Semantic Business Analysis Model
Considering Association Rules Mining

Anna Rozeva[1], Boryana Deliyska[1] and Roumiana Tsankova[2]
[1]*Department of Business Management, University of Forestry, 10 Kliment Ohridski blv., Sofia, Bulgaria*
[2]*Department of Management, Technical University, 8 Kliment Ohridski blv.Sofia, Bulgaria*
arozeva@hotmail.com, delijska@gmail.com, rts@tu-sofia.bg

Keywords: Business Analysis, Knowledge Discovery in Databases, Association Rules, Domain Ontology, Semantic Model, Ontology Reasoning.

Abstract: Deriving models for intelligent business analysis by generation of knowledge through data mining techniques has proved to be highly theoretically researched and practically implemented topic in the field of decision support and business intelligence systems in the last decade. A general data mining task concerns discovery and description of relationships among items recorded in business transactions. The model of association rules is the one most implemented for revealing such relationships. In order to increase the decision support value of the output associative models the necessity for capturing and involving semantics from the domain of discourse has emerged. Ontologies represent the tool for structuring the concepts and their relationships as knowledge for a subject area that was established with the growth of the Semantic Web. The paper is intended to design a framework for implementing ontologies in the association rule analysis model that provides for involving semantics in the extracted rules by means of initial verification and optimization of the mining task by database scheme ontology and exploration of rules' interestingness by ontology reasoning process.

1 INTRODUCTION

Modern business analysis is inevitably information based. Therefore it faces the problem of dealing with continuously growing amounts of structured or unstructured data from a variety of sources. The main challenge consists in getting the "big picture" out of it for the sake of best serving the decision making. The general technology for processing data resulting in deriving summarized models is data mining (Larose, 2006). Models contain knowledge about a related domain. Mining tasks perform mainly classification, clustering or extraction of associations. The model describing associations between items on the basis of their mutual occurrence in transactions has proven to be of particular value for business analysis. It's represented by rules relating certain items X and Y with assigned support and confidence (Maragatham and Lakshmi, 2012). These rules represent new knowledge and are derived by processing raw data from transactions by data mining techniques which is also often referred to as knowledge discovery in databases (KDD) (Frawley et al, 1992). Models extracted by summarizing significant amounts of data are related to instances of objects from the domain of discourse and hence they lack the abstractness that is inherent to models in general.

At the same time knowledge about domains exist which is accumulated, stored for being used and shared through the resources of the Semantic Web. The knowledge represents conceptualization that articulates abstractions of certain state in reality (Guizzardi, 2007). The tool for its engineering is referred to as ontology. Obviously ontologies capture the domain semantics. The task is to map ontologies to extracted analytical models for verifying their correctness and for inferring new knowledge. "The role to be played by ontologies in KDD (and even their mere usability) depends on the given mining task and method, on the stage of the KDD process, and also on some characteristics of the domain and dataset" (Svatek and Rauch, 2006, p. 163). We propose a framework for implementation of domain ontologies in association rule mining with the goal for mining task verification on the database scheme, extracted rules conceptualization and further refinement through domain ontologies.

Rozeva A., Deliyska B. and Tsankova R.
Semantic Business Analysis ModelConsidering Association Rules Mining.
DOI: 10.5220/0004461801220127
In *Proceedings of the Second International Symposium on Business Modeling and Software Design (BMSD 2012)*, pages 122-127
ISBN: 978-989-8565-26-6

The remainder of the paper is organized as follows: The second section is a review on approaches for application of ontologies in mining databases for association rules extraction and results obtained. The third section presents a framework for extracting semantic association rules analysis model for knowledge generation from a database by ontology reasonong. The fourth section presents application results on sample database and ontology instantiations. The last section concludes with discussion on the effect for the semantic enrichment of the business analysis model.

2 ASSOCIATION RULES AND ONTOLOGIES

Association rule as defined by (Agrawal et al., 1993) is a triple (*I, S, C*), where I is an implication of the form $X \rightarrow Y$ denoting *if X then also Y*, *S* and *C* are interestingness measures for support and confidence. *X* and *Y* are items in database transactions and the rule correlates the presence of both sets of items in transactions, i.e. transactions that contain items of *X* tend to contain items of *Y* as well. S indicates the statistical significance indicating the percent of transactions that contain items of both *X* and *Y*. *C* measures rule's strength as the probability of their mutual occurrence. The extracted association rules have support and confidence greater than the predefined ones.

Ontology is represented in (Maedche, 2002) as 5-tuple of the form (C, R, H^C, F, A), where C is a set of concepts, R is a set of nonhierarchical relations among concepts, H^C is taxonomy of the concept hierarchy that defines relations among concepts c_1 and c_2 of type 'is-a' and 'has-a' mainly. F is a function that instantiates the relationships from R and A is a set of axioms that describes constraints. The definition is a formal description of the concepts and their hierarchical relationships in a specific domain as a piece of reality. Further on it's instantiated for an element of the domain by application ontology. Task ontology is designed and implemented with the purpose of modelling the knowledge for solving specific task within the application as shown in (Deliyska and Manoilov, 2012).

The ontology support of the KDD process has been within the scope of a lot of studies recently. In (Gottgtroy, Kasabov and MacDonell, 2004) a general framework for the mutual interdependence of ontology building and maintenance and the knowledge discovery is suggested. It's argued that ontologies facilitate each stage of the knowledge discovery process from improving the quality of the source data, feature selection by navigation through hierarchy and finally production of improved results by reasoning within ontology's links and relationships.

Framework for implementation of domain knowledge into association rules generation is proposed in (Antunes, 2008). It provides for formulation of constraints that control the mining process by using domain ontology. Two constraint types are defined, i.e. interestingness and content. While the interestingness measure refers to quantitative conditions on the frequency of mutual appearance of the items in transactions, the content constraint considers characteristics of the items that are present in the domain ontology. They are qualified as taxonomical when based on restriction among concepts, defined in the ontology taxonomy. Non-taxonomical constraints are referred to as relational and they are based on the ontology relations among concepts. The constraints guide the knowledge discovery process, providing the desired level of abstraction.

The framework presented in (Bellandi et al., 2007) provides for the extraction of constraint-based multilevel association rules with ontology support. The constraints for the mining process are defined from domain ontology as domain specific. The ontology is used for filtering the transaction instances sourcing the mining process. The system architecture involves interpretation module translating user constraints and passing them to an ontology query engine for excluding non-interesting rules and for presenting the interesting ones at the relevant abstraction level. It's stated that this approach improves association rules support and provides for decreasing the amount of useless rules discovered when source data are sparse.

While the frameworks discussed so far address the input of the mining process, the one proposed by (Marinica and Guillet, 2010) considers the association rules post mining phase with the aim to decrease the number of delivered rules so that they are useful and understandable for the user. An approach for pruning and filtering the discovered rules is designed. Ontologies are used for the integration of user knowledge at the post processing stage. Besides this the quality of discovered rules can be validated at different points in an interactive process by the domain expert.

The notion of multidimensional association rules has been introduced in (Wu et. al., 2007). The

definition refers to the star scheme of the data warehouse as source of the transactions for the mining process. The approach is focused on the stored data and aims to overcome the lack in their structural and semantic exploration. It proposes functions for the effective maintenance of the discovered knowledge. The stated problems are solved by designing two types of ontologies. The scheme ontology contains the warehouse metadata. The domain ontology constructs the domain knowledge for the mining subject as conceptual layer and relationships among the related concepts. They are implemented at loading data in the warehouse. It's pointed out that by this approach minimization of data mining searching boundaries and prevention of repeated mining are achieved. On the other hand by extending the association rule mining to items from the domain ontology generalization of items to concepts with richer semantics is achieved.

In our previous work (Rozeva et. al., 2011) we've designed a framework for generation of knowledge models from text documents which consisted of structure and knowledge models. Current work extends the categorization knowledge model presented there by exploring the association rules model. It is designed on mining set containing transaction items and acquires business semantics by reference to ontologies. The results obtained will support its involvement in mining a text document corpus. The related work review presented is the background for designing a semantic analysis model.

3 SEMANTIC ANALYSIS MODEL DESIGN

The proposed framework implements ontological reference both in the step of setting up the input to the association rules mining task and extending the value of extracted patterns. The functionality implemented in the task definition step provides for the optimization analysis of mining task parameters. It examines the input and predictable item sets and performs reasoning on designed database scheme ontology for ensuring non-redundant rule generation. On the basis of exploring key-based and hierarchical dependencies included in the scheme ontology, both the input and predictable item sets will be optimized. At the evaluation stage obtained rules are explored by reasoning on provided domain ontologies. The goal is to put the focus on the interesting rules. Such rules are considered the ones with items belonging to different domains. The proposed architecture is shown in Figure 1.

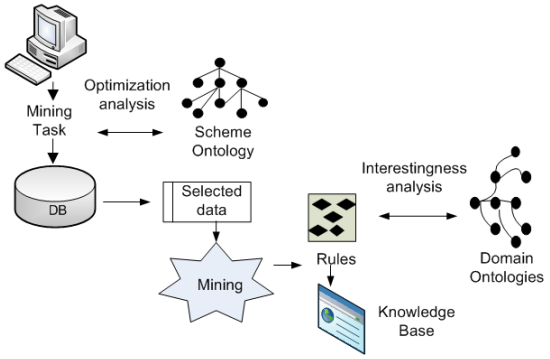

Figure 1: Framework of semantic analysis model.

The ontological reference at the front and back ends of the association rule generation process provides for the reduction of the number of rules obtained and for enhancing their value for business analysis purposes.

3.1 Scheme Ontology Design

The scheme ontology contains metadata of the database scheme. An excerpt of the designed scheme ontology is shown in Figure 2.

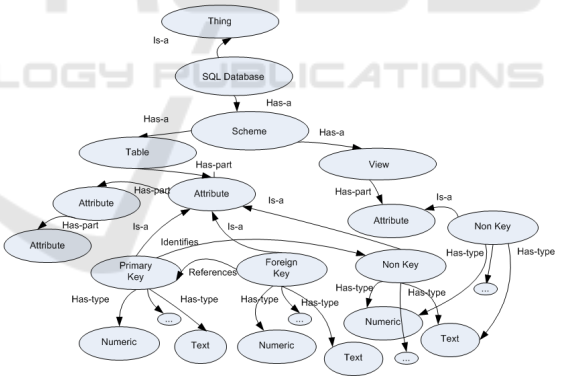

Figure 2: Database scheme ontology.

The top ontology concept *SQL Database* has subconcept *Scheme* and its subconcepts are the database objects *Table* and *View*. Their concept relationships to the ontology root are 'Has-a'. The concepts *Table* and *View* have subconcept *Attribute*. It is related to the superconcepts by 'Has-part' relationship. The *Attribute* concept which is subsumed by the *Table* concept has three subconcepts, i.e. *Primary key*, *Foreign key* and *Non-key*, related to it with 'Is-a' relationships. The *Attribute* subconcept of the *View* concept has just a

Non-key subconcept. The primary, foreign and non-key concepts have properties which are name and value type. Value types are: *numeric, text, date*, etc., which have 'Is-of-type/Has-type' relationships to the *Attribute* concept. The *Attribute* concept which is subsumed by the *Table* concept may represent hierarchy with levels defined by *Attribute* concept. The relationships between the hierarchy levels are of type 'Has-part'. For shortness concept instances are not shown in the scheme ontology.

A mining task MT for association rules specifies database tables, views or table and view related with many-to-one relationship; input and predictable attribute(s). The MT query-like representation adapted from (Wu et al., 2007) is:

```
Mine Association Rules
InputSet {IAttr1, IAttr2,…}
PredictSet {PAttr1, PAttr2, …}
From CaseTable Inner Join NestedTable
With MinSup%, MinConf%
```

The aim of the MT query optimization analysis is to identify input attributes which are functionally dependent. The dependency type is specified in the scheme ontology as being either on the primary key or between levels in a hierarchy.

3.2 Ontological Optimization of Mining Query Definition

The optimization of MT definition is proposed to be performed by reasoning on the database scheme ontology. A description logic reasoning tool Pellet is described in (Sirin, E., Parsia, B. et al., 2007). Reasoning concerns finding implicit facts in the ontology on the basis of explicitly stated facts. Basic reasoning tasks refer to proving satisfiability of a concept, subsumption of concepts, check an individual as instance of a concept, retrieving all the individuals that are instances of a concept and finding all the concepts an individual belongs to. Answering queries over ontology classes and instances for finding more general/ specific classes and retrieving the individuals matching it is a basic service performed in ontology reasoning. The reasoning tasks for checking MT definition retrieve the individuals of the *Attribute* concept and their descriptions and axioms. Mining query parameters are checked to match some of the retrieved individuals. Further on parameters are checked for consistency against the descriptions and constraints of the examined concept. Consistencies on primary key and hierarchical inclusion are considered. If the parameter set is consistent on the examined descriptions then the dependent parameter has to be removed from the mining query item set. For shortness parameter set with 2 items is considered. Further on is the specification of the ontology reasoning process:

```
InputParameterSet: {IAttr1, IAttr2}, SchemeOntology;
Concept→'Attribute';
Retrieve individuals of 'Attribute' → ABox;
Retrieve descriptions for ABox → TBox;
Check IAttr1∈ABox,
Check IAttr2∈ABox;
Check ∃ 'Identifies'.IAttr1,IAttr2 ⇒ Outputset {IAttr1};
Check ∃ 'Has-part'.IAttr1⊃IAttr2 ⇒Outputset {IAttr1}.
```

The optimization of mining query definition performed by ontological reasoning process decreases model's size and training time by preventing the generation of redundant association rules.

3.3 Analysis of Rules Interestingness

The interestingness analysis of mined association rules has been adapted from (Marinica and Guillet, 2010). We propose to perform it by reasoning on domain ontologies. The analysis is targeted at:

- Conceptualization / individualization along the domain ontology taxonomy;
- Filtering obvious rules, i.e. with the same subsuming concept.

The first task provides for the generalization / specialization of the left side (condition) and the right side (consequent) items of the extracted rules by implementing the subsumption or individual retrieval reasoning operations on the domain ontology. This analysis can be applied when the condition and consequent items refer to the same or to different domain ontologies. The second task aims at focusing on non-obvious rules. Obvious are considered rules with condition and consequent items having the same subsuming concept. The rule involving such items represents association between items from a common domain. The semantic value that is added by the analysis consists in revealing associations between items from different domains. The associations are considered more interesting when the condition and the consequent domains differ to the greatest extent possible. The measure

for the differentiation is the number of subsumptions that are to be performed for reaching their common subsuming concept as in the following reasoning process:

```
InputRuleSet: R, DomainOntology: DO;
R:{Condition, Consequent};
DO:Statement;
[a, b rdf:Statement;
rdf:subject:s;
rdf:predicate:p;
rdf:object:o]
Condition:a.o ∩ Consequent:a.o
⇒ {Condition, Consequent}⊂ a.s;
R → Non interesting;
Condition:a.o;
Consequent:b:o;
⇒ {Condition, Consequent} ⊄ a.s;
R → Interesting.
```

4 SEMANTIC MODEL IMPLEMENTATION

The proposed approach for semantic association rule model design has been implemented on purchase transaction tables from the sample database (Microsoft SQL Server Database Product Samples, 2012). The database scheme ontology from Figure 2 has been filled in with the corresponding instances. Administrative location and product domain ontologies have been designed. The MT queries that have been defined and the result from ontological reasoning analysis are presented in Table 1.

Association rules model has been trained with the Apriori algorithm with input and predictable parameters CustId and ProdId. The minimum probability was set to 10% and minimum support to 1%. The number of generated rules without and with optimization as shown in the last column is approximately 50%. By varying the support and probability values the rule number will be different.

Enhancement of model interestingness is performed by filtering the non-obvious rules. They are identified as belonging to different subsuming concepts. The case of rules that involve single condition item is considered.

Table 1: Mining task optimization.

MT Input	MT Predict	Optimization	Rules
CustId, CustName	ProdId	CustId, ProdId	158→49
CustId, Country	ProdId	CustId, ProdId	111→51

The task is performed by extraction of mined rules from the model as RDF triples {id, condition, consequent} from the tree-like structure shown in Figure 3.

ATTRIBUTE_NAME	ATTRIBUTE_VALUE	SUPPORT	PROBABILITY
Customer Key	11223	8	0,00053767054237516
v Assoc Seq Line Items(Sport-100)	Existing	8	0,00053767054237516

Figure 3: Model's node content.

The tree nodes are from the following types: with condition item only, with consequent item only and with both condition and consequent item values. Each node has the respective probability and support values attached. The nodes with both condition and consequent items are filtered and the RDF triples are obtained. The extracted subsuming concepts of the rule items are compared and the rule is either kept or discarded. An initial model with ProdId as input and predictable parameter and resulting predicted associations is shown in Figure 4.

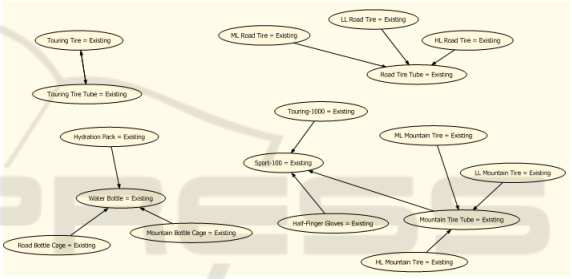

Figure 4: Initial model rules.

By applying interestingness reasoning on the Product domain ontology where the Product concept is subsumed by the Category concept the model shown in Figure 5 has been obtained. The rules which remained after performing the analysis display associations between ontologically remote items only.

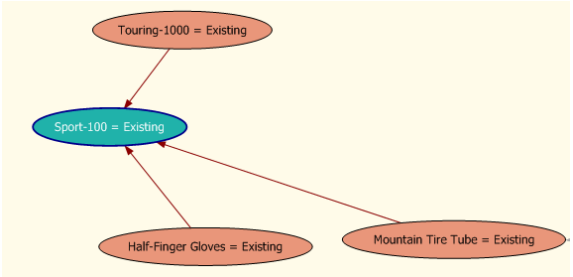

Figure 5: Interesting rules.

By subsumption operation on the RDF triples the rules can be further ontologically generalized.

5 CONCLUSIONS

The analysis model for extraction of associations between items from business transactions stored in a database presented in the paper introduces an innovative approach for capturing and using domain knowledge. It is intended to filling the gap between definition of the analysis task and the interpretation of obtained results and the examined business domain. Ontologies have been recognized and widely adopted as model for capturing this background knowledge. The proposed framework for designing semantic analysis association rules model implements two types of ontologies that provide background knowledge for the data source structure and the domain of discourse. The ontology content is made use of by means of reasoning process based on description logic. The reasoning on the data source ontology provides for the support and optimization of mining task definition. Key dependent and hierarchy related query parameters are identified by the reasoning process and discarded for the sake of generating non-redundant set of rules. Domain ontology reasoning is implemented for tuning rule interestingness. Interesting rules are considered those involving items from different domains. Reasoning process procedures have been presented. The proposed methodology has been evaluated on sample transaction database with reasoning on instantiated structure and domain ontologies.

Future work is intended in refining the reasoning process in order to be applied further on to mining associations between terms extracted from text document corpus with available ontology referring to e-Governance services. Application of the designed framework in automatic generation of ontologies will be researched as well.

ACKNOWLEDGEMENTS

The paper presents results of the project “Research and Education Centre for e-Governance” funded by the Ministry of Education in Bulgaria.

REFERENCES

Agrawal, R., Imielinsky, T. and Swami, A., 1993, Mining Association Rules between Sets of Items in Large Databases., In *Proc. ACM SIGMOD Conf. on Management of Data*, p.207-216.

Antunes, C., 2008, An Ontology-Based Framework For Mining Patterns In The Presence Of Background Knowledge, In *Proceedings of International conference on advanced intelligence (ICAI 2008)*, Post and Telecom Press, Baijing, China, p.163-168.

Bellandi, A., Furletti, B., Grossi, V. and Romei, A., 2007, Ontology-driven association rule extraction: A case study, In *Proceedings Workshop Contexts and Ontologies: Representation and reasoning in 16th European conference on artificial intelligence*, p.1-10.

Deliyska, B., Manoilov, P., 2012, Ontologies in intelligent learning systems, In Jin, Q. (ed.), *Intelligent learning systems and advancements in computer-aided interaction*, Information science reference. An imprint of IGI Global, p.31-48.

Gottgtroy, P., Kasabov, N. and MacDonell, S., 2004, An ontology driven approach for knowledge discovery in biomedicine, In *Proc. 8th Pacific Rim Int’l Conf. on Artificial Intelligence*.

Guizardi, G., 2007. On ontology, ontologies, conceptualizations, modelling languages and (meta) models. In Vasilecas, O. (eds.), In *Frontiers in artificial intelligence and applications, databases and information systems IV*, Amsterdam, The Nederlands: IOS Press, p.18-39.

Frawley, W., Piatetsky-Shapiro, G., and Matheus, C., 1992, Knowledge Discovery in Databases: An Overview. *AI Magazine*, Vol. 13, p.57-70.

Larose, D., 2006. *Data mining methods and models,* Wiley-IEEE Press.

Maedche, A., 2002, *Ontology Learning for the Semantic Web*, Kluwer Academic Publishers.

Maragatham, G., Lakshmi, M., 2012, A recent review on association rule mining, *Indian journal of computer science and engineering*, Vol. 2(6), p.831-836

Marinica, C. and Guillet, F., 2010, Knowledge-Based Interactive Postmining of Association Rules Using Ontologies, *IEEE Transactions On Knowledge And Data Engineering*, Vol. 22, No. 6, p. 784-797.

Microsoft SQL Server Database Product Samples, *AdventureWorks2008_SR4*, http://msftdbprodsamples.codeplex.com/releases/view/37109, Retrieved March 24th 2012.

Rozeva, A., Ivanov, M. and Tsankova, R., 2011, Business modelling for generation of knowledge from explicit data, In *Proceedings of First International Symposium on Business Modeling and Software Design (B.Shishkov, ed.),* Sofia, Bulgaria, p.114-121

Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A. and Katz, Y., 2007, Pellet: A Practical OWL-DL Reasoner, *Web Semantics*, 5 (2), p.51-53

Svatek V., Rauch J., 2005, Ontology-Enhanced Association Mining. In: *Semantics, Web and Mining, Joint International Workshops, EWMF 2005and KDO 2005*, p.163–179.

Wu, C.-A., Lin, W.-Y., Tseng, M.-C. and Wu, C.-C., 2007, Ontology-Incorporated Mining of Association Rules in Data Warehouse, *Journal of Internet Technology*, Vol.8, No4, p.1-9.