model is usually the output of AI. However, in the process of constructing an AI model, it is usually necessary to repeatedly debug parameters and obtain different model versions. In addition, the new data that is continuously imported will also have an impact on the model. Some models may take up much space. Therefore, a set of management methods is required for AI models, which can depend on

data management technology. For example, the research team from Renmin University of China deconstructed and modeled the machine learning training process from the perspective of data management. They proposed a research framework of data management technology that supports machine learning and managed and optimized data selection, data storage, data access, automatic optimization and system implementation in the process of machine learning.

(4) Data management and analysis support AI system construction. With the increasing influence of AI, it should not only be available for scientists who master AI technology. Using big data technology and rich experience of building large-scale systems in the field of database can help build AI systems with low thresholds and strong versatility and enhance the inclusiveness of AI technology. For example, the research team from Harbin Institute of Technology studied the application of existing information to query prediction of knowledge graphs to preload and cache data, improving the response efficiency of the system. They proposed the method of extracting SparQL query in the sequence form and used Seq2Seq to analyze and predict the data, revealing good effect through massive experiments. In addition, the research team from Tianjin University proposed a knowledge graph database (KGDB) management system with the unified data model and query language. They introduced a unified storage scheme, enabled untyped triple storage and realized the interoperability between two knowledge graph query languages. The results verify that the system saves 30% of the storage space on average compared with gStore and Neo4j and the query efficiency can be improved by two orders of magnitude at most. The research team from the Computer Network Information Center of the Chinese Academy of Sciences proposed a fusion management system, PandaDB, of distributed data based on the intelligent property graph model, which realized efficient storage and management of structured/unstructured data and provided the AI operator extension mechanism. With regard to entity disambiguation and visualization of multiple heterogeneous academic graphs, highly concurrent response can be achieved by large-scale attribute filtering and the AI operator extension mechanism of the system. The performance of the system is good and stable.

In summary, this special issue focuses on the core database technology and discusses the role of data management and analysis technique in promoting the AI research. It particularly explores the optimization of AI in data and compute-intensive links by the theoretical technology of data management analysis and the support from the design concept and development experience of the data management system for the construction of a general AI platform. It highlights the support of data management and analysis technique for AI in data storage, algorithm optimization, model management, model service and system construction.

Brief introductions to five representative papers included in this special issue are as follows:

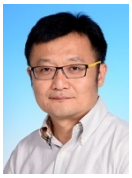
Node Embedding Research over Temporal Graph proposed an adaptive temporal graph embedding, ATGEB. Combined with the information propagation characteristics in the temporal graph, an adaptive cluster method was proposed for the node active frequency, and bidirectional multi-tree and a node sampling strategy were designed. Desired experimental results had been achieved in terms of node clustering, reachability prediction, and node classification in temporal graphs.

Time Series Data Cleaning under Multi-speed Constraints proposed a time series data repairing method under multi-speed constraints and used dynamic programming to solve the problem of data anomalies with the optimal repairing path, there by verifying the feasibility and effectiveness of this method. Especially, it can improve the quality of artificial intelligence results.

Local Semantic Structure Captured and Instance Discriminated by Unsupervised Hashing introduced a deep unsupervised hashing learning framework based on semantic structure preservation and instance discrimination, which guided the learning of hash coding while learning the semantic structure. It has been verified that the framework can effectively improve the discriminative ability of hash coding.

PandaDB: Intelligent Management System for Heterogeneous Data designed a fusion management system, PandaDB, of distributed data based on the intelligent property graph model, which realized efficient storage management of structured/unstructured data, provided a flexible AI operator extension mechanism and had the ability to ad hoc query the internal information of multiple heterogeneous data.

KGDB: Knowledge Graph Database System with Unified Model and Query Language developed a knowledge graph database (KGDB) management system with the unified data model and query language. It proposed a unified storage scheme, enabled untyped triple storage and realized the interoperability between two knowledge graph query languages. The results verify that the system saves 30% of the storage space on average compared with gStore and Neo4j and the query efficiency can be improved by two orders of magnitude at most.



Lei Chen, Ph.D., Chair Professor of The Hong Kong University of Science and Technology, Ph.D. Supervisor, IEEE Fellow, ACM Distinguished Scientist and winner of the National Natural Science Foundation of China for Distinguished Young Scholars (Overseas), serves as the Director of the HKUST MOE/MSRA Information Technology Key Laboratory and the Director of the Big Data Institute of The Hong Kong University of Science and Technology. He is mainly engaged in the research on big data, database, swarm intelligence and machine learning. He is also the editors-in-chief of *VLDB Journal*, associate editor-in-chief of *IEEE Transaction on Data and Knowledge Engineering* and member of the VLDB Council. He won the "SIGMOD Test of Time Award" and the "VLDB2014 Excellent Demonstration Award".

Hongzhi Wang, Ph.D., professor of Harbin Institute of Technology, Ph.D. supervisor, assistant dean of Honors School of HIT, distinguished member of CCF, is mainly engaged in the research on big data management and analysis, database system and data governance. He presided over more than 10 projects, including key projects of the National Natural Science Foundation of China and the National Key Technology Research and Development Program, and published more than 200 papers. He serves as the chairman of CCF Harbin Branch, standing committee member of CCF Technical Committee on Databases, secretary general of ACM SIGMOD China, member of CCF Task Force on Big Data, member of CCF-TC Computer Applications and expert of ACM Data Science Subject Standards Compilation Group. He won the first prize of Natural Science Award of Heilongjiang Province, the first prize of Science and Technology Progress Award of Higher Education of Ministry of Education and Youth Science and Technology Award of Heilongjiang Province.

Yongxin Tong, Ph.D., professor of Beihang University, Ph.D. supervisor, senior member of CCF, is mainly engaged in the research on big data, databases, federated learning, spatiotemporal big data computing and crowd intelligence. He led more than 10 projects, including those supported by National Natural Science Foundation of China,National Outstanding Youth Science Foundation, and National Key R&D Program of China, and published more than 80 papers. He won the first "DAMO Academy Young Fellow" by Alibaba Group, "VLDB2014 Excellent Demonstration Award" and "the Champion of KDD Cup 2020 RL Track".

Hong Gao, Ph.D., professor of Harbin Institute of Technology, Ph.D. supervisor, director of the Provincial Key Laboratory of Big Data Science and Engineering, is mainly engaged in the research on massive data calculation and analysis, social network analysis, management and analysis of spatiotemporal series data, data quality, perception data collection of Internet of Things and distributed perception data calculation. She participated in more than 20 major projects, including the major projects of National Natural Science Foundation of China, the key projects of National Natural Science Foundation of China and key research and development topics of the Ministry of Science and Technology, and published more than 200 academic papers.