Data mining, also referred to as knowledge discovery in databases (KDD), is a process of finding new, interesting, previously unknown, potentially useful, and ultimately understandable patterns from very large volumes of data. Data mining is a discipline which brings together database systems, statistics, artificial intelligence, machine learning, parallel and distributed processing and visualization between other disciplines (Fayyad et al., 1996; Hand & Kamber, 2001; Hernandez Orallo et al., 2004).

Nowadays, one of the most important and challenging problems in data mining is the definition of the prior knowledge; this can be originated from the process or the domain. This contextual information may help select the appropriate information, features or techniques, decrease the space of hypothesis, represent the output in a most comprehensible way and improve the whole process.

Therefore we need a conceptual model to help represent to this knowledge. According to Gruber’s ontology definition—explicit formal specifications of the terms in the domain and relations among them (Gruber, 1993, 2002); we can represent the knowledge of knowledge discovery process and knowledge about domain. Principally, ontologies are used for communication (between machines and/or humans), automated reasoning, and representation and reuse of knowledge (Cimiano et al., 2004). As a result, ontological foundation is a precondition for efficient automated usage of knowledge discovery information.

Thus, we can perceive the relation between Ontologies and data mining in two manners:

- From ontologies to data mining, we are incorporating knowledge in the process through the use of ontologies, i.e. how the experts comprehend and carry out the analysis tasks. Representative applications are intelligent assistants for discover process (Bernstein et al., 2001, 2005), interpretation and validation of mined knowledge, Ontologies for resource and service description and knowledge Grids (Cannataro et al., 2003; Brezany et al., 2004).
- From data mining to Ontologies, we include domain knowledge in the input information or use the ontologies to represent the results. Therefore the analysis is done over these ontologies. The most characteristic applications are in medicine, biology and spatial data, such as gene representation, taxonomies, applications in geosciences, medical applications and specially in evolving domains (Langley, 2006; Gottgtroy et al., 2003, 2005; Bogorny et al., 2005).

When we can represent and include knowledge in the process through ontologies, we can transform data mining into knowledge mining.
Figure 1 shows our vision of data mining with ontologies cycle.

- **Metadata ontologies:** These ontologies establish how this variable is constructed i.e. which was the process that permit us to obtain its value, and it can vary using another method. Of course this ontology must also express general information about the variable as is treated.
- **Domain ontologies:** These ontologies express the knowledge about application domain.
- **Ontologies for data mining process:** These ontologies codify all knowledge about the process, i.e., select features, select the best algorithms according to the variables and the problem, and establish valid process sequences (Bernstein, 2001, 2005; Cannataro, 2003, 2004).

According with Gomez-Perez and Manzano-Macho (2003) the different methods and approaches, which allow the extraction of ontologies or semantics from database schemas can be classified on three areas, main goal, techniques used and sources used for learning. With regard to the attributes of each area they are the following for summary of ontology learning methods from relational schema are:

- **Main goal**
  - To map a relational schema with a conceptual schema
  - To create (and refine) an ontology
  - To create ontological instances (from a database)
  - Enhance ad hoc queries
• **Techniques used**
  - Mappings
  - Reverse engineering
  - Induction inference
  - Rule generation
  - Graphic modeling

• **Sources used for learning**
  - Relational schemas (of a database)
  - Schema of domain specific databases
  - Flat files
  - Relational databases

In next paragraphs we explain in more detail these three classes of ontologies based on earlier works from different knowledge fields.

**Domain Ontology**

The models on many scientists work to represent their work hypotheses are generally cause effect diagrams. Models make use of general laws or theories to predict or explain behavior in specific situations. Currently these cause effect diagrams can be without difficulty translated to ontologies, by means of conceptual maps which discriminate taxonomy organized as central concepts, main concept, secondary concepts, specific concepts.

Discovery systems produce models that are valuable for prediction, but they should also produce models that have been stated in some declarative format, that can be communicated clearly and precisely, which helps people understand observations, in terms that they find well known (Bridewell, 2006; Langley, 2002, 2006). Models can be from different appearances and dissimilar abstraction level, but the more complex the fact for which they account, the more important that they be cast in some formal notation with an unambiguous interpretation. And of course these new knowledge can be easily communicated and updated between systems and Knowledge databases. In particular into data mining field knowledge can be represented in different formalisms, e.g. rules, decision trees, cluster, known as models. Discovery systems should generate knowledge in a format that is well known to domain users.

There are an important relation between knowledge structures and discovery process with learning machine. The formers are important outputs of discovery process, and are important inputs to discovery (Langley, 2000). Thus knowledge plays as crucial a role as data in the automation of discovery. Therefore, ontologies provide a structure capable of supporting the knowledge representation about domain.

**Metadata Ontologies**

As Spyns et al. (2002) affirm ontologies in current computer science language are computer-based resources that represent agreed domain semantics. Unlike data models, the fundamental asset of ontologies is their relative independence of particular applications, i.e., an ontology consists of relatively generic knowledge that can be reused by different kinds of applications/tasks.

In opposition a data model represents the structure and integrity of the data elements of the, in principle “single”, specific enterprise application(s) by which it will be used. Consequently, the conceptualization and the vocabulary of a data model are not intended a priori to be shared by other applications (Gottgtroy et al., 2005).
Similarly, in data modeling practice, the semantics of data models often constitute an informal accord between the developers and the users of the data model—including when a data warehouse is designed—and, in many cases, the data model is updated as it evolves when particular new functional requirements pop up without any significant update in the metadata repository. Both ontology model and data model have similarities in terms of scope and task. They are context dependent knowledge representation, that is, there doesn’t exist a strict line between generic and specific knowledge when you are building ontology. Moreover, both modeling techniques are knowledge acquisition intensive tasks and the resulted models represent partial account of conceptualizations (Gottgtroy et al., 2003).

In spite of the differences, we should consider the similarities and the fact of data models carry a lot of useful hide knowledge about the domain in its data schemas, in order to build ontologies from data and improve the process of knowledge discovery in databases. Due the fact data schemas do not have the required semantic knowledge to intelligently guide ontology construction has been presented as a challenge for database and ontology engineers (Gottgtroy et al., 2003).

**Ontologies for Data Mining Process**

Vision about KDD process is changing over time. In its beginnings the main objective was to extract a valuable pattern from a fat file as a play of try and error. As time goes by, researchers and fundamentally practitioners discuss the importance of a priori knowledge, the knowledge and understandability about the problem, the choice of the methodology to do the discovery, the expertise in similar situations and an important question arises up to what existent is such inversion on data mining projects worthwhile?

As practitioners and researchers in this field we can perceive that expertise is very important, knowledge about domain is helpful and it simplify the process. To do more attractive the process to managers the practitioners must do it more efficiently and reusing experience. So we can codify all statistical and machine learning knowledge with ontologies and use it.

Bernstein et al. (2001) have developed the concept of intelligent assistant discovery (IDA), which helps data miners with the exploration of the space of valid data mining processes. It takes advantage of an explicit ontology of data-mining techniques, which defines the various techniques and their properties. Main characteristics are (Bernstein et al., 2005).

- A systematic enumeration of valid DM processes, so they do not miss important, potentially fruitful options.
- Effective rankings of these valid processes by different criteria, to help them choose between the options.
- An infrastructure for sharing data mining knowledge, which leads to what economists call network externalities.

Cannataro and colleagues have done another interesting contribution to this kind of ontologies. They developed an ontology that can be used to simplify the development of distributed knowledge discovery applications on the Grid, offering to a domain expert a reference model for the different kind of data mining tasks, methodologies and software available to solve a given problem, helping a user in finding the most appropriate solution (Cannataro et al., 2003, 2004). Authors have adopted the Enterprise Methodology (Corcho et al., 2003).
RESEARCH WORKS IN THE TOPIC

The next paragraphs will describe the most recently research works in data mining with ontologies field.

Singh, Vajirkar, and Lee (2003) have developed a context aware data mining framework which provide accuracy and efficacy to data mining outcomes. Context factors were modeled using ontological representation. Although the context aware framework proposed is generic in nature and can be applied to most of the fields, the medical scenario provided was like a proof of concept to our proposed model.

Hotho, Staab and Stumme (2003) have showed that using ontologies as filters in term selection prior to the application of a K-means clustering algorithm will increase the tightness and relative isolation of document clusters as a measure of improvement.

Pand and Shen (2005) have proposed architecture for knowledge discovery in evolving environments. The architecture creates a communication mechanism to incorporate known knowledge into discovery process, through ontology service facility. The continuous mining is transparent to the end user; moreover, the architecture supports logical and physical data independence.

Rennolls (2005, p. 719) have developed an intelligent framework for data mining, knowledge discovery and business intelligence. The ontological framework will guide to user to choice of models from an expanded data mining toolkit, and the epistemological framework will assist to user in interpreting and appraising the discovered relationships and patterns.

On domain ontologies, Pan and Pan (2006) have proposed ontobase ontology repository. It is an implementation, which allows users and agents to retrieve ontologies and metadata through open Web standards and ontology service. Key features of the system include the use of XML metadata interchange to represent and import ontologies and metadata, the support for smooth transformation and transparent integration using ontology mapping and the use of ontology services to share and reuse domain knowledge in a generic way.

Recently, Bounif et al. (2006) have explained the articulation of a new approach for database schema evolution and outline the use of domain ontology. The approach they have proposed belongs to a new tendency called the tendency of a priori approaches. It implies the investigation of potential future requirements besides the current requirements during the standard requirements analysis phase of schema design or redesign and their inclusion into the conceptual schema. Those requirements are determined with the help of a domain ontology called “a requirements ontology” using data mining techniques and schema repository.

BOOK ORGANIZATION

This book is organized into three major sections dealing respectively with implementations, findings, and frameworks.

Section I: Implementations includes applications or study cases on data mining with ontologies.

Chapter I, TODE: An Ontology-Based Model for the Dynamic Population of Web Directories by Sofia Stamou, Alexandros Ntoulas, and Dimitris Christodoulakis studies how we can organize the continuously proliferating Web content into topical categories, also known as Web directories. Authors have implemented a system, named TODE that uses Topical Ontology for Directories’ Editing. Also TODE’s
performance is evaluated; experimental results imply that the use of a rich topical ontology significantly increases classification accuracy for dynamic contents.

Chapter II, *Raising, to Enhance Rule Mining in Web Marketing with the Use of an Ontology* by Xuan Zhou and James Geller introduces Raising as an operation which is used as a preprocessing step for data mining. Rules have been derived using demographic and interest information as input for data mining. The Raising step takes advantage of interest ontology to advance data mining and to improve rule quality. Furthermore, the effects caused by Raising are analyzed in detail, showing an improvement of the support and confidence values of useful association rules for marketing purposes.

Chapter III, *Web Usage Mining for Ontology Management* by Brigitte Trousse, Marie-Aude Aufaure, Bénédicte Le Grand, Yves Lechevallier, and Florent Masseglia proposes an original approach for ontology management in the context of Web-based information systems. Their approach relies on the usage analysis of the chosen Web site, in complement of the existing approaches based on content analysis of Web pages. One major contribution of this chapter is then the application of usage analysis to support ontology evolution and/or web site reorganization.

Chapter IV, *SOM-Based Clustering of Multilingual Documents Using an Ontology* by Minh Hai Pham, Delphine Bernhard, Gayo Diallo, Radja Messai, and Michel Simonet presents a method which make use of Self-Organizing Map (SOM) to cluster medical documents. The originality of the method is that it does not rely on the words shared by documents but rather on concepts taken from ontology. The goal is to cluster various medical documents in thematically consistent groups. Authors have compared the results for two indexing schemes: stem-based indexing and conceptual indexing.

**Section II:** Findings comprise more theoretical aspects of data mining with ontologies such as ontologies for interpretation and validation and domain ontologies.

Chapter V, *Ontology-Based Interpretation and Validation of Mined Knowledge: Normative and Cognitive Factors in Data Mining* by Ana Isabel Canhoto, addresses the role of cognition and context in the interpretation and validation of mined knowledge. She proposes the use of ontology charts and norm specifications to map how varying levels of access to information and exposure to specific social norms lead to divergent views of mined knowledge. Domain knowledge and bias information influence which patterns in the data are deemed as useful and, ultimately, valid.

Chapter VI, *Data Integration Through Protein Ontology* by Amandeep S. Sidhu, Tharam S. Dillon, and Elizabeth Chang discuss conceptual framework of Protein Ontology that has a hierarchical classification of concepts represented as classes, from general to specific; a list of attributes related to each concept, for each class; a set of relations between classes to link concepts in ontology in more complicated ways than implied by the hierarchy, to promote reuse of concepts in the ontology; and a set of algebraic operators to query protein ontology instances.

Chapter VII, *TtoO: Mining a Thesaurus and Texts to Build and Update a Domain Ontology* by Jo-siane Mothe and Nathalie Hernandez introduces a method re-using a thesaurus built for a given domain, in order to create new resources of a higher semantic level in the form of an ontology. The originality of the method is that it is based on both the knowledge extracted from a thesaurus and the knowledge semiautomatically extracted from a textual corpus. In parallel, authors have developed mechanisms based on the obtained ontology to accomplish a science-monitoring task. An example is provided in this chapter.

Chapter VIII, *Evaluating the Construction of Domain Ontologies for Recommender Systems Based on Texts* by Stanley Loh, Daniel Lichtnow, Thyago Borges, and Gustavo Piltcher, investigates different aspects in the construction of domain ontology to a content-based recommender system. The chapter
discusses different approaches so as to construct the domain ontology, including the use of text mining software tools for supervised learning, the interference of domain experts in the engineering process and the use of a normalization step.

Section III: Frameworks includes different architectures for different domains in data warehousing or mining with ontologies context.

Chapter IX, Enhancing the Process of Knowledge Discovery in Geographic Databases Using Geo-Ontologies by Vania Bogorny, Paulo Martins Engel, and Luis Otavio Alvares introduces the problem of mining frequent geographic patterns and spatial association rules from geographic databases. A large amount of natural geographic associations are explicitly represented in geographic database schemas and geo-ontologies, which have not been used so far in frequent geographic pattern mining. The main goal of this chapter is to show how the large amount of knowledge represented in geo-ontologies as prior knowledge can be used to avoid the extraction of patterns previously known as noninteresting.

Chapter X, Ontology-Based Construction of Grid Data Mining Workflows by Peter Brezany, Ivan Janciak, and A Min Tjoa, introduces an ontology-based framework for automated construction of complex interactive data mining workflows. The authors present their solution called GridMiner Assistant (GMA), which addresses the whole life cycle of the knowledge discovery process. In addition, conceptual and implementation architectures of the framework are presented and its application to an example taken from the medical domain is illustrated.

Chapter XI, Ontology-Based Data Warehousing and Mining Approaches in Petroleum Industries by Shastri L. Nimmagadda and Heinz Dreher. Complex geo-spatial heterogeneous data structures complicate the accessibility and presentation of data in petroleum industries. Data warehousing approach supported by ontology will be described for effective data mining. Ontology based data warehousing framework with fine-grained multidimensional data structures facilitates mining and visualization of data patterns, trends, and correlations hidden under massive volumes of data.

Chapter XII, A Framework for Integrating Ontologies and Pattern-Bases by Evangelos Kotsifakos, Gerasimos Marketos, and Yannis Theodoridis propose the integration of pattern base management systems (PBMS) and ontologies. It is as a solution to the need of many scientific fields for efficient extraction of useful information from large databases and the exploitation of knowledge. Authors use a case study of data mining over scientific (seismological) data to illustrate their proposal.

BOOK OBJECTIVE

This book aims at publishing original academic work with high quality scientific papers. The key objective is to provide to data mining students, practitioners, professionals, professors and researchers an integral vision of the topic. This book specifically focuses on those areas that explore new methodologies or examine real study cases that are ontology-based.

The book describes the state-of-the-art, innovative theoretical frameworks, advanced and successful implementations as well as the latest empirical research findings in the area of data mining with ontologies.
AUDIENCE

The target audience of this book is readers who want to learn how to apply data mining based on ontologies to real world problems. The purpose is to show users how to go from theory and algorithms to real applications.

The book is also geared toward students, practitioners, professionals, professors and researchers with basic understanding in data mining. The information technology community can increase its knowledge and skills with these new techniques.

People working on the Knowledge Management area such as engineers, managers, and analysts can read it, due to the fact that data mining, ontologies and knowledge management areas are linked straightforwardly.

REFERENCES


