Rough Sets in Data Warehousing
Extended Abstract

Dominik Ślęzak, Jakub Wróblewski, Victoria Eastwood, and Piotr Synak

Infobright Inc., Poland & Canada
www.infobright.com,
{slezak,jakubw,victoriae,synak}@infobright.com

The theory of rough sets [15,16], based on the universal framework of information systems, provides a powerful model for representing patterns and dependencies both in databases and in data mining. On the one hand, although there are numerous rough set applications to data mining and knowledge discovery [10,18], the usage of rough sets inside the database engines is still quite an uncharted territory. On the other hand, however, this situation is not so exceptional given that even the most well-known paradigms of machine learning, soft computing, artificial intelligence, and approximate reasoning are still waiting for more recognition in the database research, with huge potential in such areas as, e.g., physical data model tuning or adaptive query optimization [2,3].

Rough set-based algorithms and similar techniques can be applied to improve database performance by employing the automatically discovered dependencies to better deal with query conditions [5,9]. Another idea is to use available information to calculate rough approximations of data needed to resolve queries and to assist the database engine in accessing relevant data [20,24]. In our approach, we partition data onto *rough rows*, each consisting of 64K of original rows. We automatically label rough rows with compact information about their values on particular columns, often involving multi-table cross-relationships. One may say that we create a new information system where objects take the form of rough rows and attributes correspond to various flavors of rough row information. A number of database operations can be fully or partially processed within such a new system, with an access to the original data pieces still available, whenever required on top of rough row information. Such a framework seems to actually fit the paradigms of rough and granular computing [1,17], where calculations on granules are additionally allowed to interact with those on single items.

The above ideas guided us towards implementing the fully functional database product, with interfaces provided via integration with MySQL [13,14] and with internals based on such trends in database research as columnar stores [8,11] and adaptive compression [6,22]. Relying on relatively small, flexible rough row information enabled us to become especially competitive in the field of analytical data warehousing, where users want to analyze terabytes of data in a complex, dynamically changing fashion. We realize though that we should keep comparing ourselves against other strategies of using data about data [4,12] and redesigning various dependency/pattern/metadata/index structures originally defined over single rows to let them work at our rough row level [7,19]. In particular, searching

C.-C. Chan et al. (Eds.): RSCTC 2008, LNAI 5306, pp. 505–507, 2008.

for most efficient though reasonably compact types of rough row information can be interpreted by means of feature extraction and selection [21,23], which will additionally inspire us to refer to the rough set methods in future.

References

1. Bargiela, A., Pedrycz, W.: Granular Computing: An Introduction. Springer, Heidelberg (2003)
2. Chaudhuri, S., Narasayya, V.R.: Self-Tuning Database Systems: A Decade of Progress. In: VLDB 2007, pp. 3–14 (2007)
3. Deshpande, A., Ives, Z.G., Raman, V.: Adaptive Query Processing. Foundations and Trends in Databases 1(1), 1–140 (2007)
4. Grondin, R., Fadeitchev, E., Zarouba, V.: Searchable archive. US Patent 7, 243, 110 (2007)
5. Haas, P.J., Hueske, F., Markl, V.: Detecting Attribute Dependencies from Query Feedback. In: VLDB 2007, pp. 830–841 (2007)
6. Holloway, A.L., Raman, V., Swart, G., DeWitt, D.J.: How to barter bits for chronons: compression and bandwidth trade offs for database scans. In: SIGMOD 2007, pp. 389–400 (2007)
7. Ioannidis, Y.E.: The History of Histograms (abridged). In: VLDB 2003, pp. 19–30 (2003)
8. Kersten, M.L.: The Database Architecture Jigsaw Puzzle. In: ICDE 2008, pp. 3–4 (2008)
9. Kerdprasop, N., Kerdprasop, K.: Semantic Knowledge Integration to Support Inductive Query Optimization. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2007. LNCS, vol. 4654, pp. 157–169. Springer, Heidelberg (2007)
10. Lin, T.Y., Cercone, N. (eds.): Rough Sets and Data Mining. Kluwer, Dordrecht (1996)
11. MacNicol, R., French, B.: Sybase IQ multiplex - designed for analytics. In: VLDB 2004, pp. 1227–1230 (2004)
12. Metzger, J.K., Zane, B.M., Hinshaw, F.D.: Limiting scans of loosely ordered and/or grouped relations using nearly ordered maps. US Patent 6 973, 452 (2005)
13. MySQL 5.1 Reference Manual: Storage Engines, `dev.mysql.com/doc/refman/5.1/en/storage-engines.html`
14. MySQL Business White Papers: Enterprise Data Warehousing with MySQL, `www.scribd.com/doc/3003152/Enterprise-Data-Warehousing-with-MySQL`
15. Pawlak, Z.: Rough sets: Theoretical aspects of reasoning about data. Kluwer, Dordrecht (1991)
16. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Inf. Sci. 177(1), 3–27 (2007)
17. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. Wiley, Chichester (2008)
18. Polkowski, L., Skowron, A. (eds.): Rough Sets in Knowledge Discovery. Parts 1 & 2. Physica-Verlag (1998)
19. Ślęzak, D.: Searching for dynamic reducts in inconsistent decision table. In: IPMU 1998, vol. 2, pp. 1362–1369 (1998)
20. Ślęzak, D., Wróblewski, J., Eastwood, V., Synak, P.: Brighthouse: An Analytic Data Warehouse for Ad-hoc Queries. In: VLDB 2008 (2008)
21. Świniarski, R.W., Skowron, A.: Rough set methods in feature selection and recognition. Pattern Recognition Letters 24(6), 833–849 (2003)

22. Wojnarski, M., Apanowicz, C., Eastwood, V., Ślęzak, D., Synak, P., Wojna, A., Wróblewski, J.: Method and System for Data Compression in a Relational Database. US Patent Application, 2008/0071818 A1 (2008)
23. Wróblewski, J.: Analyzing relational databases using rough set based methods. In: IPMU 2000, vol. 1, pp. 256–262 (2000)
24. Wróblewski, J., Apanowicz, C., Eastwood, V., Ślęzak, D., Synak, P., Wojna, A., Wojnarski, M.: Method and System for Storing, Organizing and Processing Data in a Relational Database. US Patent Application, 2008/0071748 A1 (2008)