applied sciences

MDPI

Editorial

Big Data Analysis and Visualization: Challenges and Solutions

Kwan-Hee Yoo [1], Carson K. Leung [2,*] and Aziz Nasridinov [1]

[1] Department of Computer Science, Chungbuk National University, Cheongju-si 28644, Chungcheongbuk-do, Korea
[2] Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
* Correspondence: kleung@cs.umanitoba.ca or carson.leung@umanitoba.ca

1. Introduction

Big data have become a core technology to provide innovative solutions in numerical applications and services in many fields. This calls for *big data analysis*. It is this process of examining these big data to discover information—such as hidden patterns, unknown correlations, market insights, and customer preferences—that can be useful for making various different business decisions. Recent advances in machine learning (including deep learning) and data mining have improved to the point where these techniques are used for analyzing big data from various applications in healthcare, manufacturing, social life, etc.

Moreover, big data have also been investigated using *big data visualization* as well as big data analytics. In particular, different visual analytical tools assist in visualizing new meanings and interpreting the big data, and thus helping to explore big data and simplify complex big data analytics processes.

This Special Issue on big data analysis and visualization discusses some challenges on analyzing and visualizing big data, as well as solutions to tackle these challenges. Solutions include:

- Big data preprocessing techniques (e.g., acquisition, integration, cleaning);
- Novel algorithms for big data analysis;
- Data mining and machine learning (e.g., deep learning) for big data analysis;
- Applications of computer vision techniques in big data analysis;
- Visual analytics of big data;
- Visualization techniques for supporting the big data analysis process;
- Data structures for big data visualization;
- Case studies and applications of big data visualization in a wide variety of fields.

Citation: Yoo, K.-H.; Leung, C.K.; Nasridinov, A. Big Data Analysis and Visualization: Challenges and Solutions. *Appl. Sci.* **2022**, *12*, 8248. https://doi.org/10.3390/app12168248

Received: 23 June 2022
Accepted: 17 August 2022
Published: 18 August 2022

2. Challenges and Solutions for Big Data Analysis and Visualization

We received more than 30 submissions for this Special Issue. After rigorous reviews by independent international reviewers, 11 refereed articles were selected for inclusion in this Special Issue. They cover different aspects of the challenges and solutions for big data analysis and visualization. Here, we give an overview of these challenges and solutions. Note that some of them focus on the fundamentals of big data analysis, while others describe interesting big data applications.

2.1. Big Data Analysis

Nowadays, *massively parallel* tasks—which make use of a large number of compute processors to simultaneously compute complicated mathematical operations in parallel—are needed for numerous big data applications. To perform these tasks, platforms such as Compute Unified Device Architecture (CUDA) and Open Computing Language (OpenCL) are widely used because they enhance the throughput of massively parallel tasks. Observing the demand in high-level abstractions and platform independence across these massively parallel computing platforms, a new cross-platform abstraction layer for single-source heterogeneous computing with C++ template-level abstractions for OpenCL, called

SYCL, was introduced. As there are several SYCL implementations from various vendors, Shin et al. [1] compared the large-scale data computing performance of these SYCL heterogeneous parallel processing layer implementations, especially when handling well-known massively parallel tasks. They analyzed their characteristics, strengths and weaknesses when computing various types of mathematical operations and/or data sizes. These comparison results help users to select a suitable SYCL implementation to maximize kernel performance in large-scale big data parallel computing and visualization applications.

Moreover, as demands for computing power grow, data processing for efficient resource management in traditional workstations has been under pressure. Consequently, a tremendous volume of big data has been processed using parallel computing. As an example, HTCondor—which is an open-source high-throughput computing (HTC) software framework for coarse-grained distributed parallelization of computationally intensive tasks—provides researchers with computing power to conduct data analysis. Although such a framework works effectively in a traditional computing cluster environment, an efficient methodology is required to meet the growing demand for computing with limited resources. Kong et al. [2] consolidated job schedulers in traditional independent scientific workflows. Their approach integrates those clusters that can share their computing power based on a priority policy. While sharing worker nodes, it maintains the resources allocated to each group. It also utilizes the historical data of user usage for analyzing problems that may have occurred during job execution caused by resource sharing and/or the actual operating results. Experimental results show the benefits of integrating resources and sharing limited computing powers by multiple scientific groups.

Note that high-performance computing (HPC) reduces overall job execution time and increases the capacity to solve large-scale complex problems using many distributed computing resources to solve problems through parallel computation. As a flagship tool in HPC, the job scheduler in a supercomputer is mainly responsible for distributing and managing the resources of large systems. Yoon et al. [3] conducted a case study on analyzing the execution log of the job scheduler over a period. It aims to improve execution time and resources in HPC, and thus reduces the idle time of jobs. The simulation results of the Tachyon2 system reveal that the causes of delayed jobs were strongly related to resource waiting. To elaborate, an increase in idle resources, which must be ready when applying for large-scale job, affected and significantly delayed the execution of the whole job. Consequently, Yoon et al. presented a backfilling algorithm to use available resources and reduce execution time for backfilled tasks, thus improving the performance of the overall scheduler.

Furthermore, graph neural networks (GNNs) can successfully process graph-based data. Their iterative information aggregation from neighbors can be considered as a special form of Laplacian smoothing. However, many GNNs fall into an over-smoothing problem—i.e., the learned representations become indistinguishable when the model goes deeper. In other words, the GNNs may be unable to explore the global graph structure effectively. To address this problem, Yan et al. [4] presented a graph neural network model—in particular, *graph dilated networks with rejection mechanism* (*GraphDRej*). The dilated graph convolution kernel captures the high-level graph structure, and the rejection mechanism addresses the over-smoothing problem. The experimental results show that GraphDRej led to a higher accuracy compared to several existing GNN-based methods.

Skyline queries are popular in many big data analysis and visualization applications. The skyline query conducts computations using a domination test. It tests if a data point has a value that is better than the other values in at least one dimension, and is not worse than others in the remaining dimensions. Such a skyline query can be used in constructing efficient queries based on data from various fields. However, with an increase in the number of dimensions or the volume of data, the overall performance of naïve skyline queries may degrade due to the higher comparison cost among data. Although attempts have been made to solve this problem using index structures, these index structures are heavily influenced by dimensionality and data volume. Choi et al. [5] presented a *hash*

index-based skyline query processing (*HI-Sky*) to overcome the aforementioned shortcomings. When computing the skyline, HI-Sky manages data through the hash index and improves its performance by reducing unnecessary dominance tests and eliminating unnecessary data comparisons. The theoretical and experimental results show that HI-Sky improves skyline query performance when compared with prevalent methods.

It is natural for humans to use hand gestures to interact with computers. For example, people usually use gestures to express meanings and their thoughts in daily conversations, as well as for numerous applications in various fields. As technology advances, hand gesture recognition is becoming popular in human–computer interactions (HCI). Tran et al. [6] presented a system—which uses an RGB-D camera (i.e., a red–green–blue color depth-sensing camera) and a three-dimensional convolution neural network (3DCNN)—for real-time hand gesture recognition and identification. It extracts fingertip locations and recognizes gestures in real time. To demonstrate the accuracy and robustness of the system interface, experiments were conducted to evaluate hand gesture recognition across various gestures. The results show that the system led to a high accuracy in hand gesture recognition, which is promising for gesture-based interfaces to interact between humans and computers by hand in the future.

In the current era of Industry 4.0 (aka Fourth Industrial Revolution), many companies are focusing on securing artificial intelligence (AI) technology by enhancing their competitiveness through machine learning (which is the core technology of AI) and allowing computers to acquire high-quality data through self-learning. To enhance their competitiveness, many companies are securing high-quality big data. As such, the volume of digital information has rapidly increased throughout the world. The presentation of the value quality index of each data attribute has become meaningful because users may be interested in evaluating data quality from their viewpoint to determine whether the data are suitable for their use. Hence, Jang et al. [7] presented a study on data profiling (for both structured and unstructured data), with a focus on both the *attribute value quality index* (*AVQI*) and the *structured data value quality index* (*SDVQI*).

2.2. Big Data Applications

One of the application areas for big data analysis and visualization is in marine wildlife. Marine resources are generally valuable assets that are protected from *illegal, unreported, and unregulated* (*IUU*) *fishing and overfishing*. To detect IUU fishing and overfishing, Kim and Lee [8] presented a convolutional neural network (CNN)-based method to identify fishing equipment—for the fishing ships in operation—from an automatic identification system (AIS)-based trajectory data of fishing ships. This deep learning-based fishing gear-type identification method identifies six fishing gear type groups (namely, drift gill net, longline, purse seine, single trawl, stow net, and traps) from AIS-based ship movement data and environmental data. More specifically, the method preprocesses and handles messages with different messaging interval lengths, contaminated messages, and missing messages for data trajectories. It uses a sliding window-based data slicing technique to capture complicated dynamic patterns in trajectories of fishing gear types and for generating the training data set. Then, the prediction module of this CNN-based method suggests a putative fishing gear type with features that can be extracted from the input trajectory data by the two feature extraction modules (which extract features from fishing ship movement data and from environmental data, respectively). The experimental results of a real-life trajectory data set containing 1380 fishing ships collected over a year show that this CNN-based method led to a higher daily performance index (DPI)—i.e., higher accuracy—than the existing support vector machine (SVM)-based models in identifying the types of fishing gears.

In addition to protecting marine wildlife from IUU fishing and overfishing, it is also important to protect Internet users from harmful online contents. For instance, Song and Kim [9] presented a multimodal stacking scheme to detect harmful online pornographic content. The stacking scheme uses a bi-directional recurrent neural network (RNN) with a

16-layered dilated convolutional network called VGG-16 (from Oxford's Visual Geometry Group (VGG)) to implicitly express the signal change patterns over time within each input and to extract the implicative auditory and visual features. It trains an audio classifier by using only the implicative auditory features; it also trains a video classifier by using only the implicative visual features. Moreover, it trains a fusion classifier by using both the auditory and visual features. Afterwards, these three classifiers are stacked—in a serial order of the fusion classifier, video classifier, and audio classifier—in the ensemble. Consequently, the resulting stack reduces the false negative errors scheme for quick online detections. The experimental results show that this stacking scheme led to a higher true positive rate and accuracy, a lower false negative rate, and a shorter detection time than an existing detection scheme.

A third big data analysis and visualization application is career choice prediction. Career choice plays an important role in life planning of university/college students. Traditionally, professional career appraisers made use of questionnaires or diagnoses in quantifying potential factors that influence career choices. However, given the diversity of goals and ideas of each student, this traditional approach can be changing for proper forecast of their career choices. Observing that the behavioral data of students may reflect their career choices, Nie et al. [10] presented a model called ACCBOX (i.e., Approach Cluster Centers Based On eXtreme gradient boosting (XGBoost)) to mine the potential behavior of college students from campus big data and predict their career choices. The experimental results from 13 million behavioral data records of over 4000 students show that ACCBOX outperformed existing prediction models with a higher accuracy, F1 score, precision, and recall.

A fourth big data analysis and visualization application area is agriculture. As onions are produced in different regions of South Korea, Cho et al. [11] presented an agricultural data analysis method—which uses a functional regression model—to predict weights of the field onions during their growth stages. Specifically, in their functional regression model, they first used (a) onion weight on growth stages as a response variable and (b) six environmental factors (namely, average temperature, average ground temperature, rainfall, wind speed, sunshine, and humidity) as explanatory variables. They then defined a least minimum integral squared residual (LMISE) measure in estimating the function regression coefficient, and applied a principal component regression analysis in deriving an estimate to minimize this defined LMISE measure. Experiments—including graphical and correlation analysis, as well as functional regression analysis—were conducted on real-life data collected from farmers in different regions of South Korea. The results reveal that appropriate sunshine and ground temperature are essential. Moreover, low humidity and rainfall, as well as appropriate temperature and wind, promote onion growth.

3. Summary

Big data have become a core technology for providing innovative solutions in numerical applications and services in many fields. This calls for big data analysis and visualization. The former examines these big data to discover information that can be useful for making various business decisions. The latter assists in visualizing new meanings and interpreting big data, and thus helps explore the data and simplify complex big data analytics processes. In this article, we summarized and highlighted some challenges and solutions in big data analysis and visualization, as well as their applications, which are covered in the 11 articles in the current Special Issue on "big data analysis and visualization".

Author Contributions: Conceptualization, K.-H.Y., C.K.L. and A.N.; writing—original draft preparation, C.K.L.; writing—review and editing, K.-H.Y., C.K.L. and A.N.; project administration, K.-H.Y., C.K.L. and A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by (a) Natural Sciences and Engineering Research Council of Canada (NSERC) and (b) University of Manitoba.

Acknowledgments: This Special Issue would not have been possible without the help and effort of many people and organizations. We thank the editors and staff of *Applied Sciences* at MDPI for their support. We also express our thanks to all authors who have contributed to this Special Issue and all reviewers who have provided constructive comments and suggestions to these authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shin, W.; Yoo, K.; Baek, N. Large-scale data computing performance comparisons on SYCL heterogeneous parallel processing layer implementations. *Appl. Sci.* **2020**, *10*, 1656. [CrossRef]
2. Kong, B.; Ryu, G.; Bae, S.; Noh, S.; Yoon, H. An efficient approach to consolidating job schedulers in traditional independent scientific workflows. *Appl. Sci.* **2020**, *10*, 1455. [CrossRef]
3. Yoon, J.; Hong, T.; Park, C.; Noh, S.; Yu, H. Log analysis-based resource and execution time improvement in HPC: A case study. *Appl. Sci.* **2020**, *10*, 2634. [CrossRef]
4. Yan, B.; Wang, C.; Guo, G. Graph dilated network with rejection mechanism. *Appl. Sci.* **2020**, *10*, 2421. [CrossRef]
5. Choi, J.; Hao, F.; Nasridinov, A. HI-sky: Hash index-based skyline query processing. *Appl. Sci.* **2020**, *10*, 1708. [CrossRef]
6. Tran, D.; Ho, N.; Yang, H.; Baek, E.; Kim, S.; Lee, G. Real-time hand gesture spotting and recognition using RGB-D camera and 3D convolutional neural network. *Appl. Sci.* **2020**, *10*, 722. [CrossRef]
7. Jang, W.; Lee, S.; Kim, J.; Gim, G. A study on data profiling: Focusing on attribute value quality index. *Appl. Sci.* **2019**, *9*, 5054. [CrossRef]
8. Kim, K.; Lee, K. Convolutional neural network-based gear type identification from automatic identification system trajectory data. *Appl. Sci.* **2020**, *10*, 4010. [CrossRef]
9. Song, K.; Kim, Y. An enhanced multimodal stacking scheme for online pornographic content detection. *Appl. Sci.* **2020**, *10*, 2943. [CrossRef]
10. Nie, M.; Xiong, Z.; Zhong, R.; Deng, W.; Yang, G. Career choice prediction based on campus big data—Mining the potential behavior of college students. *Appl. Sci.* **2020**, *10*, 2841. [CrossRef]
11. Cho, W.; Na, M.; Park, Y.; Kim, D.; Cho, Y. Prediction of weights during growth stages of onion using agricultural data analysis method. *Appl. Sci.* **2020**, *10*, 2094. [CrossRef]