Topics for Undergraduate Theses & Accelerated Master's Programs

Edurdo Cunha de Almeida
Departamento de Informática, Universidade Federal do Paraná, Curitiba, Brazil
eduardo@inf.ufpr.br, eduardo.almeida@ufpr.br

Abstract

In this document, we present proposed topics for undergraduate final projects (TCC1 & TCC2) and master's dissertations, intended for supervision within the Graduate Program in Informatics at UFPR. Master's projects are expected to be completed within one year after admission to the program.

I. Data Profiling

Title: *Discovery of Business Rules.*

Brief description: Data models propose constraints that restrict the values accepted in a database. There are various types of constraints, ranging from attribute domain restrictions to complex rules involving multiple predicates. However, providing high-quality rules is a challenging task with exponential cost in the worst case scenario both in discovery time and in search space [1]. This complexity motivated extensive research into the development of automated solutions [1]–[10], referred to as data constraint discovery, or informally as business rule discovery. In this line of work, we will study various rule discovery algorithms, predicate processing operations, computational costs, and quality measures for the discovered rules.

Title: *Discovery of Data Violations.*

Brief description: The discovery of data violations is a critical task in the data cleaning process. Dirty data can impact tasks such as artificial intelligence, database design, data compression, and query processing, among others [2]. In general, constraint-based data cleaning involves two main steps: error detection and error correction [11]. In this line of work, we will focus on error detection through the identification of data constraint violations. Several studies have used relational DBMSs to detect such violations [12]–[15], translating business rules into SQL statements. In this line of work, we will study algorithms for discovering data violations, algorithms and structures for representing violations (such as graphs and prefix trees), and the process of translating these violations into SQL queries.

II. Database-Hardware Co-Design

Title: *FPGA Data Processing*.

Brief description: *Field Programmable Gate Arrays* (FPGAs) are programmable integrated circuits that enable high parallelism in practical data processing applications. Recent literature presents circuit designs for evaluating SQL predicates involving arithmetic and logical operations [16]–[19]. In this line of work, we will study data processing techniques and SQL operations accelerated by FPGAs with the goal of eliminating intermediate data structures, thereby reducing memory requirements and potentially achieving performance improvements by orders of magnitude [20].

References

[1] X. Chu, I. F. Ilyas, and P. Papotti, "Discovering denial constraints," *Proceedings of the VLDB Endowment*, vol. 6, no. 13, pp. 1498–1509, 2013.
[2] A. Martin, E. C. de Almeida, O. Romero, and A. Queralt, "How and why false denial constraints are discovered," *Proceedings of the VLDB Endowment*, vol. 18, no. 10, pp. 3477 – 3489, 2025.
[3] T. Bleifuß, S. Kruse, and F. Naumann, "Efficient denial constraint discovery with hydra," *Proceedings of the VLDB Endowment*, vol. 11, no. 3, pp. 311–323, 2017.
[4] E. H. Pena and E. C. de Almeida, "Bfastdc: A bitwise algorithm for mining denial constraints," in *Database and Expert Systems Applications: 29th International Conference, DEXA 2018, Regensburg, Germany, September 3–6, 2018, Proceedings, Part I 29*. Springer, 2018, pp. 53–68.
[5] E. H. Pena, E. C. De Almeida, and F. Naumann, "Discovery of approximate (and exact) denial constraints," *Proceedings of the VLDB Endowment*, vol. 13, no. 3, pp. 266–278, 2019.
[6] E. Livshits, A. Heidari, I. F. Ilyas, and B. Kimelfeld, "Approximate denial constraints," *Proc. VLDB Endow.*, vol. 13, no. 10, pp. 1682–1695, 2020. [Online]. Available: http://www.vldb.org/pvldb/vol13/p1682-livshits.pdf
[7] R. Xiao, Z. Tan, H. Wang, and S. Ma, "Fast approximate denial constraint discovery," *Proceedings of the VLDB Endowment*, vol. 16, no. 2, pp. 269–281, 2022.
[8] E. H. Pena, F. Porto, and F. Naumann, "Fast algorithms for denial constraint discovery," *Proceedings of the VLDB Endowment*, vol. 16, no. 4, pp. 684–696, 2022.
[9] C. Qian, M. Li, Z. Tan, A. Ran, and S. Ma, "Incremental discovery of denial constraints," *The VLDB Journal*, vol. 32, no. 6, pp. 1289–1313, 2023.
[10] L. Bian, W. Yang, J. Xu, and Z. Tan, "Discovering denial constraints based on deep reinforcement learning," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 120–129.

[11] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data cleaning: Overview and emerging challenges," 2016, p. 2201–2206. [Online]. Available: https://doi.org/10.1145/2882903.2912574
[12] W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, "Conditional functional dependencies for capturing data inconsistencies," vol. 33, no. 2, pp. 6:1–6:48, 2008. [Online]. Available: https://doi.org/10.1145/1366102.1366103
[13] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, "HoloClean: Holistic data repairs with probabilistic inference," vol. 10, no. 11, pp. 1190–1201, 2017.
[14] F. Geerts, G. Mecca, P. Papotti, and D. Santoro, "Cleaning data with llunatic," vol. 29, no. 4, pp. 867–892, 2020. [Online]. Available: https://doi.org/10.1007/s00778-019-00586-5
[15] W. Fan, C. Tian, Y. Wang, and Q. Yin, "Parallel discrepancy detection and incremental detection," vol. 14, no. 8, pp. 1351–1364, 2021.
[16] R. Mueller, J. Teubner, and G. Alonso, "Streams on wires: a query compiler for fpgas," *Proc. VLDB Endow.*, vol. 2, no. 1, p. 229–240, Aug. 2009. [Online]. Available: https://doi.org/10.14778/1687627.1687654
[17] W. Jiang, M. Parvanov, and G. Alonso, "Swiftspatial: Spatial joins on modern hardware," 2023. [Online]. Available: https://arxiv.org/abs/2309.16520
[18] B. Sukhwani, H. Min, M. Thoennes, P. Dube, B. Iyer, B. Brezzo, D. Dillenberger, and S. Asaad, "Database analytics acceleration using fpgas," in *Proceedings of the 21st International Conference on Parallel Architectures and Compilation Techniques*, ser. PACT '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 411–420. [Online]. Available: https://doi.org/10.1145/2370816.2370874
[19] H. Kong, W. Lu, Y. Chen, J. Wu, Y. Zhang, G. Yan, and X. Li, "DOE: database offloading engine for accelerating SQL processing," *Distributed Parallel Databases*, vol. 41, no. 3, pp. 273–297, 2023. [Online]. Available: https://doi.org/10.1007/s10619-023-07427-z
[20] S. L. Marques Filho, "Discovering denial constraints using boolean patterns," in *SIGMOD Companion*, ser. SIGMOD '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 281–283. [Online]. Available: https://doi.org/10.1145/3555041.3589392