An ontology driven data mining process

Laurent Brisson, Martine Collard

AN ONTOLOGY DRIVEN DATA MINING PROCESS

Laurent BRISSON
Institut TELECOM, TELECOM Bretagne, CNRS FRE 3167 LAB-STICC, Technopôle Brest-Iroise, France
laurent.brisson@telecom-bretagne.eu

Martine COLLARD
INRIA Sophia Antipolis, 2004 route des Lucioles, 06902 BP93 Sophia Antipolis, France
University of Nice Sophia Antipolis, France
martine.collard@sophia.inria.fr

Keywords: Data mining, Knowledge integration, Ontology Driven Information System

Abstract: This paper deals with knowledge integration in a data mining process. We suggest to model domain knowledge during business understanding and data understanding steps in order to build an ontology driven information system (ODIS). We present the KEOPS Methodology based on this approach. In KEOPS, the ODIS is dedicated to data mining tasks. It allows using expert knowledge for efficient data selection, data preparation and model interpretation. In this paper, we detail each of these ontology driven steps and we define a part-way interestingness measure that integrates both objective and subjective criteria in order to evaluate model relevance according to expert knowledge.

1 Introduction

In knowledge discovery from data, methods and techniques are developed for discovering specific trends in a system or organization business by analyzing its data. The real advantage for decision making relies on the add-on provided by comparing extracted knowledge against a priori domain knowledge. Integrating domain *a priori* knowledge during the data mining process is currently an important research issue in the data mining field.

In this paper, we present KEOPS methodology based on an ontology driven information system which integrates *a priori* knowledge all along the data mining process in a coherent and uniform manner. We detail each of these ontology driven steps and we define a part-way interestingness measure that integrates both objective and subjective criteria in order to evaluate model relevance according to expert knowledge.

The paper is organized in six sections. Section 2 presents the issue addressed and KEOPS main characteristics. Section 3 is devoted to ontology driven information systems. In section 4, the KEOPS methodology is presented step by step. In Section 5, we comment some results which demonstrate the relevance of the approach. We conclude in Section 6.

2 Knowledge integration in data mining

The Data Mining process described according to the CRISP-DM model (Chapman et al., 2000) is presented as both iterative and interactive. The iterative nature is due to the way processes run cycling test-error experiments. Indeed data miners have to repeat the pre-processing steps of domain understanding, data understanding and data preparation until final models are considered relevant. The interactive nature is inherent to a data mining activity since communications with experts is necessary for understanding domain and data and for interpreting results. Issues in evaluating and interpreting mining process results are currently big research challenges. In order to avoid useless iterations on preliminary tasks and facilitate model interpretation, one solution is to explore deeply expert knowledge and source data in order to formalize them in conceptual structures and exploit these structures both for robust data preparation and for flexible model interpretation.

In the literature, partial solutions for domain knowledge interpretation are proposed for optimizing pre-processing steps (Kedad and Métais, 2002). For model evaluation, detailed studies have been devoted to interestingness measures (McGarry, 2005).

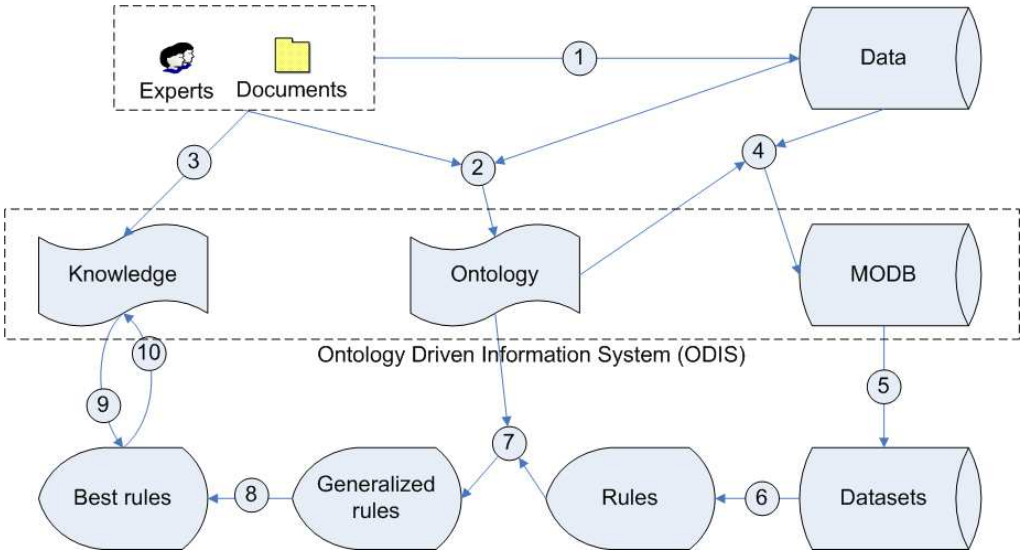

Figure 1: KEOPS methodology

A consensus among researchers is now established to consider objective interestingness versus subjective interestingness. Objective interestingness is traditionally evaluated by a variety of statistic indexes while subjective interestingness is generally evaluated by comparing discovered patterns to user knowledge or *a priori* convictions of domain experts. In this paper we present the KEOPS methodology based on an ontology driven information system which addresses the knowledge integration issue. The system relies on three main components: an ontology, a knowledge base and a mining oriented database rebuilt from source raw data. These components allow to model domain concepts and relationships among them. They are used to pre-process data and to identify mapping between discovered patterns and expert knowledge.

3 Ontology Driven Information System (ODIS)

An ontology driven information systems is an information system (IS) which relies mainly on an *explicit* ontology. This ontology may underlie all aspects and components of the information system. An ODIS contains three kinds of components: application programs, information resources and user interfaces. (Guarino, 1998) discusses the impact of an ontology on an information system according to temporal and structural dimension.

The temporal dimension refers to ontology role during IS construction and run-time. If we have a set of reusable ontologies, the semantic content expressed can be transformed and translated into an IS component. Even if the volume of ontology knowledge available is modest it may nevertheless help a designer in a conceptual analysis task. This task consists frequently of redesigning an existing information system. This approach fits the needs of data mining tasks where an operational database has to be transformed into datasets before the data mining modeling step.

The structural dimension refers to each information system component which may use the ontology in a specific way.

- Database component: at development time, an ontology can play an important role in requirement analysis and conceptual modeling. The resulting conceptual model can be represented as a computer processable ontology mapped to a concrete target platform (Ceri and Fraternali, 1997). Usually, IS conceptual schemes (CS) are created from scratch, wasting a lot of time and resources.

- Interface components may be assisted by ontologies which are used to generate personalized interfaces or to manage user profiles (Guarino et al., 1998; Penarrubia et al., 2004).

- Application program components use implicit knowledge in order to perform a task. However, this knowledge is often hardcoded in software. Ontologies may provide a formal base helping to access domain knowledge.

4 KEOPS methodology

KEOPS is a methodology which drives data mining processes by integrating expert knowledge. These are the goals addressed:

- To manage interactions between knowledge and data all along the data mining process: data preparation, datasets generation, modeling, evaluation and results visualization.
- To evaluate extracted models according to domain expert knowledge.
- To provide easy navigation throughout the space of results.

KEOPS (cf. fig. 1) is based upon an ontology driven information system (ODIS) set up with four components:

- An application ontology whose concepts and relationships between them are dedicated to domain and data mining task.
- A Mining Oriented DataBase (MODB): a relational database whose attributes and values are chosen among ontology concepts.
- A knowledge base to express consensual knowledge, obvious knowledge and user assumptions.
- A set of information system components - user interfaces, extraction algorithms, evaluation methods - in order to select the most relevant extracted models according to expert knowledge.

KEOPS methodology extends the CRISP-DM process model by integrating knowledge in most steps of the mining process. The initial step focuses on business understanding. The second step focuses on data understanding and activities in order to check data reliability. Data reliability problems are solved during the third step of data preparation. The fourth step is the evaluation of extracted models. In this paper we don't focus on modeling step of CRISP-DM model since we ran CLOSE algorithm (Pasquier et al., 1999) which extracts association rules without domain knowledge.

4.1 Business understanding

During business understanding step, documents, data, domain knowledge and discussion between experts lead to assess situation, to determine business objectives and success criteria, and to evaluate risks and contingencies. However this step is often rather informal.

KEOPS methodology requires to build an ontology driven information system during the next step, data understanding. Consequently an informal specification of business objectives and expert knowledge is henceforth insufficient. Thus, it is necessary to formalize expert knowledge during business understanding. We chose to state knowledge with production rules, also called "if ... then ..." rules. These rules are modular, each defining a small and independent piece of knowledge. Furthermore, they can be easily compared to extracted association rules. Each knowledge rule has some essential properties to select the most interesting association rules:

- Knowledge confidence level: five different values are available to describe knowledge confidence according to a domain expert. These values are ranges of confidence values: 0-20%, 20-40%, 40-60%, 60-80% and 80-100%. We call confidence the probability for the rule consequence to occur when the rule condition holds.
- Knowledge certainty:
  - Obvious: knowledge cannot be contradicted.
  - Consensual: domain knowledge shared among experts.
  - Assumption: knowledge the user wants to check.

Since the description of expert interview methodology in order to capture knowledge is beyond the scope of this paper, the reader should refer to (Becker, 1976).

4.2 Data understanding

Data understanding means selection and description of source data in order to capture their semantic and reliability. During this step, the ontology is built in order to identify domain concepts and relationships between them (the objective is to select among data the most interesting attributes according to the business objectives), to solve ambiguities within data and to choose data discretization levels.

Consequently, the ontology formalizes domain concepts and information about data. This ontology is an application ontology; it contains the essential knowledge in order to drive data mining tasks. Ontology concepts are related to domain concepts, however relationships between them model database relationships. During next step, data preparation (cf. section 4.3), a relational database called Mining Oriented DataBase (MODB) will be built.

In order to understand links between the MODB and the ontology it is necessary to define notions of domain, concept and relationships:

- Domain: This notion in KEOPS methodology, refers to the notion of domain in relational theory.

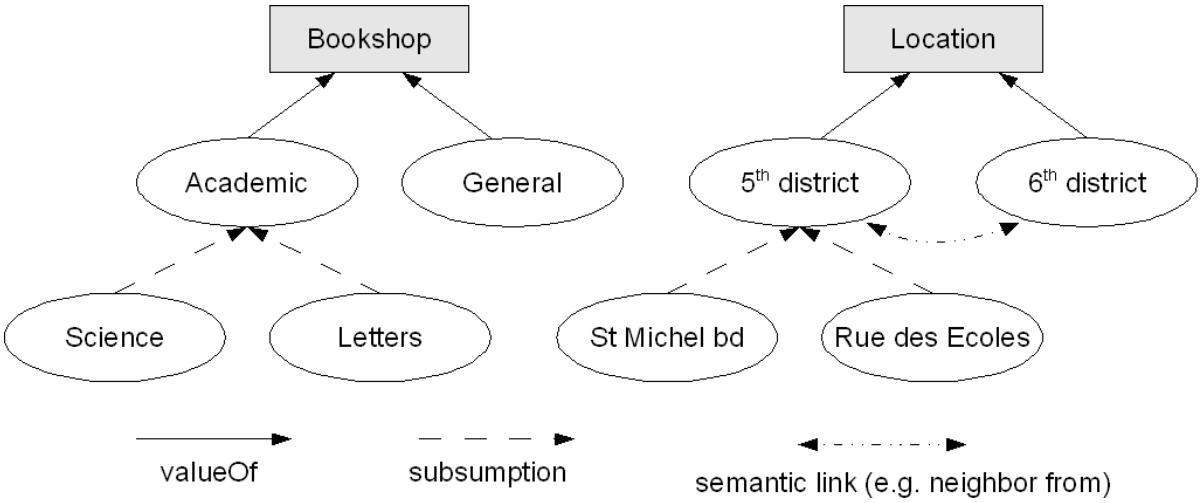

Figure 2: Bookshop ontology snapshoot

A domain represents a set of values associated to a semantic entity (or concept).

- Concept: Each concept of the ontology has a property defining its role. There exist two classes of concepts: attribute concepts and value concepts.
  - An *attribute concept* is identified by a name and a *domain*.
  - Each value of domain is called a *value concept*.

  Thus a domain is described by an attribute concept and by value concepts organized into a taxonomy. Each MODB attribute is linked to one and only one attribute concept and takes its values in the associated domain. In figure 2 "Bookshop" is an attribute concept, "Academic" a value concept and the set {Academic, General, Sciences, Letters} defines "Bookshop" domain.

- Relationships: There exists three kinds of relationships between concepts:
  - A data-related relationship: "valueOf" relationship between an attribute concept and a value concept. The set of value concepts linked to an attribute concept with "valueOf" relationship define a domain within the MODB.
  - Subsumption relationship between two value concepts. A concept subsumed by another one is member of the same domain. This relationship is useful during data preparation (to select data granularity in datasets), reduction of rule volume (to generate generalized association rules, see 4.4.1, comparison between models and knowledge (to consider sibling and ancestor concepts) and final results visualization.
  - Semantic relationships between value concepts. These relationships could be order, composition, exclusion or equivalence relationships. They can be used to compare extracted models and knowledge and to visualize results.

KEOPS methodology aims to extract interesting models according user knowledge. Consequently, it is necessary during ontology construction to be careful with some usual concerns in data mining:

- Aggregation level: like data, ontology concepts have to represent disjoint domains.
- Discretization level: ontology concepts have to model various solutions for data discretization. Bad choices may affect modeling step efficiency.
- Data correlation: if concepts are strongly related into the MODB, extracted models might be trivial and uninteresting.

Since these concerns are beyond the scope of this paper, the reader should refer to (De Leenheer and de Moor, 2005) for a better insight on concept elicitation and (Berka and Bruha, 1998) for a better insight on discretization and grouping.

Example Let's take the case of a bookstore company with several bookshops in Paris which plan to improve customer relationships. Bookshops may be specialized in a field like "academic" or not (general) (see figure 2). Bookshops are located geographically. Data are provided on bookshops, customers and sales.

Source Data	Attribute Concept	Value Concept
St Denis Shop	Bookshop	Academic
St Denis Shop	Location	St Michel bd
Rive Gauche 5th	Bookshop	General
Rive Gauche 5th	Location	5th District

Table 1: Bookshop ontology concept elicitation

Table 1 shows a way for mapping source values to ontology concepts.

4.3 Data preparation

Data preparation is very iterative and time consuming. The objective is to refine data: discretize, clean and build new attributes and values in the MODB. During this step, KEOPS suggests building MODB by mapping original data with ontology concepts. The database contains only bottom ontology concepts. The objective is to structure knowledge and data in order to process efficient mining tasks and to save time spent into data preparation. The idea is to allow generation of multiple datasets from the MODB, using ontology relationships without another preparation step from raw data. Furthermore, during ODIS construction, experts can express their knowledge using the ontology which is consistent with data.

4.3.1 Mining Oriented Database (MODB) Construction

Databases often contain several tables sharing similar information. However, it is desirable that each MODB table contains all the information semantically close and it's important to observe normal forms in these tables. During datasets generation, it's easy to use join in order to create interesting datasets to be mined. However these datasets don't have to observe normal forms.

4.3.2 Datasets generation

It's often necessary, in a data mining process, to step back to data preparation. Algorithms were proposed to choose relevant attributes among large data sources. However, sometimes results don't satisfy user expectations and datasets have to be built again to run new tests. KEOPS methodology suggests using the ontology in order to describe domain values and relationships between these values. Consequently, various datasets could be generated according to expert user choices.

The ontology driven information system allows choosing all data preparation strategies providing various datasets from the same source values. A dataset is built using the following operators:

- Traditional relational algebra operators: projection, selection and join.
- Data granularity: this operator allows choosing, among ontology, concepts which will be in the mining oriented database.

In order to generate datasets we developed software whose inputs are MODB and user parameters and outputs are new datasets. The user can graphically select relational algebra operator and data granularity. While database attributes and values are also ontology concepts KEOPS methodology and KEOPS software make easier the data preparation task.

4.4 Evaluation

This step assesses to what extent models meet the business objectives and seeks to determine if there is some business reason why these models are deficient. Furthermore, algorithms may generate lots of models according to parameters chosen for the extraction. That's why *evaluation* is an important task in KEOPS methodology in order to select the most interesting models according to expert knowledge.

4.4.1 Rule volume reduction

We choose an association rule extraction algorithm which can generate bases containing only minimal non-redundant rules without information loss. Then, these rules are filtered to suppress semantic redundancies. KEOPS methodology is based on Srikant's *generalized association rules* definition (Srikant and Agrawal, 1995). These rules are minimal because they forbid all irrelevant relationships within their items. We give a formal definition below:

Let $\mathcal{T}$ be a taxonomy of items. $R : A \rightarrow C$ is called *generalized association rule* if:

- $A \subset \mathcal{T}$
- $C \subset \mathcal{T}$
- No item in C is an ancestor of any item in A or C
- No item in A is an ancestor of any item in A

Consequently relationships appearing within these rules are semantic and generalization relationships from C items to A items. The objective is to maximize information level in minimal rules. The last step consists of replacing a set of these rules by a more generalized one.

Kind of knowledge	Rule R informative level		
	More than K	Similar	Less than K
Obvious	weak	none	none
Consensual	medium	weak	weak
Assumption	**strong**	medium	medium

Table 2: Interestingness measure if confidence levels are similar

4.4.2 Rule interestingness evaluation

KEOPS methodology suggests comparing extracted rules with expert's knowledge. Extracted rules having one or more items that are in relationship with some knowledge rules items (i.e. value concepts are linked in the ontology) have to be selected. Then, for each pair knowledge rule/extracted rule:

- Extracted rule antecedant coverage is compared to knowledge rule antecedent coverage, then extracted rule consequent coverage is compared to knowledge rule consequent coverage.
- By coverage comparison the most informative rule is deduced, i.e. rule predicting the largest consequence from the smallest condition.
- IMAK interestingness measure is applied (Brisson, 2007). This measure evaluates rule quality considering relative confidence values, relative information levels and knowledge certainty (see section 4.1).

Thus, ontology driven information systems are useful in order to formalize domain concepts, to express knowledge, to generate models and to facilitate knowledge and models ontology-based comparison.

Example Let us assume that a domain expert makes the following assumption: *"If a student wants to buy a book about JAVA he comes to an academic bookshop.'* and gives it a 60%-80% estimation of confidence. Let us assume that the extracted rule is slightly different because it says that *"Every young customer buying a book about J2EE comes to an academic bookshop"* and has 75% confidence.

Assumption K book='JAVA' $\wedge$ buyer='student' $\rightarrow$ bookshop='Academic'

Extracted Rule R book='J2EE' $\wedge$ buyer='youngs' $\rightarrow$ bookshop='Academic'

According KEOPS methodology these two rules are said to be comparable because at least one extracted rule item is in relationship with a knowledge rule item: 'youngs' is more general than 'student' and 'JAVA' is more general than 'J2EE'. Then, the algorithm compares the coverage of these two rules in order to evaluate the more informative one. Let's make the assumption that R is more informative than K. Since these two rules have similar confidence we can use table 2 in order to evaluate extracted rule interestingness (similar tables for various confidence levels are presented in (Brisson, 2007)). While the knowledge is an assumption, the interestingness degree of the extracted rule is **strong**.

5 Experiments

Although we illustrated in this paper the KEOPS methodology with bookstore example, we run experiments on real data provided by French Family Allowance Office (CAF: Caisses d'allocations familiales). In this section we don't express some specific rules about allowance beneficiaries behavior (because of privacy) but only extracted rules reliability. These results show we are able to select relevant rules to provide to experts for final human evaluation.

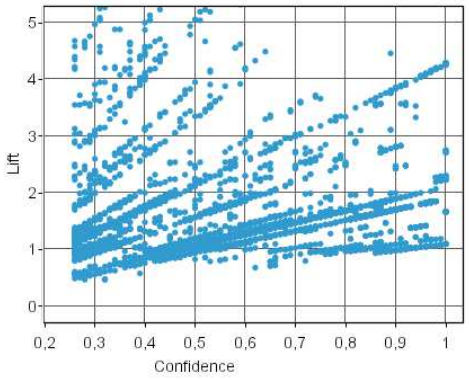

Figure 5: Confidence vs Lift of all of the extracted rules

CAF data were extracted during 2004 in the town of Grenoble (France). Each row describes one contact between the office and a beneficiary with 15 attributes and data about 443716 contacts were provided. We ran CLOSE algorithm and extracted 4404 association

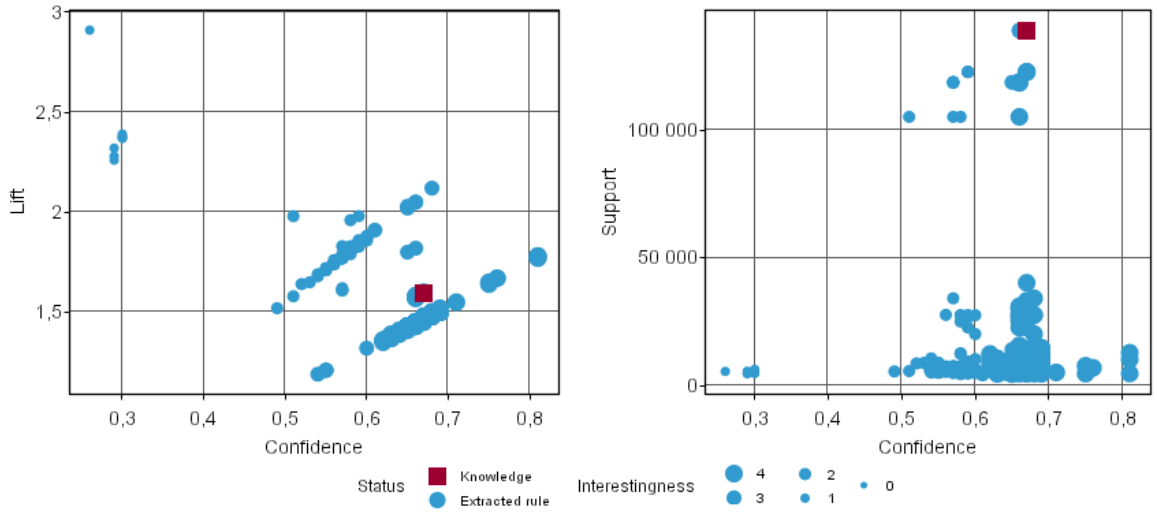

Figure 3: Extracted rules (dots) matching knowledge rule 335 (square) (IMAK interestingness value increase with dot size) a) Confidence vs Lift - b) Confidence vs Support

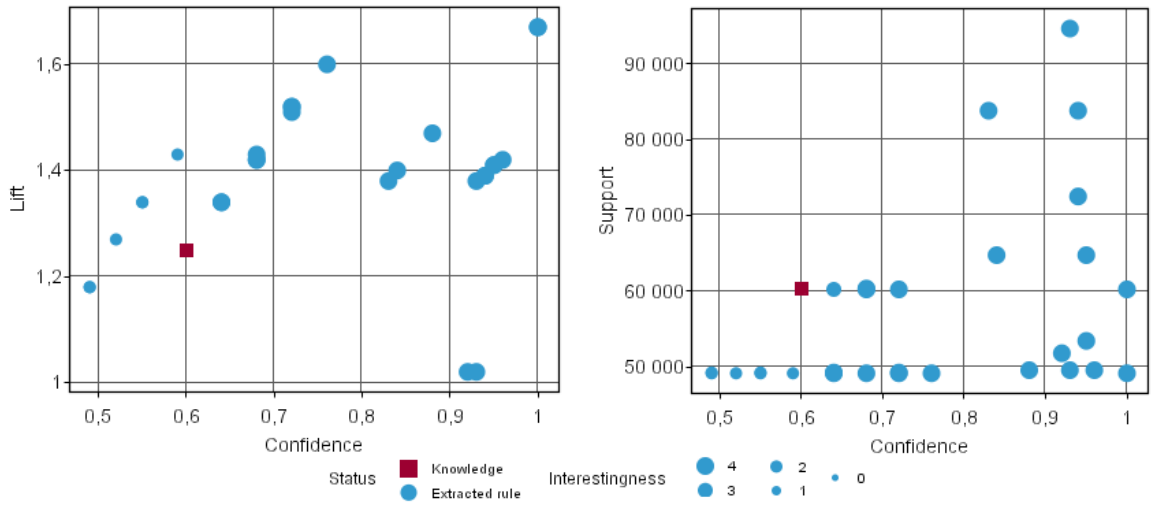

Figure 4: Extracted rules (dots) matching knowledge rule 565 (square) (IMAK interestingness value increase with dot size) a) Confidence vs Lift - b) Confidence vs Support

rules. The interestingness measure, IMAK, helps to filter the best ones. Figure 5 plots 4404 rules according to confidence and lift.

Experiments illustrated by figure 3 and 4 compare these rules to a specific knowledge. We may observe that among all of the extracted rules only few of them are selected. Selection condition is to match the knowledge and to have an interestingness value greater than 0. In these figures interestingness value is illustrated by the dot size.

In figure 3 lift of selected rules is greater than 1 and often greater than knowledge lift (lift equals 1 at independency). Furthermore, some extracted rules have a better confidence but a smaller support: they illustrated the discovery of rare events which could be very interesting for expert users.

Figure 4 shows some results for another specific knowledge. We may observe again that only few rules are selected. These rules offer various tradeoff to select rare events (low support and high confidence) or general rules (high support and good confidence) to provide to domain experts.

As future work, we plan to evaluate rules selected by KEOPS software with the help of some expert groups who are able to validate their semantic relevance.

6 Conclusion

Managing domain knowledge during the data mining process is currently an important research issue in the data mining field. In this paper, we presented the so-called KEOPS methodology for integrating expert knowledge all along the data mining process in a coherent and uniform manner.

We built an ontology driven information system (ODIS) based on an application ontology, a knowledge base and a mining oriented database rebuilt from source raw data. Thus, expert knowledge is used during business and data understanding, data preparation and model evaluation steps. We show that integrating expert knowledge during the first step, gives experts a best insight upon the whole data mining process. In the last step we introduced IMAK, a part-way interestingness measure that integrates both objective and subjective criteria in order to evaluate models relevance according to expert knowledge.

We developed KEOPS software in order to run experiments. Experimental results show that IMAK measure helps to select a reduced rule set among data mining results. These rules offer various tradeoff allowing experts to select rare events or more general rules which are relevant according to their knowledge.

REFERENCES

Becker, H. S. (1976). *Sociological Work: Method and Substance*. Transaction Publishers, U. S.

Berka, P. and Bruha, I. (1998). Discretization and grouping: Preprocessing steps for data mining. In *PKDD*, pages 239–245.

Brisson, L. (2007). Knowledge extraction using a conceptual information system (excis). In *Ontologies-Based Databases and Information Systems*, volume 4623 of *Lecture notes in computer science*, pages 119 – 134, Berlin, Heidelberg. Springer.

Ceri, S. and Fraternali, P. (1997). *Designing Database Applications with Objects and Rules: The IDEA Methodology*. Series on Database Systems and Applications. Addison Wesley.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide. In *SPSS Inc*.

De Leenheer, P. and de Moor, A. (2005). Context-driven disambiguation in ontology elicitation. In Shvaiko, P. and Euzenat, J., editors, *Context and Ontologies: Theory, Practice and Applications*, pages 17–24, Pittsburgh, Pennsylvania. AAAI, AAAI Press.

Guarino, N. (1998). *Formal Ontology in Information Systems*. IOS Press, Amsterdam, The Netherlands. Amended version of previous one in Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy.

Guarino, N., Masolo, C., and Vetere, G. (1998). Ontoseek: Using large linguistic ontologies for gathering information resources from the web. Technical report, LADSEB-CNR.

Kedad, Z. and Métais, E. (2002). Ontology-based data cleaning. In *NLDB '02: Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers*, pages 137–149, London, UK. Springer-Verlag.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *Knowl. Eng. Rev.*, 20(1):39–61.

Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999). Closed set based discovery of small covers for association rules. In *Actes des 15mes journes Bases de Donnes Avances (BDA'99)*, pages 361–381.

Penarrubia, A., Fernandez-Caballero, A., Gonzalez, P., Botella, F., Grau, A., and Martinez, O. (2004). Ontology-based interface adaptivity in web-based learning systems. In *ICALT '04: Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04)*, pages 435–439, Washington, DC, USA. IEEE Computer Society.

Srikant, R. and Agrawal, R. (1995). Mining generalized association rules. In *VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases*, pages 407–419, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.