Panel: Pattern management challenges

Panos Vassiliadis

Univ. of Ioannina, Dept. of Computer Science,
45110, Ioannina, Hellas
E-mail: pvassil@cs.uoi.gr

1 Introduction

The increasing opportunity of quickly collecting and cheaply storing large volumes of data, and the need for extracting concise information to be efficiently manipulated and intuitively analyzed, are posing new requirements for data management systems in both industrial and scientific applications. A usual approach to deal with such huge volume of data is to reduce the available data to knowledge artifacts (i.e., clusters, rules, etc.), also called **patterns**, through data processing methods (pattern recognition, data mining, knowledge extraction) that reduce their number and size to make them manageable from humans while preserving as much as possible their hidden / interesting information. In order to efficiently and effectively deal with patterns, academic groups and industrial consortiums have devoted efforts towards the modeling, storage, retrieval, analysis and manipulation of patterns with results mainly in the area of standards, inductive databases and pattern-base management systems.

Notably, two main lines of research have been pursued, by different research groups. The first line of research, focused around the notion of the **Pattern-Base Management System** (*PBMS*), has been pursued by the members of the European PANDA project [1]. In the PANDA approach, patterns are concise and rich in semantics representations of data. Patterns are differentiated from data: the former lie in the PBMS, where customized operations (e.g., similarity checks) are to be applied, whereas the latter reside in traditional data stores. PANDA has come up with a generic model and languages for patterns [2,3] as well as results for internal data representation in the PBMS. In the second approach, focused around the notion of **inductive databases** which are databases that, in addition to data, also contain patterns (i.e., generalizations extracted from the data) [4,5]. Inductive databases allow the user to perform database mining as an extension of traditional database querying (i.e., involving the querying of data, patterns and combinations of these two). Cinq [6] is a European project that addressed both practical and theoretical issues of inductive querying. Finally, it is worth mentioning that there is already a variety of **industrial standards** for the modeling of patterns, although they are rather in primitive status so far [7].

Naturally, research in the field is still recent and a wide variety of research topics remains open. A fundamental question involves **the possibility of devising a generic model that can cover a significant number of pattern types**. Why is this generalization so important? First, it allows the ability to represent *any* potential pattern within its expressive power. Of course, this is not a sufficient property by

itself: most importantly, a generic model can allow the possibility to define generic operations over patterns. For example, if one would like to detect how similar is a set of association rules with respect to a decision tree, then, a single similarity operation over a generic model would be sufficient. The only other alternative would be to define pairwise similarity operations for all pattern types for whom this would make sense (which does not obviously scale up as the number of pattern types grows).

Naturally, a second question rises: which would these operations be? **Is there a set of common/popular operations that can be applied over a wide range of patterns?** One can think of examples like similarity test, hypothesis testing, prediction in the future, cross-over from patterns to data and so on. Of course, this ad hoc list should have to fit smoothly in the context of a model.

Finally, even if the aforementioned were accomplished, there are still important issues to resolve. Even if genericity resolves problems at the logical level (i.e., it allows the smooth integration of query language and data mining manipulations) it is not obvious how patterns are to be represented at the physical level. Therefore, obvious research issues rise, with particular focus on the software architecture, indexing and so on.

On the grounds of the aforementioned topics, the panel 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”/operations over patterns?
- How would you manage patterns?
- How would you prescribe a research agenda for pattern management?

2 Panelists

- B. Catania, Univ. of Genova. Prof. Catania has worked in the area of deductive, constraint and object oriented databases, as well as XML data management. Prof. Catania is also involved in the PANDA project.
- D. Keim, Univ. of Konstanz. Prof. Keim has been extensively involved with data visualization techniques for large databases, and with data mining algorithms, too. Prof. Keim is also involved in the PANDA project.
- C. Robardet, Univ. of Lyon. Prof. Robardet has worked on the area of clustering with emphasis on the comparison of clustering algorithms. Prof. Robardet has been involved in the Cinq project.
- Y. Theodoridis, CTI & Univ. of Piraeus. Prof. Theodoridis has worked in the area of spatial indexing and spatiotemporal data management. Prof. Theodoridis has been the coordinator of the PANDA project.

3 What do you think patterns are?

The topic involved discussing a definition of patterns, along with a discussion on their usage and the potential stakeholders who would exploit their existence.

The panelists related to the PANDA project would stick with the same definition of patterns: *patterns are a concise and compact representation of data, rich in semantics*. On the other hand, there was a different understanding from the part of the Cinq approach, where *local patterns* (like association rules) are distinguished from *models* (like clusters) on the basis of using correct algorithms for extracting the former and heuristics for extracting the latter.

A second interesting issue was raised concerning the relationship among patterns and metadata, with patterns appearing to carry a lot more semantics than metadata do.

In terms of the stakeholders concerned and the usage of patterns, there was a broad agreement that patterns are needed almost everywhere, in order to extract extra knowledge from the bulk of available data. Highlighted areas of interest involve astronomical, biological, and telecom data along with applications exploiting semantically rich information like grid, and semantic web applications.

4 Is it possible to devise a generic model for patterns?

The topic involved discussing whether there exist common characteristics that patterns share and whether there is a possibility of devising a generic model over these characteristics.

The panel practically rejected the possibility of being able to devise a common generic model that covers all cases, although one might detect a set of common characteristics in all patterns (like structure, relationship to underlying data and so on). Still, the study of these characteristics involves associating semantic context with patterns and a meta- level perspective on their management.

There was the opinion that not only is this genericity rather hard to obtain but it also hides a potential impact of lack of semantics. As a complement to this, it was also stated that different kinds of patterns should be expressed in different models that can be instantiated. In fact, there was also the opinion that it is probably impossible to obtain some structure for patterns in the traditional (i.e., relational) understanding of the term.

The panelists also agreed that extensibility is the key to handle the diversity of patterns. The extensibility mechanisms should be used however, in order to combine patterns of different nature.

5 What do you think are the main “queries”/operations over patterns?

This question addresses the possibility of having common operations applied to all kinds of patterns. In this context, other questions were raised such as the current and

future exploitation of patterns and the main requirements in terms of query languages and visualization techniques.

According to the Cinq approach, the way people handle patterns now is very dependent on the domain of application. In most cases though, one could argue that (a) simple observation is the most likely treatment and (b) the integration of patterns with the respective data is another possible operation. A classification of potential operations was also suggested, involving data retrieval, pattern processing, data and pattern cross-over queries, and data mining queries.

There was also the opinion that it is pointless to ask on *common*, but rather on *useful* or *important* operations. Similarity check on patterns is one such operation with pattern update and synchronization being other interesting operations, too. An interesting issue was raised at this point, that whenever speaking about similarity or combination of patterns, this should make sense in the first place.

In terms of language requirements, the necessity for a calculus and an algebra for pattern management were clear. This kind of languages, with an emphasis on formality, was a major concern, since it can also be the guide for optimization. Still, the practical necessity for SQL extensions was also suggested (in terms of how easily one can implement a language on top of existing technology).

Visualization was also an issue of broad discussion. Visualization was identified as a major necessity, since there are “visual” operations that cannot really be expressed with SQL-like extensions.

More radical opinions were also expressed concerning algebras that are driven by the visual operations and imprecise query answering based either on visual operations or Google-like queries. The main motivation behind these approaches would be the desire to bring the user directly in contact with a pattern-base management system rather than through some GUI developed by a third party. Still, there was quite extended criticism on these approaches on the grounds of imprecision of the answers.

6 How would you manage patterns?

If one would built a pattern-base management system (a-la RDBMS) how would the software architecture be? Is (object) relational technology enough? These were the main questions which drove the discussion for this topic.

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. Special focus should be paid to index structures for pattern management.

An interesting observation was made by more than one panelists, regarding the third point: it appears that although ORDB technology can manage patterns, this is not always the most efficient technique to follow. In fact, it was reported that there are cases where data are dumped in files for more efficient processing.

7 How would you prescribe a research agenda for pattern management?

Coming to 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. An accompanying question concerned the role of the database community, other scientific communities and the industry in this context.

The main research areas that were identified were:

- Visualization;
- Operations on patterns (but still, operations that make sense);
- Modeling and querying languages (possibly in a QBE fashion);
- Internal operations of a pattern-base management system, like similarity algorithms, indexing methods and synchronization;
- Support for the whole process of pattern management where people extract, process and store patterns.

Naturally, the database community is expected to play a key role in this line of research. Moreover, other scientific communities can be key drivers for this technology (rather than simply 'data providers'). Collaboration with the statistics and machine learning communities should be valuable, too. Finally, industry should also play a key role, not only due to any potential financing, but also due to the necessity of standards for the whole pattern management process.

Acknowledgments

Naturally, many thanks go to the panelists for their participation and to the audience for a vivid discussion. The moderator would also like to warmly thank Dr. Theodore Dalamagas, Manolis Terrovitis and Dr. Spiros Skiadopoulos for their assistance in keeping notes of the panel discussions.

References

1. The PANDA Project, http://dke.cti.gr/panda, 2002
2. Stefano Rizzi, Elisa Bertino, Barbara Catania, Matteo Golfarelli, Maria Halkidi, Manolis Terrovitis, Panos Vassiliadis, Michalis Vazirgiannis, Euripides Vrachnos, "Towards a Logical Model for Patterns", in Proc. 22nd Int'l Conference on Conceptual Modeling (ER'03), Chicago, IL, October 2003.
3. E. Bertino, B. Catania, A. Maddalena, "Towards a Language for Pattern Manipulation", in Proc. 1st International Workshop on Pattern Representation and Management (PaRMa'04), Crete, Greece, March 2004.
4. T. Imielinski and H. Mannila, "A database perspective on knowledge discovery", *Communications of the ACM*, vol. 39(11), pp. 58–64, 1996.
5. L. De Raedt, "A perspective on inductive databases", *SIGKDD Explorations*, vol. 4(2), pp. 69–77, 2002.
6. The Cinq Project, http://www.cinq-project.org/

7. M. Vazirgiannis et al. A Survey on Pattern Application Domains and Pattern Management Approaches. PANDA Technical Report TR-2003-01, February 2003. Available at [1].