Proceedings 3rd European Conf. on Principles of Data Mining and Knowledge Discovery PKDD'99,
Prague (CZ), 15-18 September 1999, Springer LNAI 1704, pp. 382-383.

Query Languages for Knowledge Discovery in Databases

Jean-François Boulicaut

Institut National des Sciences Appliquées de Lyon
LISI Bâtiment 501
F-69621 Villeurbanne cedex, France
mailto:Jean-Francois.Boulicaut@insa-lyon.fr
http://www.insa-lyon.fr/People/LISI/jfboulic

Abstract

Discovering knowledge from data appears as a complex iterative and interactive process containing many steps: understanding the data, preparing the data set, discovering potentially interesting patterns (mining phase), postprocessing of discovered patterns and finally putting the results in use. Different kinds of patterns might be used and therefore different data mining techniques are needed (e.g., association and episode rules for alarm analysis, clusters and decision trees for sales data analysis, inclusion and functional dependencies for database reverse engineering, etc).

This tutorial adresses the challenge of supporting KDD processes following a querying approach. Following Imielinski and Mannila [6], second generation data mining systems might support the whole process by means of powerful query languages. We propose not only a state of the art in that field but also introduce a research perspective, the so-called inductive database framework [3]. It is out of the scope of our presentation to consider coupling architectures between data mining algorithms and database management systems. Instead, we focuse on user written queries that capture their needs during a KDD process. The popular association rules mining processes is used to illustrate most of the concepts.

The use of query languages to select data for a mining task seems obvious. However, crucial issues like cleaning, sampling, supporting multidimensional data manipulation or on-line analytical processing queries when preparing a data set are still to be discussed. Indeed, available query languages offer a rather poor support for that. Next, mining phases quite often provide a huge amount of extracted patterns though just a few of them can be of practical interest for the end-user. Within the rule mining domain, one copes with that problem by selecting patterns w.r.t. interestingness measures, objective ones (e.g., confidence [1], J-measure [10], conviction [5]) or subjective ones (e.g., templates [2] or more generally queries). Indeed, it is possible to use standard query languages like SQL3 or OQL to query rule databases and perform typical postprocessing like inclusive/exclusive selection of rules or rule cover computations (elimination of "redundancy"). However, more or less specialized query languages have been

proposed like M-SQL [7] or MINE RULE [8]. These languages enable to select the data, to specify mining tasks and to perform some postprocessing as well. The main ideas concerning query evaluation are considered.

This leads us to the concept of inductive database and general-purpose query languages for KDD applications. An inductive database is a database that in addition to data contains intensionnaly defined generalizations about the data. Using the simple formalization from [3], it is possible to define query languages that satisfies the closure property, i.e., the result of a query on an inductive database is, again, an inductive database. The whole discovery process can be then be modelized by means of a sequence of queries (on data, on patterns or linking data to patterns). This gives rise to optimization techniques (compiling scheme) for KDD processes. Among others, we demonstrate the possibilities according to a research proposal for a rule-based language [4].

Implementating inductive databases for various classes of patterns is still an open problem. However, the research about association rule mining has demonstrated that a technology (see e.g., [9]) is now available for such a simple but still important class of patterns.

References

1. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In: Proc.*SIGMOD'93*, pages 207 – 216, May 1993. ACM Press.
2. M. Klemettinen A Knowledge Discovery Methodology for Telecommunications Network Alarm Databases. PhD thesis, Report A-1999-1, University of Helsinki (FIN), January 1999.
3. J.-F. Boulicaut, M. Klemettinen, and H. Mannila. Modeling KDD Processes within the Inductive Database Framework. In: Proc. *DaWak'99*, August 29-September 2nd 1999. Florence (I), Springer-Verlag. To appear.
4. J.-F. Boulicaut, P. Marcel, and C. Rigotti. A Query driven knowledge discovery in multidimensional data. Research Report, INSA Lyon, LISI, July 1999, 15 p., Submitted.
5. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In: Proc. *SIGMOD'97*, pages 255 – 264, 1997. ACM Press.
6. T. Imielinski and H. Mannila. A Database Perspective on Knowledge Discovery. *Communications of the ACM*, 39(11):58 – 64, November 1996.
7. A. Virmani. Second Generation Data Mining: Concetps and Implementation. PhD thesis, Rutgers University (USA), April 1998.
8. R. Meo, G. Psaila, and S. Ceri. A new SQL-like Operator for Mining Association Rules. In: Proc. *VLDB'96*, pages 122 – 133, September 1996. Morgan Kaufmann.
9. R. Ng, L. Lakshmanan, J. Han, and A. Pang. Exploratory Mining and Pruning Optimizations of Constrained Associations Rules. In: Proc. *SIGMOD'98*, pages 13 – 24, 1998. ACM Press.
10. P. Smyth and R. M. Goodman. An Information Theoretic Approach to Rule Induction from Databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4):301 – 316, August 1992.