Report on the International Workshop on Pattern Representation and Management (PaRMa'04)

Yannis Theodoridis[1], Panos Vassiliadis[2]

[1] Univ. of Piraeus, and Computer Technology Institute, Hellas, ytheod@unipi.gr

[2] Univ. of Ioannina, Hellas, pvassil@cs.uoi.gr

1 Introduction

The increasing ability to quickly collect and cheaply store large volumes of data, and the need for extracting concise information to be efficiently manipulated and intuitively analyzed, are posing new requirements for Database Management Systems (DBMS) in both industrial and scientific applications. A common approach to deal with huge data volumes is to reduce the available information to knowledge artifacts (i.e., clusters, rules, etc.), hereafter called *patterns*, through data processing methods (pattern recognition, data mining, knowledge extraction). Patterns reduce the number and size of the original information to manageable size while preserving as much as possible its hidden / interesting content. In order to efficiently and effectively deal with patterns, academic groups and industrial consortiums have recently devoted efforts towards modeling, storage, retrieval, analysis and manipulation of patterns with results mainly in the areas of *Inductive Databases* and *Pattern Base Management Systems* (PBMS).

This report focuses on the International Workshop on Pattern Representation and Management (PaRMa) held in conjunction with the EDBT'04 Conference in Heraklion, Crete, Hellas, on March 18th, 2004. A summary of the papers and the discussions held during the workshop is given.

The main theme for the workshop was "From Data to Patterns". In response to the call for papers, 15 papers were submitted. The Program Committee selected 9 papers based on their scientific contribution, novelty and relevance to the workshop. Due to its obvious database-centric perspective, three papers on modelling and pattern manipulation languages were accepted. A second area of interest involves the management of patterns with a particular focus on pattern similarity and prediction. Finally, following a long tradition in the area of data mining, the workshop program has included three papers on systems and algorithms for pattern extraction.

The electronic edition of the workshop proceedings is available from the CEUR series: http://ceur-ws.org/Vol-96/

2 Languages for patterns

Languages for defining and manipulating patterns are one of the most important issues in Pattern Base (PB) management. Three papers were presented on this topic, all attempting to introduce general approaches for pattern management.

Bertino et al. [2] propose a Pattern Manipulation Language (PML) and a Pattern Query Language (PQL). The PANDA model for pattern representation (based on the proposal of [8]) is adopted. In this approach, patterns are *instances-of* pattern types and *members-of* pattern classes. The model is generic enough to represent any type of pattern by means of a *structure* (e.g., an association rule has a head and a body), *underlying data* that it represents, a *formula* that annotates it with explicit semantics, and statistical *measures* (e.g., confidence and support). The concept of *mining function* (to specify the mining algorithm to be used for extraction) is also introduced. Then, the authors define PML and PQL. PML operations over patterns or classes include insertion, deletion and update. PQL is defined over classes and is closed. PQL operations include projection, selection, identity, shallow/deep value equality, similarity, drill-down and roll-up, decomposition, join, drill-through and covering (with the two last being *cross-over* operators for relating patterns with raw data).

Bouros et al. [3] propose *PatManQL*, a language to manipulate patterns and data in hierarchical catalogs. The paper presents modeling constructs to represent hierarchical catalogs, based on the concept of *Tree-Structured Relations* (TSRs). Then, the PatManQL language is presented, involving operators to manipulate path-like patterns (select, project, Cartesian product, union, intersection, difference over TSRs). Finally, the development of a system implementing the above operators is also given. The prototype includes an interpreter, a query execution engine, a storage mechanism (using XML files and MySQL relational tables) and a graphic result interface.

Rizzi [9] discusses the conceptual modeling of pattern-bases, based on UML. Using [8], the author addresses the issue of how can UML be used and extended to model PBs from the static / functional / dynamic points of view. *Static modeling* describes PBs from a structural point of view, with particular emphasis on expressing complex relationships between patterns. *Functional modeling* represents the processes used to establish the relationship between data and patterns (DFD, UML use case diagrams,

UML activity diagrams). *Dynamic modeling* describes how patterns and raw data are kept in synchronization (state charts, UML sequence diagrams). The overall conclusion is that a broad range of issues related to PBs can be expressively represented without giving up the syntax and semantics of UML.

3 Pattern similarity

Similarity is one of the key issues in pattern representation and management since a number of procedures and tuning mechanisms, including indexing, query optimization, update and synchronization, assume a formal definition of pattern similarity measure. Two papers were presented in this category.

Ntoutsi [6] discusses similarity issues in data and pattern spaces motivated by data mining and PBMS domains. Formally, the measurement of (dis-) similarity is mapped to graph assignment problem, which, although well-researched, is of high complexity (Hungarian method is $O(n^3)$). Experiments on data (100 transactions X 10 items per transaction) and the corresponding pattern sets were shown, uncovering the behavior of similarity measures as well as how similarity in the data space is related to similarity in the pattern space.

Bartolini et al. [1] propose a framework for the comparison of complex patterns. This framework is based on the PANDA logical model for pattern representation [8]. First, the similarity between simple patterns is defined as an aggregation formula between structure and measure components. Then, the similarity between complex patterns (based on simple ones) is defined as the aggregation of similarities between simple components. A prototype was also implemented defining foundation classes that guarantee the generality of the proposed framework. Case studies included a comparison of clustering schemas and a comparison of web sites (reduced to similarity between graphs).

4 Finding 'interesting' patterns

Although the efficient extraction of patterns has been studied extensively in the data mining literature, the detection of interesting patterns is another 'hot' task in the research agenda. Two papers were presented on this topic.

Keim and Wawryniuk [5] address the issue of extracting patterns from data which have categorical and numerical attributes (e.g., census data). After reviewing related work (distance- and mean- based association rules), the authors propose an algorithm that identifies *interesting itemsets*, where 'interestingness' means that the itemset has an impact on the numerical attribute, i.e., distribution different from overall distribution.

Heilmann et al. [4] tackle the problem of finding interesting spatial patterns in network data by exploiting ideas from the visualization and geo-spatial visualization areas. Visualization not only offers a high degree of user satisfaction and confidence but also yields better results. In fact, there exist correlations between data that cannot be expressed easily with some formal language whereas they are straightforward through a visual representation. The authors also presented some example for the effective visualization of network data in an important area of application, the analysis of e-mail traffic.

5 Systems - Applications

The study of patterns in different application domains reflects the need for efficient pattern representation and management. Two papers were presented in this category.

Pelekis et al. [7] propose *Fuzzy-Miner*, a fuzzy system for solving classification problems. Fuzzy-Miner consists of a fuzzification interface (that converts crisp input into fuzzy singletons), a knowledge base (that consists of a data base and a rule base with a set of if-then rules), a decision making logic module (which employs fuzzy if-then rules from the rule base to infer the output by a fuzzy reasoning system) and, finally, a defuzzification interface (that defuzzifies the output). The quality of the system, in terms of classification success, was experimentally evaluated using a real dataset (Athens Stock Exchange data).

Robardet et al. [9] exploit pattern databases for gene expression data analysis. In particular, the authors design inductive querying techniques that extract interesting sets of objects (biological situations) and associated sets of attributes (genes). The methodology utilizes a hierarchical classification method to cluster concepts and then the obtained hierarchy of concepts is visualized.

6 Panel Discussion

Following the general motivation for having a workshop on pattern management, a panel took place at the end of the workshop, in order to summarize the state of the art in this novel research field, as well as the main impressions from the workshop and discuss the research agenda of the topic. In this section we will briefly sketch the main highlights of the discussions; a more detailed report is available [11].

The panel was moderated by Panos Vassiliadis with the participation of Barbara Catania, Daniel Keim, Céline Robardet and Yannis Theodoridis. The panelists addressed the following issues:

- What do you think patterns are? Is it possible to devise a generic model for patterns?
- What do you think are the main "queries" and operations over patterns?
- How would you manage patterns?
- How would you prescribe a research agenda for pattern management?

In terms of discussing the nature of patterns, there was a (naturally expected) reference to the definition of [8] for a compact representation of data that is rich in semantics. At the same time, there was also a discussion of the inductive database approach for simple patterns vs. models. The discussion that took place converged towards two main

results: (a) there exist common characteristics in patterns (like structure, relationship to underlying data and so on) and (b) it is quite hard, yet visionary, to come up with a generic model for patterns.

The difficulty in providing a common, generic set of operations also appeared as the outcome of the second topic. During the discussion, it was pointed out that the way people handle patterns now is dependent on the domain of application. Still, one could possibly identify common characteristics, in the sense that in all application domains, people spend quite a lot of time either observing patterns or combining patterns with data in order to deduce knowledge. At the same time, the opinion that it is pointless to ask for *common*, but for *useful* or *important* operations, was also expressed. A vivid discussion followed, over the nature of the query paradigm and the languages that can be adopted for this purpose.

On the issue of how to build a system to manage patterns, the general feeling can be summarized as follows: (a) building such a system from scratch is not worthwhile; (b) on the contrary, Object-Relational technology should be the basis over which such a system should be built; (c) possibly, a mix of ORDB and semi-structured/XML database systems could outperform pure ORDB technology. An interesting observation was made by more than one panelist regarding the third point: it appears that although ORDB technology can manage patterns, this is not always the most efficient technique to follow

Arriving at the concluding question of the panel, the panelists where asked to express their opinions on the main topics / challenges / opportunities / pitfalls for research in pattern management. The main research areas that were identified were: (a) *visualization*; (b) *operations on patterns* (possibly departing from the traditional relational operations, if this is necessary in order to facilitate typical/interesting user tasks); (c) *modeling and querying languages* (possibly in a QBE fashion); (d) internal *operations of a pattern-base management system*, like similarity algorithms, indexing methods and synchronization (with respect to the changes of the underlying data); (e) support for the whole process of pattern management where people *extract, process and store patterns*.

7 Summary

The lack of database techniques supporting the management of patterns has been the main motivation behind the PaRMa'04 workshop. We believe that the papers presented along with the panel discussions have contributed towards this research direction.

We now summarize the main outcomes of the workshop:

- Although it appears too hard to obtain a generic model and language for all kinds of patterns, it is worthwhile to explore this possibility even for a broad subset of cases (e.g., data mining results). Approaches towards the logical and conceptual modeling, as well as the querying of patterns present a valid research challenge.
- The discovery of pattern similarity and pattern visualization appear to be a couple of approaches that are highly interesting. Both approaches aim at extracting useful knowledge at the pattern space through different means and provide interesting research topics, as well.
- There are several other research opportunities, with the building of systems and the exploration of their internal operations and data representation mechanisms being the most prominent one. Returning to our starting point, the provision of database-like support to the management of patterns is an open, novel and wide area of research.

Acknowledgments

On behalf of the Program Committee we would like to thank all the authors of submitted papers and all the referees for their careful work. The support of the European Commission through the IST/FET Programme and the PANDA Working Group (IST-2001-33058) is also gratefully acknowledged. Finally, we would like to express our gratitude to the members of the Organizing Committee of EDBT'04 for their support in organizing this workshop.

References

[1] I. Bartolini, P. Ciaccia, M. Patella, "A Framework for the Comparison of Complex Patterns", in [11].

[2] E. Bertino, B. Catania, A. Maddalena, "Towards a Language for Pattern Manipulation", in [11].

[3] P. Bouros, T. Dalamagas, T. Sellis, M. Terrovitis, "PatManQL: A language to manipulate patterns and data in hierarchical catalogs", in [11].

[4] R. Heilmann, D. Keim, C. Panse, J. Schneidewind, "Finding Spatial Patterns in Network Data", in [11].

[5] D. Keim, M. Wawryniuk, "Identifying Most Predictive Items", in [11].

[6] I. Ntoutsi, "The notion of similarity in data and pattern space", in [11].

[7] N. Pelekis, B. Theodoulidis, I. Kopanakis, "Fuzzy Miner A Fuzzy System for Solving Pattern Classification Problems", in [11].

[8] S. Rizzi et al, "Towards a Logical Model for Patterns", *Proc. of the 22nd Int'l Conference on Conceptual Modeling* (ER'03), Chicago, IL, USA, October 2003.

[9] S. Rizzi, "UML-Based Conceptual Modeling of Pattern-Bases", in [11].

[10] C. Robardet, R. Pensa, J. Besson, J. Ecois, "Using classification and visualization on pattern databases for gene expression data analysis", in [11].

[11] Y. Theodoridis, P. Vassiliadis (eds.), *Proc. of the Int'l Workshop on Pattern Representation and Management* (PaRMa'04), Crete, Greece, March 2004. http://ceur-ws.org/Vol-96/