Dipartimento di Informatica e Scienze dell'informazione
Università degli Studi di Genova

Dottorato di Ricerca in Informatica

Pattern Based Management: Data Models and Architectural Aspects

- PhD THESIS PROGRESS REPORT -

Anna Maddalena

December 2004

Contents

1 Introduction

My thesis deals with pattern management. Patterns are concise, but rich in semantic, representation of data (data mining results, multimedia features, Web structural information, etc.). Only few approaches exist to deal with heterogeneous patterns even if this is an hot topic in most data-intensive and distributed applications. The main goals of my thesis are: (i) definition of a pattern data model and languages to cope with different types of patterns; (ii) development of a prototype system implementing the proposed framework; (iii) investigation concerning the usage of the proposed model and languages in centralized and advanced distributed environments.

In Section 2 the goals of my PhD thesis are summarized and Section 3 gives a synthetic overview of my work. Then, in Sections 4 and 5 achieved goals and work in progress are discussed, respectively. In Section 6, a future work plan is presented. Finally, Section 7 reports a possible structure of the index of my PhD Thesis.

2 Thesis Goals

In this section, the goals of my PhD thesis are summarized. At the same time, a re-modulation of the objectives identified in [1] is presented, according to hints and comments given by the PhD evaluation committee during the presentation of my Phd proposal in January 2004.

The goals can be grouped in three different tasks: task "T1", concerning pattern modeling issues; task "T2", concerning pattern languages issues; task "T3", concerning architectural issues. Logical issues are taken into account in tasks "T1" and "T2", whereas architectural issues are taken into account in task "T3". Besides these three tasks, there is also an overall activity concerning the analysis of existing work dealing with pattern (knowledge) management. Since this research activity is a background for all the issues we plan to investigate, it can be considered as a task "T0".

2.1 Background

T0 - *Related work analysis*: the basic aim of this task is to analyzed existing proposals for pattern management in order to perform a comparative study pointing out peculiarities of the framework we want to propose.

2.2 Logical modeling goals

T1 - *Patterns model*: development of a pattern data model to represent pattern types and their instances. In order to achieve this goal, the following activities should be considered:

1. Definition of a pattern model by identifying:
   (a) basic pattern logical model concepts (e.g. pattern types and patterns);
   (b) interesting relationships between patterns aimed at increasing the expressivity of the proposed logical model, but which also improve reusability, extensibility, and impact query flexibility (i.e., specialization hierarchy, refinement and composition relationships).
2. Extension of the model with advanced concepts supporting *time data management*. To deal with synchronizations problems, the pattern logical model has to be enriched with temporal aspects in order to be able to manage the notions of transaction and

validity time. In the pattern based management system (PBMS) context, the pattern transaction time is the instant of time at which a pattern is generated and inserted in the PBMS. On the other hand, the concept of validity time has to be defined with respect to the validity time associated with the set of raw data the pattern represents.

T2 - *Languages for pattern manipulation*: development of languages for manipulating and querying patterns represented according to the proposed model. In particular:

1. *Pattern Manipulation Language (PML)*: development of a language to enable users to insert new patterns generated from raw data or from scratch into the PBMS, delete, and modify existing patterns. In particular, traditional manipulation operations, such as insertion, deletion, and update have to be re-interpreted in the context of a pattern-based system.
2. *Pattern Query Language (PQL)*: development of a language supporting the retrieval of patterns from the PBMS. In particular, the following issues will be considered:
   (a) Definition of an algebra for patterns: analysis and formalization of an algebra for patterns manipulation.
   (b) Definition of a calculus for patterns: analysis and formalization of a calculus for patterns manipulation.
   (c) Standard representation of PQL and PML using SQL or XML syntax.
   (d) Equivalence between the calculus and the algebra: formal demonstration of the equivalence of the proposed languages.
   (e) Expressive power of the proposed languages: investigation of the expressive power of the proposed query language.
3. *Query optimization*: this activity concerns optimization issues. In particular, rewriting strategies to improve the patterns query evaluation process will be identified and analyzed.

2.3 Architectural goals

From an architectural point of view, the main goals of the thesis are listed in the following.

T3 - Architectural aspects concerning pattern management will be investigated. In particular:

1. Definition of a testbed architecture for pattern management. A prototype implementation of a pattern-based system will be developed and a case study will be provided.
2. Analysis of the usage of the proposed framework in distributed architecture, with special reference to GRID environments.

3 Working plan for identified goals

The goal listed in Section 2 can be classified as follows:

- **Achieved**. Goals that have been achieved before December 2004.
- **In progress**. Goals in progress that we plan to complete within March 2005.

Goal	Achieved	In progress	Future
T0	(X)	X	
T1.1.a	X		
T1.1.b	X		
T2.1	X		
T2.2.a	X		
T2.2.b	(X)	X	
T2.2.c			X
T2.2.d		X	
T2.2.e		X	
T2.3		(X)	X
T3.1	(X)	X	X
T3.2			X

Table 1: Overview of the goals

- **Future Work**. Goals to be investigated in the last year of my PhD, i.e. from March 2005 to March 2006.

Table 1 shows a synthetic overview of the state of my work. In particular, for each goal identified in Section 2, the table points out its current state. The symbol (X) indicates that several preliminary activities concerning the goal have been conducted even if the goal has not been completed yet.

4 Activity report and achieved goals

In the following, I will briefly summarize the activities carried out until now and the obtained results.

4.1 Task T0: Related work analysis

Since the problem of patten management is very interesting and widespread, many researchers both from the academic world and the industrial one are devoting efforts on this. The areas mainly involved concern standards for pattern representation and exchange, inductive databases, and knowledge and pattern management systems. Of course, the target issue in pattern management in each area can be different. For example, standardization efforts mainly deal with pattern representation in order to achieve interoperability and pattern exchange among different systems. On the other hand, efforts in the area of knowledge management systems try to address theoretical and practical aspects concerning modeling and manipulation of patterns. In order to get a clear picture of the existing work, during 2004, a deep analysis of existing proposals has been conducted.

In the area of standards, several existing standardization efforts for modeling patterns have been analyzed [15, 16, 18]. From a more theoretical point of view, the most relevant research effort in the literature concerning pattern management is in the field of inductive databases, i.e., databases that, in addition to data, store and manage patterns [17, 13]. Many research efforts have been put in the CINQ project [20], aimed to define the basic theoretical and practical issues of inductive database querying process. Finally, a recent unified framework for data mining and

analysis [14] has been analyzed. Under this framework an iterative knowledge discovery (KDD) process is supported.

Unfortunately, although several efforts have been invested in order to deal with patterns, no coherent approach has been proposed and convincingly implemented for an heterogeneous model yet. Thus, in many cases the proposed approaches deal with a fixed set of pattern types and either representation or querying purposes are taken into account. Moreover, issues concerning the manipulation and querying of patterns themselves, and additional mining process over them have not analyzed in depth and no effective and efficient solutions have been identified.

4.2 Task T1: Patterns model

Definition of the model (Goal T1.1). Some research work leading to the formalization of the concepts at the basis of the logical model for patterns has been conducted during 2004. In particular, the model presented in [1, 2, 5], which is heterogeneous and suitable to represent a wide variety of patterns including a-priori and a-posteriori ones[1], has been refined and formalized by using the complex value model proposed in [12].

To this end, formal definitions of the three basic model concepts (i.e., pattern types, patterns, and classes) have been provided and hierarchical relationships between pattern types (and, therefore, patterns) have been taken into account.

Two components of patterns require a specific formalization: the *data source*, representing the set of source data from which the pattern has been extracted, and the pattern *formula*, representing the relationship between a pattern and the raw data from which it has been generated. Both components can be intensionally represented by using ad hoc languages that we plan to formally define.

Finally, the relation between patterns and data they represent has been studied in depth. The exact relationship between patterns and data has been modeled as an independent component of the system, called *intermediate mapping*. On the other hand, the formula represents an approximate relationship describing a region in the data source space approximately represented by the pattern.

The results of this research activities are reported in [7, 9, 11].

Extension of the model with temporal features (Goal T1.2). Since source data change with high frequency, an important issue consists in determining whether existing patterns, after a certain time, still represent the data source from which they have been generated. To deal with with semantic and temporal validity issues the model proposed in [2, 19] has been extended. To this end, two different time information can be considered: a *transaction time* and a *validity period*.

The *transaction time* is the time the pattern "starts to live" in the system. For a-priori patterns, it is the instant when the user inserts the pattern in the system; for a-posteriori patterns, it is the instant when the pattern is extracted from raw data and inserted in the system. It is automatically computed by the PBMS and cannot be changed by the user, thus it is just recorded in system catalogs. On the other hand, the *validity period* is the interval $[StartTime : tg, EndTime : tg)$ in which the pattern can be considered reliable with respect to raw data. The validity period can be queried by the user, thus it must

[1] We recall that *a-posteriori* patterns are extracted form raw data by using data mining tools, whereas *a-priori* patterns are known by the user or the knowledge administrator in advance and used, for example, to check how well some data set is represented by them.

be inserted in the model. Thus, each pattern defined according to [19] is extended with a new component vt, representing the actual pattern validity period according to the chosen granularity.

Pattern temporal validity specifies that the pattern is *assumed* to be valid in a certain period, i.e, its validity period. However, since raw data change with a high frequency, the pattern, in its validity period, may not correctly represent raw data it is associated with. To this end, also the concept of *semantic validity* with respect to a data source and the notion of *safety*, for patterns that are both temporally and semantically valid in a certain instant of time have been introduced.

The logical model extended with temporal information associate with patterns has been presented in [8].

4.3 Task T2: Languages for patterns

Pattern Manipulation Language (Goal T2.1). A Pattern Manipulation Language (PML) is used to generate patterns from raw data and to insert them in the PBMS, to delete and to update patterns. A preliminary proposal for a PML has been presented in [4].

After the extension of the model for patterns with temporal informations (see above), the pattern manipulation language has been extended with several features dealing with time information. As the PML, the Temporal PML must support primitives to generate patterns from raw data, to insert them in the PBMS, to delete, and to update patterns. These operations have to be defined by taking into account validity issues and differences between a-posteriori and a-priori patterns. A temporal pattern manipulation language (TPML) has been proposed in [8].

Pattern Query Language (Goal T2.2). The Pattern Query Language (PQL) supports the retrieval of patterns from the PBMS. Besides operators querying patterns, it is also important, for real application purposes, to support operations binding patterns with raw data and exploiting the relationship between them. Such operations are known as *cross-over queries* since for their execution both the PBMS and the DBMS where raw data rely have to be used.

Algebra (Goal T2.2.a). Many research activities have been conducted in order to define an algebra for the Pattern Query Language (PQL). The first proposal of a set of algebraic operators dealing with patterns has been presented in [4]. Predicates forcing an intensional or an extensional evaluation of certain pattern components have been introduced. Intensional predicates can be checked by exploiting only intensional form of the pattern components. This means that it is possible to solve them locally to the PBMS, without accessing raw data. On the other side, extensional predicates can be checked only by accessing both patterns and data. Thus, they require an execution process involving both the PBMS and the DBMS (cross-over processing).

A more formal algebraic approach for the querying processing has been exploited in [7]. Here the notion of 'query' over the PBMS has been formally defined and a classification of query operators has been provided. Such a classification is based on which system - the database or the pattern base management system - is required to answer the query. In particular, four different query classes have been identified: (i) data base query operators, (ii) pattern base query operators, (iii) cross-over data base operators, and (iv) cross-over pattern base operators.

Recently, the algebraic approach proposed in [4, 7] has been formalized by exploiting traditional complex value algebraic operators [12]. The results of this research activity are reported in [9].

After the extension of the model for patterns with temporal informations (see above), the pattern query language algebra has been extended with several features dealing with time information. A temporal pattern manipulation language (TPML) taking into account temporal information associated with patterns has been proposed in [8]. In particular, the set of pattern predicates has been extended and several predicates concerning pattern validity have been introduced.

Calculus (Goal T2.2.b) A calculus for patterns has already been analyzed and it is under definition. It is a complex value calculus and it is based on the approach presented in [12]. In particular, the notion of query of the calculus has been introduced and well formed expressions for such a calculus have been identified. The results of this research activity are reported in [11].

4.4 Task T3: Architectural aspects

Pattern management system prototype (Goal T3.1). Due to the characteristics of the proposed PBMS, the object-relational database technology seems the more suitable to develop a PBMS prototype. Indeed, in order to support cross-over operations, we need to store both raw data, that can be stored as tables in a relational DBMS, and patterns, that are complex data and can be implemented as objects in an object-oriented DBMS. Thus, the Oracle environment, which is object-relational, has been chosen for this purpose [21].

A high-level, preliminary design of the system prototype has been performed in order to understand the main parts composing the system. The focus of this basic design activity has been the analysis of each model component in order to identify how they can be represented using the object-relational technology.

Some preliminary research activities concerning the analysis of a pattern management case study in the context of the Web have also been conducted. These results are presented in [6] and in [3].

In particular, the proposal in [6] shows how the proposed pattern modeling framework can be used for representing and manipulating cluster results, extracted by means of the analysis of Web-server logs (i.e., click sequences) and applicational logs (i.e., users profile information).[2]

Also the proposal in [3] deals with patterns derived from the Web, since it focus on patterns resulting from the clickstream analysis.

5 Work in progress

In this section current research activities will be discussed.

[2]This work have been performed before the proposal of the pattern model extended with temporal features [8], thus it is based on the model for representing patterns presented in [2, 19].

5.1 Task T0: Related work analysis

A chapter, titled *Pattern Management: Practice and Challenges*, for the upcoming book *Processing and Managing Complex Data for Decision Support*, where all the analyzed approaches are presented and compared is in preparation [10].

5.2 Task T2: Languages for patterns

Calculus (Goal T2.2.b). The calculus has been used to formalize the definition of the intensionally described components of the model: the datasource and the formula components. In this way, the expressive power of the proposed calculus can be exploited and no other dedicated languages are needed. In particular, we are currently defining two different sub-calculi: $\mathcal{DS}$-calc and $\mathcal{F}$-calc. $\mathcal{DS}$-calc is used to express the query which intensionally represents the datasource of a certain pattern, whereas $\mathcal{F}$-calc is suitable to express the pattern formula component.

Equivalence between algebra and calculus (Goal T2.2.d). The formal proof of the equivalence between algebra and calculus is under development.

Expressive power analysis (Goal T2.2.e) The sub-calculus $\mathcal{F}$-calc (see above) can be extended integrating several logical theories. In this way, its expressive power can be varied.

We are currently investigating which logical theories can be used for this purpose and which is the resulting expressive power.

5.3 Task T3: Architectural aspects

Pattern management system prototype (Goal T3.1). The physical design and the implementation of the system prototype are under development.

The focus of this design activity is the investigation concerning how requirements identified in the previous design phase can be represented using the technology chosen for the prototype implementation (i.e. the Oracle technology). To this end, each model concept (i.e., pattern type, pattern, and classes) has been mapped into Oracle concepts. Thus, each pattern type is implemented by defining a corresponding Abstract Data Type (ADT). Then, each pattern is an object instantiating the ADT corresponding to its pattern type. Finally, pattern classes are implemented by way of typed table of a specific ADT (i.e., the one corresponding to the pattern type over which the class has been defined). Concerning the datasource component of a pattern, we have chosen to implement it by using a view definition, i.e. an Oracle query over raw data. On the other side, concerning the implementation of the formula component of a pattern, we have chosen to implement it by using a method. The behavior of such a method, i.e. its body, has to be specified by the user by using an Oracle PL/SQL syntax.

The development of the system prototype using Oracle technology has been assigned as a master thesis work and it has already started [22].

6 Future work

In this section, the main goals to complete my PhD work are presented. Mainly, in the last year of my PhD activity I will work on tasks "T2" and "T3".

6.1 Task T2: Languages for patterns

Standard representation of PML and PQL using a standard syntax (Goal T2.2.c). I plan to devote some efforts in translating the proposed languages for patterns (i.e., the *Pattern Manipulation Language* and the *Pattern Query Language*) in a standard syntax. Probably, due to its flexibility and diffusion, the XML language will be chosen as a target language.

Query optimization (Goal T2.3). I plan to investigate problems concerning query optimization. Particularly, I wish to identify and propose some query strategies that can be efficiently used by the pattern based management system in order to improve the query processing performance.

Moreover, in order to identify efficient query solving plans, I plan to investigate relationships that may exist between the system managing data (i.e. the DBMS) and the ones managing patterns (i.e. the PBMS). Thus, strategies of query execution will be specialized with respect to the fact that the query itself requires the access to the DBMS or the PBMS or both (as in the case of cross-over queries). In this way, the interaction between the two systems can be improved.

6.2 Task T3: Architectural aspects

Pattern management system prototype (Goal T3.1). The prototype of the pattern management system will be completed and implemented, and a case study will be provided.

In particular, to finalize the prototype implementation, issues concerning the implementation of the proposed PML and PQL will be taken into account and primitives to manipulate and query patterns will be provided. Concerning the PQL, we plan to exploit both an intensional approach and an extensional one, providing support for queries involving both types of predicates.

Usage of the proposed framework in distributed architecture (Goal T3.2). I plan to investigate how the proposed framework can be efficiently used in the context of distribute architectures, considering GRID environment as a special case.

7 Thesis structure

In the following, the structure of my PhD thesis is provided, according to the goals pointed out in Section 2.

Introduction

1. A Short Introduction to Patterns Management
2. Research Problems and Objectives
3. Overview of the Dissertation

Chapter 2: Related work

1. Models
2. Languages
3. Architectures

Chapter 2: A Logical Model for Patterns

1. An Informal Approach
   (a) Requirements
   (b) Basic concepts: Pattern Types, Patterns, and Pattern Classes
2. A Formal Approach
   (a) Patterns Types, Patterns, and Classes
   (b) Relationships among Patterns
3. Temporal Extensions of the Model

Chapter 3: Languages for Pattern Management

1. PML: the Pattern Manipulation Language
2. PQL: the Pattern Query Language
   (a) APQL: the Pattern Query Algebra
   (b) CPQL: the Pattern Query Calculus
      - $\mathcal{DS} - calc$: datasource description language
      - $\mathcal{F} - calc$: formula description language
   (c) Equivalence between APQL and CPQL
   (d) PQL Expressive Power
3. Logical optimization issues

Chapter 4: A Prototype of the Pattern Management System

1. The Design
2. The Development
3. A Case Study

Chapter 5: Usage of the PBMS in centralized and distributed architectures

Conclusions

1. Summary of the contributions
2. Topics for further research

Bibliography

References

Personal references

[1] A. Maddalena. Pattern Based Management: Data Models and Architectural Aspects. PhD Thesis Proposal, December 2003.

[2] E. Bertino, B. Catania, M. Golfarelli, M. Halkidi, A. Maddalena, S. Skiadopoulos, S. Rizzi, M. Terrovitis, P. Vassiliadis, M. Vazirgiannis, and E. Vrachnos. The Logical Model for Patterns. Technical Report TR-2003-02, PANDA, 2003.

[3] A. Maddalena and B. Catania. Pattern di Clickworld gestiti con il modello PANDA. Technical report, CLICKWORLD, 2003.

[4] E. Bertino, B. Catania, and A. Maddalena. Toward a Language for Pattern Manipulation and Querying. In *Proc. of the Int. Workshop on Pattern Representation and Management (PaRMa'04)*, in conjunction with EDBT 2004 Conference, Heraklion - Crete, Greece, March 18, 2004.

[5] A. Maddalena. Pattern Based Management: Data Models and Architectural Aspects. In *Proc. of the ICDE/EDBT Ph.D. Workshop*, in conjunction with EDBT 2004 Conference, Heraklion - Crete, Greece, March 18, 2004. LNCS 3268, Springer-Verlag Berlin Heidelberg.

[6] B. Catania and A. Maddalena. A Framework for Cluster Management. In *Proc. of the Int. Workshop on Clustering Information over the Web (ClustWeb)*, in conjunction with EDBT 2004 Conference, Heraklion - Crete, Greece, March 18, 2004. LNCS 3268, Springer-Verlag Berlin Heidelberg.

[7] M. Terrovitis, P. Vassiliadis, S. Skiadopoulos, E. Bertino, B. Catania, and A. Maddalena. Modeling and Language Support for the Management of Pattern-Bases. In *Proc. of the 16th Int. Conf. on Scientific and Statistical Database Management (SSDBM'04)*, Santorini Island, Greece, June 21-23, 2004.

[8] B. Catania, A. Maddalena, M. Mazza, E. Bertino, and S. Rizzi. A Framework for Data Mining Pattern Management. In *Proc. of the 15th European Conference on Machine Learning and 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'04)*, Pisa, Italy, September 20-24,2004. LNAI 3202, pp. 87-98.

[9] M. Terrovitis, P. Vassiliadis, S. Skiadopoulos, B. Catania, A. Maddalena, and S. Rizzi. Modeling and Language Support for the Management of Pattern-Bases. In preparation.

[10] B. Catania and A. Maddalena. Pattern Management: Practice and Challenges. Chapter in the upcoming book *Processing and Managing Complex Data for Decision Support*. In preparation.

[11] A. Maddalena and B. Catania. Towards a Calculus for Patterns. In preparation.

General references

[12] S. Abiteboul and C. Beeri. The Power of Languages for the Manipulation of Complex Values. *VLDB Journal: Very Large Data Bases*, 4(4):727–794, 1995.

[13] T. Imielinski and H. Mannila. A Database Perspective on Knowledge Discovery. *Communications of the ACM*, 39(11):58–64, 1996.

[14] T. Johnson, L.V.S. Lakshmanan, and R.T. Ng. *The 3W model and algebra for unified data mining*. In Proc. of the 26th VLDB, 2000.

[15] Common Warehouse Metamodel (CWM). `http://www.omg.org/cwm`, 2001.

[16] ISO SQL/MM Part 6. `http://www.sql-99.org/SC32/WG4/Progression_Documents/FCD/fcd-datamining-2001-05.pdf`, 2001.

[17] L. De Raedt. A Perspective on Inductive Databases. *ACM SIGKDD Explorations Newsletter*, 4(2):69–77, 2002.

[18] Predictive Model Markup Language (PMML). `http://www.dmg.org/pmmlspecs_v2/pmml_v2_0.html`, 2003.

[19] I. Bartolini et al. Toward a Logical Model for Patterns. In *Proc. of the 22nd International Conference on Conceptual Modeling (ER 2003)*, Chicago, 2003.

[20] The CINQ project. `http://www.cinq-project.org`

[21] The Oracle environment. `http://www.oracle.com/index.html`

[22] M. Mazza. Fondamenti teorici, progettazione e realizzazione di un framework per la rappresentazione e la manipolazione di pattern. Master Thesis. In preparation.