information

MDPI

Editorial

Best IDEAS: Special Issue of the International Database Engineered Applications Symposium

Peter Z. Revesz [1,2]

[1] School of Computing, College of Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA; peter.revesz@unl.edu

[2] Department of Classics and Religious Studies, College of Arts and Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

1. Introduction

Database engineered applications cover a broad range of topics including various design and maintenance methods, as well as data analytics and data mining algorithms and learning strategies for enterprise, distributed, or federated data stores. The exponentially growing amounts of commercial, governmental, and non-government organizational data provide a continued challenge for many database engineered applications. The collection of papers in this Special Issue makes several fundamental contributions to this research area.

This Special Issue is primarily based on extended versions of selected papers from the 27th International Database Engineered Applications Symposium (IDEAS) held in 2023 in Heraklion, Crete, Greece, as well as selected papers from prior IDEAS conferences. We also invited additional papers on the conference theme, and they also underwent a rigorous review process.

These invited papers included the paper "Fundamental Research Challenges for Distributed Computing Continuum Systems" by Schahram Dustdar, Professor and Head of the Distributed Systems group at the Vienna University of Technology (TU Wien), and his coworkers [1]. Schahram Dustdar was invited to contribute to this Special Issue because he has served as one of the invited speakers at several IDEAS conferences. This paper [1] lays out a bold vision for the future of distributed computing systems.

The other papers in the Special Issue cover a range of topics, as follows.

2. Federated Learning and Learning Instance Selection

In 2010, Revesz and Triplet [2] introduced the concept—though not the term—of federated learning, in which multiple entities collaborate to train a model without sharing the data due to privacy concerns. Revesz and Triplet [2] gave the example of a set of hospitals who may not share information about their cardiology patients because of patient privacy restrictions. Revesz and Triplet [2] proposed that each hospital train its own classification model on their local data, and then they share the classification models instead of the raw data. Revesz and Triplet [2] also presented several classification integration methods based on constraint databases [3]. Bonawitz et al. [4] termed this type of collaborative learning 'federated learning' and applied it to a set of mobile devices training a neural network. Federated learning has become a very active research area since then [5]. It should not be confused with federated databases [6], which also cooperate in answering queries, but not in learning.

Two papers in this Special Issue deal with the topic of federated learning. The paper "Comparative Analysis of Membership Inference Attacks in Federated and Centralized Learning" by Abbasi Tadi et al. [7] describes several methods that can be used to prevent potential attackers from inferring sensitive data by intercepting updates transmitted between training parties and a central server which maintains the common learned model.

check for updates

Citation: Revesz, P.Z. Best IDEAS: Special Issue of the International Database Engineered Applications Symposium. *Information* **2024**, *15*, 713. https://doi.org/10.3390/info15110713

Received: 22 October 2024
Revised: 2 November 2024
Accepted: 4 November 2024
Published: 6 November 2024

The paper “Exploring Federated Learning Tendencies Using a Semantic Keyword Clustering Approach” by Enguix, Carrascosa, and Rincon [8] considers identifying current trends and emerging subareas within a research area. The authors propose an automatic semantic keyword clustering method. They apply their method to the set of federated learning research papers published since 2017 and identify the fastest growing subareas.

The paper “Prototype Selection for Multilabel Instance-Based Learning” by Filippakis, Ougiaroglou, and Evangelidis [9] considers the problem of reducing the size of the training set in the case of multilabel instance-based classification learning. Here, the term “multilabel” means that each instance can belong to several classes. While there are several well-known algorithms for reducing the size of the training set in the case of single-label instance-based classification learning, the multilabel case was an open problem. Filippakis et al. [9] propose several solutions to this open problem.

We would like to point out that the authors of [9] had the highest and the authors of [7] had the second highest ranked paper at the IDEAS 2023 conference, and their journal articles are also excellent contributions to this Special Issue.

3. Data Analysis and Data Mining

Learning is closely related to data analysis and data mining. In fact, the paper “Convolutional Neural Networks Analysis Reveals Three Possible Sources of Bronze Age Writings between Greece and India” by Daggumati and Revesz [10] included training a set of convolutional neural networks (CNNs) to recognize eight Bronze Age scripts as a first step. In a second step, Daggumati and Revesz passed each script’s signs to each other script’s trained CNN. As each CNN recognized each of the foreign scripts’ signs as a local sign, a table of sign correspondences was found. Two scripts could be identified as being related if their sign correspondence table showed a one-to-one function. Based on that idea, the eight Bronze Age scripts were found to form three groups: (1) Sumerian pictograms, the Indus Valley script, and the proto-Elamite script; (2) Cretan hieroglyphs and Linear B; and (3) the Phoenician, Greek, and Brahmi alphabets. The CNN-based script similarity method of Daggumati and Revesz [10] improves on an earlier computational script similarity method based on feature vectors [11]. A better understanding of script similarities helps in the decipherment of ancient inscriptions [12,13].

The paper “Archaeogenetic Data Mining Supports a Uralic–Minoan Homeland in the Danube Basin” by Revesz [14] applies data mining to the rapidly growing archaeogenetic data. The available archaeogenetic data are often incomplete and therefore more difficult to analyze than regular genetic data. By using some novel data mining algorithms, it was possible to show that the Minoans, who formed the first Bronze Age civilization in Europe, mostly originated from the lower Danube Basin. A better understanding of the origin of the Minoans helps to narrow down the set of languages to be considered as likely cognates with the Minoan language. This could avoid resorting to brute-force methods of cryptanalysis where all possible ancient languages are considered from the Mediterranean and Black Sea areas [15]. The lower Danube Basin is a good candidate for a Proto-Uralic language area in the Neolithic.

4. Temporal Logic and Verification

Linear Temporal Logic over finite traces (LTL_f) can be used to express a set of temporal specifications Φ. Verifying that a system satisfies an LTL_f specification is a computationally difficult task. Therefore, an extended LTL_f (xtLTL_f) is proposed by Bergami, Appleby, and Morgan [16] in the paper “Quickening Data-Aware Conformance Checking through Temporal Algebras”. They describe systems by a set of traces of observed and completed labeled activities expressing one possible run of a process. Verifying that such system descriptions satisfy an xtLTL_f specification can be efficiently checked if the set of traces are first converted to a columnar data storage [16].

The paper “Streamlining Temporal Formal Verification over Columnar Databases” by Bergami [17] takes this idea further by considering the following four new operators:

ChainResponse(A,B), ChainPrecedence(A,B), AltResponse(A,B), and AltPrecedence(A,B). For example, ChainResponse(A,B) is true if the activation of A is immediately followed by the target B. Bergami [17] shows that expressions including these operators can also be checked efficiently if the traces are converted to columnar data storage.

5. Prediction, Detection and Imputation

The paper "Enhancing Flight Delay Predictions Using Network Centrality Measures" by Ajayi et al. [18] aims at improving the accuracy of predicting airplane flight delays. The authors improve the prediction accuracy by introducing a novel method based on network centrality measures that are sensitive to the structure of the flight network.

The paper "Correction of Threshold Determination in Rapid-Guessing Behaviour Detection" by Alfian et al. [19] concerns detecting whether a student is only guessing answers on a multiple-choice test. The traditional method of detecting whether a student is guessing is based on setting a fixed threshold response time, say K seconds, where K is a small number like 3 or 5 depending on the overall difficulty level of the test. If the student's response time is less than K seconds, then the student is assumed to have guessed the answer. Alfian et al. [19] criticize this K-seconds approach because the difficulty of the questions could vary on a test. They show that the accuracy of detecting guessing is improved when the threshold is a variable depending on the difficulty level of the questions.

Greco, Molinaro, and Trubitsyna already considered the challenging topic of incomplete databases in an earlier IDEAS paper [20]. Now, Shahbazian and Trubitsyna [21] address the issue again in the paper "DEGAIN: Generative-Adversarial-Network-Based Missing Data Imputation". They propose handling missing data in incomplete databases by means of data imputation, where the missing values are estimated based on the rest of the data. Generative Adversarial Imputation Nets (GAINs) can be used to generate synthetic data that are like the real data [22]. The main idea is to have a generator of fake data and a discriminator that tries to tell whether a datum is real or fake. However, Shahbazian and Trubitsyna [21] argue that there is a strong correlation among real data. Hence, a deconvolution process is needed to reduce these correlations, and then the generator and discriminator network will work more effectively. Combining deconvolution and GAIN gives rise to the name DEGAIN. We hope that DEGAIN will gain widespread acceptance in data imputation in the future.

Acknowledgments: I would like to thank the many reviewers of the papers submitted to this Special Issue. Their detailed comments and thoughtful recommendations regarding acceptance or rejection helped to maintain a high standard for this Special Issue. I also would like to thank the Section Managing Editor at MDPI for this Special Issue, for her outstanding help in every aspect of the organization work, including her help in finding some reviewers. I thank Haridimos Kondylakis for serving as editor of [14] and Xin Ning for serving as editor of [10] to avoid any conflicts of interest in the review process. Finally, I express my great appreciation for the many talented contributors to this Special Issue, as well as the authors and editors that contributed to the IDEAS 2023 conference and earlier IDEAS conferences. It was great working with the authors throughout the publication process and learning about their exciting results. I wish all of them much success in their future research.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Casamayor Pujol, V.; Morichetta, A.; Murturi, I.; Donta, P.K.; Dustdar, S. Fundamental Research Challenges for Distributed Computing Continuum Systems. *Information* **2023**, *14*, 198. [CrossRef]
2. Revesz, P.Z.; Triplet, T. Classification integration and reclassification using constraint databases. *Artif. Intell. Med.* **2010**, *49*, 79–91. [CrossRef] [PubMed]
3. Kanellakis, P.C.; Kuper, G.M.; Revesz, P.Z. Constraint query languages. *J. Comput. Syst. Sci.* **1995**, *51*, 26–52. [CrossRef]
4. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, New York, NY, USA, 30 October–3 November 2017; pp. 1175–1191.

5. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
6. Sheth, A.P.; Larson, J.A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Comput. Surv.* **1990**, *22*, 183–236. [CrossRef]
7. Abbasi Tadi, A.; Dayal, S.; Alhadidi, A.; Mohammed, N. Comparative Analysis of Membership Inference Attacks in Federated and Centralized Learning. *Information* **2023**, *14*, 620. [CrossRef]
8. Enguix, F.; Carrascosa, C.; Rincon, J. Exploring Federated Learning Tendencies Using a Semantic Keyword Clustering Approach. *Information* **2024**, *15*, 379. [CrossRef]
9. Filippakis, P.; Ougiaroglou, S.; Evangelidis, G. Prototype Selection for Multilabel Instance-Based Learning. *Information* **2023**, *14*, 572. [CrossRef]
10. Daggumati, S.; Revesz, P.Z. Convolutional Neural Networks Analysis Reveals Three Possible Sources of Bronze Age Writings between Greece and India. *Information* **2023**, *14*, 227. [CrossRef]
11. Revesz, P.Z. Establishing the West-Ugric Language Family with Minoan, Hattic and Hungarian by a Decipherment of Linear A. *WSEAS Trans. Inf. Sci. Appl.* **2017**, *14*, 306–335.
12. Revesz, P.Z. A Translation of the Arkalochori Axe and the Malia Altar Stone. *WSEAS Trans. Inf. Sci. Appl.* **2017**, *14*, 124–133.
13. Hughes-Castleberry, K. Could AI Language Models Like ChatGPT Unlock Mysterious Ancient Texts? *Discover Magazine*. 11 April 2023. Available online: https://www.discovermagazine.com/technology/could-ai-language-models-like-chatgpt-unlock-mysterious-ancient-texts (accessed on 15 April 2023).
14. Revesz, P.Z. Archaeogenetic Data Mining Supports a Uralic–Minoan Homeland in the Danube Basin. *Information* **2024**, *15*, 646. [CrossRef]
15. Nepal, A.; Perono Cacciafoco, F. Minoan Cryptanalysis: Computational Approaches to Deciphering Linear A and Assessing its Connections with Language Families from the Mediterranean and the Black Sea Areas. *Information* **2024**, *15*, 73. [CrossRef]
16. Bergami, G.; Appleby, S.; Morgan, G. Quickening Data-Aware Conformance Checking through Temporal Algebras. *Information* **2023**, *14*, 173. [CrossRef]
17. Bergami, G. Streamlining Temporal Formal Verification over Columnar Databases. *Information* **2024**, *15*, 34. [CrossRef]
18. Ajayi, J.; Xu, Y.; Li, L.; Wang, K. Enhancing Flight Delay Predictions Using Network Centrality Measures. *Information* **2024**, *15*, 559. [CrossRef]
19. Alfian, M.; Yuhana, U.L.; Pardede, E.; Bimantoro, A.N.P. Correction of Threshold Determination in Rapid-Guessing Behaviour Detection. *Information* **2023**, *14*, 422. [CrossRef]
20. Greco, S.; Molinaro, C.; Trubitsyna, I. Algorithms for computing approximate certain answers over incomplete databases. In Proceedings of the 22nd International Database Engineering and Applications Symposium, Villa San Giovanni, Italy, 18–20 June 2018; ACM Press: New York, NY, USA, 2018; pp. 1–4.
21. Shahbazian, R.; Trubitsyna, I. DEGAIN: Generative-Adversarial-Network-Based Missing Data Imputation. *Information* **2022**, *13*, 575. [CrossRef]
22. Yoon, J.; Jordon, J.; Schaar, M. GAIN: Missing data imputation using generative adversarial nets. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5689–5698.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.