KR Meets DB for Data Mining

Amedeo Napoli and **Arnaud Simon**
CRIN CNRS – INRIA Lorraine,
B.P. 239 – 54506 Vandœuvre-lès-Nancy Cedex – France
(Email: {napoli,simona}@loria.fr)

Abstract. Position paper for the Workshop on Knowledge Representation Meets Databases (KRDB'96 at ECAI'96, Budapest).

1 INTRODUCTION

In this position paper, we relate our experiments with the design of a data mining system for use in the field of medicine. The goal of the system is to analyze administrative medical data, e.g. questionnaires about children suffering from cancer, in order to discover patterns in the data that could be interpreted as units of knowledge, i.e. rules, decision trees, classes of individuals. These patterns can be related to the history, family characteristics and environment of the patients. The units of knowledge discovered will be used to provide better administrative services for patients. The data mining system is used in association with a knowledge-based system (KBS), the knowledge base of which contains both medical and administrative knowledge. The role of the KBS is to facilitate the knowledge discovery process, as illustrated in [Brachman *et al.*,1993], where a data base management system (DBMS) and a KBS are combined in a data mining system.

In this work, we are mainly interested in the combination of KBS and DBMS techniques to analyze rough data and to discover knowledge. The following questions, which are related to the topics of the workshop are addressed below:

(i) Why is an object-based representation system (see below) used instead of a relational or an object-oriented DBMS for the data mining process? This first question is related to the workshop topic "KR formalisms as schema languages": data are represented by classes and instances (or objects), and are manipulated by KBS inference formalisms for information retrieval and classification.

(ii) How can KBS and DBMS be combined to enhance the data mining process and what are the problems encountered? This second question deals with the workshop topic "Integration of relational, deductive, and object-oriented formalisms". Analogous questions about integration involves the study of relations between DBMS and description logics (DL) [Borgida and Brachman,1993] [Borgida,1995], relations between object-oriented formalisms and DL [Napoli *et al.*,1994], and the management of large knowledge bases [Karp and Paley,1995].

This paper is organized as follows: first, we briefly introduce data mining in the field of medicine, then we present object-based representation systems, and finally, we discuss the integration problems described in the two questions mentioned above.

2 DATA MINING IN THE FIELD OF MEDICINE

The goal of data mining is to obtain useful knowledge from large masses of normal data [Frawley *et al.*,1992] [Mannila,1995]. One of the basic tasks in data mining is to build descriptions of data by finding "interesting subgroups" from data. There are several forms of descriptions: rules, decision trees, class hierarchies, etc. The data mining process relies on methods and techniques borrowed from machine learning, statistics, and data base management.

A knowledge-based system providing domain knowledge can be used to increase the performances of the data mining process. For our present purpose, two main algorithms, CDP [Agrawal *et al.*,1993] and DB-LEARN [Cai *et al.*,1991] [Hu,1995], are used to analyze data and to build a decision tree for interpreting the medical data. The DB-LEARN algorithm takes advantage of domain knowledge to improve the design and the accuracy of the decision tree. The resultant decision tree yields a list of rules summarizing the medical data, which is provided for physicians for evaluation and validation.

3 OBJECT-BASED REPRESENTATION SYSTEM

In *object-based representation systems* (OBRS), real-world knowledge is represented by generic and specific objects [Napoli *et al.*,1994]. A generic object, or *class*, has an identity and is composed of a set of *properties* describing the behavioral and definitional characteristics of a real-world concept. Thus, a class has a state and a behavior and can be used to generate a set of *instances*, often simply called *objects*, describing real-world individuals (instances of real-world concepts). Classes are organized in a hierarchy $\mathcal{H} = (\mathcal{C}, \preceq, \omega)$, where $\mathcal{C}$ is a set of classes, $\preceq$ is a partial ordering and ω is the *root* of the hierarchy $\mathcal{H}$. The class ω is assumed to exist and to be the greatest element of $\mathcal{C}$ for $\preceq$. Moreover, the hierarchical organization of classes involves *knowledge* or *property sharing*, based on the transitivity of $\preceq$ and depending on the semantics of $\preceq$. Knowledge sharing can be monotonic or nonmonotonic. It is usually used to exhibit implicit knowledge for information retrieval purposes and for default reasoning, i.e. for inferring the existence and values of properties. OBRS systems combine characteristics of object-oriented systems and description logics as described below [Napoli,1994]. It can be interesting to compare OBRS characteristics with those of the generic frame protocol presented in [Karp *et al.*,1995].

In most cases, inheritance is the primary and most powerful representation primitive in object-oriented systems. However, inheritance is mainly a mechanism for knowledge sharing. To improve the deductive capabilities of inheritance, a classification tool is associated with OBRS [Napoli *et al.*,1994]. The classification mechanism is similar to the subsumption-based classification mechanism of description logics [Nebel,1990]. It is used to aid the building of class hierarchy and/or to derive new information through the classification-based reasoning cycle:

(1) *instantiation* of a new object X,

(2) *classification* of X, i.e. searching for the most specific subsumers and the most general subsumees of X,

(3) *updating operations* triggered by insertion of X in the hierarchy.

During classification, the *definitional* part of an object is considered and handled like a defined concept in terminological logics.

The capability of representing data from several viewpoints is very important in a data mining process. Thus, a particular technique called "attribute-oriented subsumption" has been added to the OBRS system [Napoli,1995].

4 DISCUSSION AND CONCLUSION

There are numerous advantages of using OBRS for data mining, e.g. conceptual model of the domain studied, validation and organization of queries and views, query and result reification [Borgida and Brachman,1993] (these advantages are also discussed in [Borgida,1995] and [Karp and Paley,1995]). An OBRS system is an intermediate system with respect to object-oriented, frame-based and description logic systems. It has the advantage of combining numerous interesting properties of the previous systems, especially in a data mining perspective. Moreover, OBRS are more flexible than relational and object-oriented DBMS, and the format of objects describing data is better-suited to handle incomplete or heterogeneous data. In addition to global inference mechanism such as inheritance and classification, it is also possible to associate with classes of objects specific methods for inference purposes and data manipulation.

In conclusion, KBS can be combined with DBMS to enhance the data mining process:

(i) the representational schemes in KBS are more flexible and efficient than the relational DB schemes,

(ii) inference tools, e.g. inheritance, classification, and methods, can be used to analyze data and to perform more sophisticated operations than simple information retrieval.

In a near future, we plan to use a translator such as DRIVER [Lebastard,1995] in order to handle all types and volumes of data (note that other techniques for translating data into objects are also presented in [Norrie *et al.*,1994] and [Karp and Paley,1995]). Up to now, we have worked with a rather small database. However, when the DB has a considerable volume, and this is a usual case, the entire set of data cannot be anymore represented by objects — classes and instances — and translation or virtual memory mechanisms have to be available.

Acknowledgement.

The authors would like to thank the "Région Lorraine" for the financial support provided for the second author.

References

[Agrawal *et al.*, 1993] R. Agrawal, T. Imielinski, and A. Swami. Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5(6):914–925, 1993.

[Borgida and Brachman, 1993] A. Borgida and R.J. Brachman. Loading Data into Description Reasoners. In *Proceedings of the ACM/SIGMOD International Conference on the Management of Data, Washington, DC, SIGMOD RECORD, 22(2)*, pages 217–226, 1993.

[Borgida, 1995] A. Borgida. Description Logics in Data Management. *IEEE Transactions on Knowledge and Data Engineering*, 7(5):671–682, 1995.

[Brachman *et al.*, 1993] R.J. Brachman, P.G. Selfridge, L.G. Terveen, B. Altman, A. Borgida, F. Halper, T. Kirk, A. Lazar, D.L. McGuinness, and L.A. Resnick. Integrated Support for Data Archaeology. *International Journal of Intelligent and Cooperative Information Systems*, 2(2):159–185, 1993.

[Cai *et al.*, 1991] Y. Cai, N. Cercone, and J. Han. Attribute-Oriented Induction in Relational Databases. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*, pages 213–228. AAAI Press / The MIT Press, Menlo Park, California, 1991.

[Frawley *et al.*, 1992] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus. Knowledge Discovery in Databases: An Overview. *The AI Magazine*, 14(3):57–70, 1992.

[Hu, 1995] X. Hu. *Knowledge Discovery in Databases: Attribute-Oriented Rough Set Approach*. PhD thesis, University of Regina, Canada, 1995.

[Karp and Paley, 1995] P.D. Karp and S.M. Paley. Knowledge Representation in the Large. In *Proceedings of the 14th IJCAI, Montréal, Canada*, pages 751–758, 1995.

[Karp *et al.*, 1995] P.D. Karp, K.L. Myers, and T. Gruber. The Generic Frame Protocol. In *Proceedings of the 14th IJCAI, Montréal, Canada*, pages 768–774, 1995.

[Lebastard, 1995] F. Lebastard. Is an object layer on a relational database more attractive than an object database. In *Working Notes of the KI'95 Workshop "Reasoning about Strutured Objects: Knowledge Representation Meets databases (KRDB'95)", Bielefield, Germany (DFKI Report D-95-12)*, pages 7–10, 1995.

[Mannila, 1995] H. Mannila. Aspects of data mining. In *Notes of the ECML'95 Workhop on Statistics, Machine Learning and Knowledge Discovery in Databases, Heraklion, Crete*, pages 1–6, 1995.

[Napoli *et al.*, 1994] A. Napoli, C. Laurenço, and R. Ducournau. An object-based representation system for organic synthesis planning. *International Journal of Human-Computer Studies*, 41(1/2):5–32, 1994.

[Napoli, 1994] A. Napoli. Studies about the Integration of Classification-Based Reasoning and Object-Oriented Programming. In F. Baader, M. Lenzerini, W. Nutt, and P.F. Patel-Schneider, editors, *Working Notes of the 1994 Description Logics Workshop*, pages 60–62. DFKI Saarbruecken, 1994.

[Napoli, 1995] A. Napoli. Objects, Classes, Specialization and Subsumption. In A. Borgida, M. Lenzerini, D. Nardi, and B. Nebel, editors, *Proceedings of the 1995 International Workshop on Description Logics, Universita di Roma (Technical Report 07.95)*, pages 52–55, 1995.

[Nebel, 1990] B. Nebel. *Reasoning and Revision in Hybrid Representation Systems.* Lecture Notes in Computer Science 422. Springer-Verlag, Berlin, 1990.

[Norrie *et al.*, 1994] M.C. Norrie, U. Reimer, P. Lippuner, M. Rys, and H.-J. Schek. Frames, Objects and Relations: Three Semantics Levels for Knowledge Base Systems. In *Working Notes of the KI'94 Workshop "Reasoning about Strutured Objects: Knowledge Representation Meets databases (KRDB'94)", Saarbrueken, Germany (DFKI Report D-94-11)*, pages 53–57, 1994.