Lecture Notes in Artificial Intelligence 2682

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Rosa Meo Pier Luca Lanzi
Mika Klemettinen (Eds.)

Database Support for Data Mining Applications

Discovering Knowledge with Inductive Queries

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Rosa Meo
Università degli Studi di Torino
Dipartimento di Informatica
Corso Svizzera, 185, 10149 Torino, Italy
E-mail: meo@di.unito.it

Pier Luca Lanzi
Politecnico di Milano
Dipartimento di Elettronica e Informazione
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
E-mail: lanzi@elet.polimi.it

Mika Klemettinen
Nokia Research Center
P.O.Box 407, Itämerenkatu 11-13, 00045 Nokia Group, Finland
E-mail: mika.klemettinen@nokia.com

Library of Congress Control Number: 2004095564

CR Subject Classification (1998): I.2, H.2, H.3

ISSN 0302-9743
ISBN 3-540-22479-3 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11301172 06/3142 5 4 3 2 1 0

Preface

Data mining from traditional relational databases as well as from non-traditional ones such as semi-structured data, Web data and scientific databases such as biological, linguistic and sensor data has recently become a popular way of discovering hidden knowledge. In the context of relational and traditional data, methods such as association rules, chi square rules, ratio rules, implication rules, etc. have been proposed in multiple, varied contexts. In the context of non-traditional data, newer, more experimental yet novel techniques are being proposed. There is an agreement among the researchers across communities that data mining is a key ingredient for success in their respective areas of research and development. Consequently, interest in developing new techniques for data mining has peaked and a tremendous stride is being made to answer interesting and fundamental questions in various disciplines using data mining.

In the past, researchers mainly focused on algorithmic issues in data mining and placed much emphasis on scalability. Recently, the focus has shifted towards a more declarative way of answering questions using data mining that has given rise to the concept of mining queries.

Data mining has recently been applied with success to discovering hidden knowledge from relational databases. Methods such as association rules, chi square rules, ratio rules, implication rules, etc. have been proposed in several and very different contexts. To cite just the most frequent and famous ones: the market basket analysis, failures in telecommunication networks, text analysis for information retrieval, Web content mining, Web usage, log analysis, graph mining, information security and privacy, and finally analysis of objects traversal by queries in distributed information systems.

From these widespread and various application domains it results that data mining rules constitute a successful and intuitive descriptive paradigm able to offer complementary choices in rule induction. Other than inductive and abductive logic programming, research into data mining from knowledge bases has been almost non-existent, because contemporary methods place the emphasis on the scalability and efficiency of algorithmic solutions, whose inherent procedurality is difficult to cast into the declarativity of knowledge base systems.

In particular, researchers convincingly argue that the ability to declaratively mine and analyze relational databases for decision support is a critical requirement for the success of the acclaimed data mining technology. Indeed, DBMSs constitute today one of the most advanced and sophisticated achievements that applied computer science has made in the past years. Unfortunately, almost all the most powerful DBMSs we have today have been developed with a focus on On-Line Transaction-Processing tasks. Instead, database technology for On-Line Analytical-Processing tasks, such as data mining, is more recent and in need of further research.

Although there have been several encouraging attempts at developing methods for data mining using SQL, simplicity and efficiency still remain significant prerequisites for further development. It is well known that today database technology is mature enough: popular DBMSs, such as Oracle, DB2 and SQL-Server, provide interfaces, services, packages and APIs that embed data mining algorithms for classification, clustering, association rules extraction and temporal sequences, such that they are directly available to programmers and ready to be called by applications.

Therefore, it is envisioned that we should be able now to mine relational databases for interesting rules directly from database query languages, without any data restructuring or preprocessing steps. Hence no additional machineries with respect to database languages would be necessary. This vision entails that the optimization issues should be addressed at the system level for which we have now a significant body of research, while the analyst could concentrate better on the declarative and conceptual level, in which the difficult task of interpretation of the extracted knowledge occurs. Therefore, it is now time to develop declarative paradigms for data mining so that these developments can be exploited at the lower and system level, for query optimization.

With this aim we planned this book on "Data Mining" with an emphasis on approaches that exploit the available database technology, declarative data mining, intelligent querying and associated issues such as optimization, indexing, query processing, languages and constraints. Attention is also paid to solution of data preprocessing problems, such as data cleaning, discretization and sampling, developed using database tools and declarative approaches, etc.

Most of this book resulted also as a consequence of the work we conducted during the development of the *cInQ* project (**c**onsortium on discovering knowledge with **In**ductive **Q**ueries) an EU funded project (IST 2000-26469) aiming at developing database technology for leveraging decision support systems by means of query languages and inductive approaches to knowledge extraction from databases. It presents new and invited contributions, plus the best papers, extensively revised and enlarged, presented during workshops on the topics of database technology, data mining and inductive databases at international conferences such as EDBT and PKDD/ECML, in 2002.

May 2004

Rosa Meo
Pier Luca Lanzi
Mika Klemettinen

Volume Organization

This volume is organized in two main sections. The former focuses on *Database Languages and Query Execution*, while the latter focuses on methodologies, techniques and new approaches that provide *Support for Knowledge Discovery Process*. Here, we briefly overview each contribution.

Database Languages and Query Execution

The first contribution is *Inductive Databases and Multiple Uses of Frequent Itemsets: The cInQ Approach* which presents the main contributions of theoretical and applied nature, in the field of inductive databases obtained in the `cInQ` project.

In *Query Languages Supporting Descriptive Rule Mining: A Comparative Study* we provide a comparison of features of available relational query languages for data mining, such as `DMQL`, `MSQL`, `MINE RULE`, and standardization efforts for coupling database technology and data mining systems, such as `OLEDB-DM` and `PMML`.

Declarative Data Mining Using SQL-3 shows a new approach, compared to existing SQL approaches, to mine association rules from an object-relational database: it uses a recursive join in SQL-3 that allows no restructuring or preprocessing of the data. It proposes a new `mine by` SQL-3 operator for capturing the functionality of the proposed approach.

Towards a Logic Query Language for Data Mining presents a logic database language with elementary data mining mechanisms, such as user-defined aggregates that provide a model, powerful and general as well, of the relevant aspects and tasks of knowledge discovery.

Data Mining Query Language for Knowledge Discovery in a Geographical Information System presents *SDMOQL* a spatial data mining query language for knowledge extraction from GIS. The language supports the extraction of classification rules and association rules, the use of background models, various interestingness measures and the visualization.

Towards Query Evaluation in Inductive Databases Using Version Spaces studies inductive queries. These ones specify constraints that should be satisfied by the data mining patterns in which the user is interested. This work investigates the properties of solution spaces of queries with monotonic and anti-monotonic constraints and their boolean combinations.

The GUHA Method, Data Preprocessing and Mining surveys the basic principles and foundations of the *GUHA* method, the available systems and related works. This method originated in the Czechoslovak Academy of Sciences of Prague in

the mid 1960s with strong logical and statistical foundations. Its main principle is to let the computer generate and evaluate all the hypotheses that may be interesting given the available data and the domain problem. This work discusses also the relationships between the GUHA method and relational data mining and discovery science.

Constraint Based Mining of First Order Sequences in SeqLog presents a logical language, *SeqLog*, for mining and querying sequential data and databases. This language is used as a representation language for an inductive database system. In this system, variants of level-wise algorithms for computing the version space of the solutions are proposed and experimented in the user-modeling domain.

Support for Knowledge Discovery Process

Interactivity, Scalability and Resource Control for Efficient KDD Support in DBMS proposes a new approach for combining preprocessing and data mining operators in a KDD-aware implementation algebra. In this way data mining operators can be integrated smoothly into a database system, thus allowing interactivity, scalability and resource control. This framework is based on the extensive use of pipelining and is built upon an extended version of a specialized database index.

Frequent Itemset Discovery with SQL Using Universal Quantification investigates the integration of data analysis functionalities into two basic components of a database management system: query execution and optimization. It employs universal and existential quantifications in queries and a vertical layout to ease the set containment operations needed for frequent itemsets discovery.

Deducing Bounds on the Support of Itemsets provides a complete set of rules for deducing tight bounds on the support of an itemset if the support of all its subsets are known. These bounds can be used by the data mining system to choose the best access path to data and provide a better representation of the collection of frequent itemsets.

Model-Independent Bounding of the Supports of Boolean Formulae in Binary Data considers frequencies of arbitrary boolean formulas, a new class of aggregates: the summaries. These ones are computed for descriptive purposes on a sparse binary data set. This work considers the problem of finding tight upper bounds on these frequencies and gives a general formulation of the problem with a linear programming solution.

Condensed Representations for Sets of Mining Queries proposes a general framework for condensed representations of sets of mining queries, defined by monotonic and anti-monotonic selection predicates. This work proves important for inductive and database systems for data mining since it deals with *sets* of queries, whereas previous work in maximal, closed and condensed representations treated so far the representation of a *single* query only.

One-Sided Instance-Based Boundary Sets introduces a family of version-space representations that are important for their applicability to inductive databases. They correspond to the task of concept learning from a database of examples when this database is updated. One-sided instance-based boundary sets are shown to be correctly and efficiently computable.

Domain Structures in Filtering Irrelevant Frequent Patterns introduces a notion of domain constraints, based on distance measures and in terms of domain structure and concept taxonomies. Domain structures are useful in the analysis of communications networks and complex systems. Indeed they allow irrelevant combinations of events that reflect the simultaneous information of independent processes in the same database to be pruned.

Integrity Constraints over Association Rules investigates the notion of integrity constraints in inductive databases. This concept is useful in detecting inconsistencies in the results of common data mining tasks. This work proposes a form of integrity constraints called *association map constraints* that specifies the allowed variations in confidence and support of association rules.

Acknowledgments

For reviewing the contributions included in this volume we relied on a group of internationally well-known experts who we wish to thank here.

Roberto Bayardo	IBM Almaden Research Center, USA
Elena Baralis	Politecnico di Torino, Italy
Christian Böhm	University for Health Informatics and Technology, Austria
Saso Dzeroski	Jozef Stefan Institute, Slovenia
Kris Koperski	Insightful Corporation, USA
Stefan Kramer	Technische Universität München, Germany
Dominique Laurent	Université F. Rabelais, Tours, France
Giuseppe Manco	ICAR-CNR, Italy
Stefano Paraboschi	Università di Bergamo, Italy
Jian Pei	The State University of New York, USA
Giuseppe Psaila	Università di Bergamo, Italy
Lorenza Saitta	Università del Piemonte Orientale, Italy
Maria Luisa Sapino	Università di Torino, Italy
Hannu Toivonen	University of Helsinki, Finland
Jiong Yang	University of Illinois at Urbana Champaign, USA
Mohammed Zaki	Rensselaer Polytechnic Institute, USA

Table of Contents

I Database Languages and Query Execution

II Support for KDD-Process