IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. IV (Nov – Dec. 2014), PP 39-45
www.iosrjournals.org

Ontology Based Data Analysing Approach for Actionable Knowledge Discovery

S.Antoinette Aroul Jeyanthi[1], Dr.S.Pannirselvam[2]

[1]*Department of Computer Science, Pope John Paul II College of Education, Pondicherry, India*
[2]*Department of Computer Science, Erode Arts & Science College, Erode, TamilNadu, India*

Abstract: *In Data Mining, the effectiveness of association rules is limited by the huge quantity of delivered rules. In this manuscript, we propose a new approach to prune and filter discovered rules. An interactive and iterative framework is designed to assist the user along the analyzing task. The manuscript focus on medical data test set for detailed analysis and formation of the ontologies. In the proposed approach, the data set of medical records is entered in the back – end database. The development of associations and the ontology is fully dynamic. In case, any symptom is searched, the proposed approach search from the back end database and creates the ontology that is dynamic in execution. It refers that the unsupervised approach of ontology generation is implemented so that the unbiased results can be achieved. The implementation of proposed work shows the dynamic results in terms of rules found and their effectiveness in the real world scenario.*

Keywords: *Association Rule, Domain Ontology, Knowledge Discovery, Pruning, Rule Schema,*

I. Introduction

In recent years, lot of attention has been drawn by new methodology called Domain Driven Data Mining (D^3M) which is intended to make data mining workable in supporting decision-making actions in the real world. Current data mining tools and techniques drive a paradigm shift from traditional data-centred hidden pattern mining to domain-driven actionable knowledge discovery (AKD). AKD must cater for domain knowledge and environmental factors, balance technical significance and business expectations from both objective and subjective perspectives and support automatically converting patterns into deliverables in business friendly and operable forms such as actions or rules.

In domain-driven framework, data mining analysts and domain-specific business analysts complement each other with regard to in-depth granularity and constrained environment through interactive system support. The involvement of domain experts and their knowledge can assist in developing highly effective domain specific data mining techniques, and reduce the complexity of knowledge discovery. The involvement of domain-related social intelligence into KDD process not only strengthens technical development and performance, but highlights business expectations and actionable capability of the identified results.

In Data Mining, the effectiveness of association rules is strappingly limited by the huge quantity of delivered rules. In this manuscript, we propose a new approach to prune and filter discovered rules. An association rule is described as the implication $X \rightarrow Y$ where X and Y are sets of items and $X \cap Y = \phi$. The strength of association rule mining rests in its ability to deliver interesting discovered knowledge that exists in data. Unfortunately, due to high dimensionality of massive data, this strength becomes its main weakness when analyzing the mining result. The huge number of discovered rules makes very difficult for a decision maker to manually outline the interesting rules. Thus, it is crucial to help the decision maker with an efficient reduction of the number of rules.

To overcome this drawback, the post-processing task was proposed to improve the selection of discovered rules. Different complementary post-processing methods may be used, like pruning, summarizing, grouping or visualization [1]. The pruning phase consists of removing uninteresting or redundant rules. In the summarizing phase summaries of rules are generated. Groups of rules are produces in the grouping phase; meanwhile the visualization phase is useful to have a better presentation.

However, most of existing post-processing methods are generally based on statistical information on database. Since rule interestingness strongly depends on user knowledge and goals these methods are not efficient enough. For instance, if the user looks for unexpected rules, all the already known rules should be pruned. Or, if the user wants to focus on specific schemas of rules, only this subset of rules should be selected.

This paper proposes an effective approach to prune and filter discovered rules. Using Domain Ontologies, we strengthen the integration of user knowledge in the post-processing task. Furthermore, an interactive and iterative framework is designed to assist the user along the analyzing task. On the one hand, we represent user domain knowledge using a Domain Ontology over database. On the other hand, a novel technique is suggested to prune and to filter discovered rules. User expectations are described by the notion of Rule

Schema and rule operators are proposed to guide user actions. Ontologies will offer a powerful representation of user knowledge, and rule schemas and rule operators a more expressive representation of user expectations in terms of rules.

II. Related Works

2.1 Post-Processing Techniques

Several approaches, integrating user knowledge, to solve the problem of huge number of discovered rules have been proposed. As early as 1994, in the KEFIR system [2], the key finding and deviation notions were suggested. Grouped in findings, deviations represent the difference between the actual value and the expected value.

Later, Klemettinen et al. proposed templates [3] to describe the form of interesting rules (inclusive templates), and those of not interesting rules (restrictive templates). Other approaches proposed to use a rule-like formalism to express user expectations, and the discovered rules are pruned/summarized comparing them to user expectations [4]).

Toivonen et al. proposed in [5] a novel technique for rule pruning and grouping based on rule covers. The notion of rule cover defines the subset of a rule set describing the same transaction row. Thus, the authors define the pruning action as the reduction of a rule set to its rule cover.

The notion of subsumed rules, discussed in [6], describes a set of rules having the same consequent and several additional conditions in the antecedent with respect to another rule. Bayardo Jr. et al. proposed a new pruning measure described as the difference between the confidences of the two rules, called Minimum Improvement. A rule is pruned if this measure is less than a pre-specified threshold, so the subsumed rule does not bring a lot of information comparing to the other rule.

In the Web domain, the paper [7] presents a framework for building behavioral profiles of individual users. Considering a set of discovered rules for each client, the authors propose an iterative rule validation process based on several operators, including rule grouping, filtering, browsing, and redundant rule elimination.

Another related approach is proposed by An et al. in [8] where the authors introduce domain knowledge in order to prune and summarize discovered rules. The first algorithm proposed use a data taxonomy, proposed by user, in order to describe the semantic distance between rules and to group rules. The second algorithm allow to group discovered rules sharing at least an item in the antecedent and in the consequent.

An original proposition was made in [9] with the exploitation of the directed hyper-graphs in order to prune singleton consequent rules. Thus, the discovered rules are represented in a directed hyper-graph called, after being pruned of cycles, Association Rules Network (ARN).

In 2007, a new methodology was proposed in [10] to prune and organize rules with the same consequent. The authors suggest transforming the database in an association-rulebase in order to extract second level association rules. Called metarules, the extracted rules $r1 \rightarrow r2$ express relations between the two association rules and help on pruning/grouping discovered rules.

2.2 Ontologies and Data Mining

Ontologies, introduced in data mining for the first time in early 2000, can be used in several ways [11]: Domain and Background Knowledge Ontologies, Ontologies for Data Mining Process, or Metadata Ontologies. Background Knowledge Ontologies organize domain knowledge and play important roles at several levels of knowledge discovery process. Ontologies for Data Mining Process codify mining process description and choose the most appropriate task according to the given problem; meanwhile, Metadata Ontologies describe the construction process of items.

In this study, we are interested in Domain and Background Knowledge Ontologies and we will present past studies related to them. The first idea of using Domain Ontologies was introduced by Srikant and Agrawal with the concept of Generalized Association Rules [12]. The authors proposed taxonomies of mined data (is-a hierarchy) in order to generalize/specify rules.

In [13], and developed in [14], it is suggested that an ontology of background knowledge can benefit all the phases of a KDD cycle described in CRISP-DM. The role of ontologies is based on the given mining task and method, and on data characteristics. From business understanding to deployment, the authors delivered a complete example of using ontologies in a cardiovascular risk domain.

Related to Generalized Association Rules, the notion of raising was exposed in [15]. Raising is the operation of generalizing rules (making rules more abstract) in order to increase support in keeping confidence high enough. This allows for strong rules to be discovered and also to obtain sufficient support for rules that, before raising, would not have minimum support due to the particular items they referred to. The difference with Generalized Association Rules is that this solution proposes to use a specific level for raising and mining.

III. The Proposed Framework

In the proposed methodology, the information set of therapeutic records is entered in the back – end database. The improvement of cooperations and the cosmology is completely progressive. In the event that, any indication is looked, the proposed methodology look from the back end database and makes the metaphysics that is dynamic in execution. It alludes that the unsupervised methodology of philosophy era is executed so the unprejudiced effects might be accomplished.

The new approach defines a new formal environment to prune and group discovered associations integrating knowledge into specific mining process of association rules. It is composed of three main parts (as shown in Fig.1). The proposed flow diagram is shown in Fig.2.

Firstly, a basic mining process is applied over data extracting a set of association rules. Secondly, the knowledge base allows formalizing user knowledge and goals. Domain knowledge allows a general view over user knowledge in database domain, and user expectations express user already knowledge over the discovered rules. Finally, the post-processing step consists in applying several operators (i.e. pruning) over user expectations in order to extract the interesting rules.

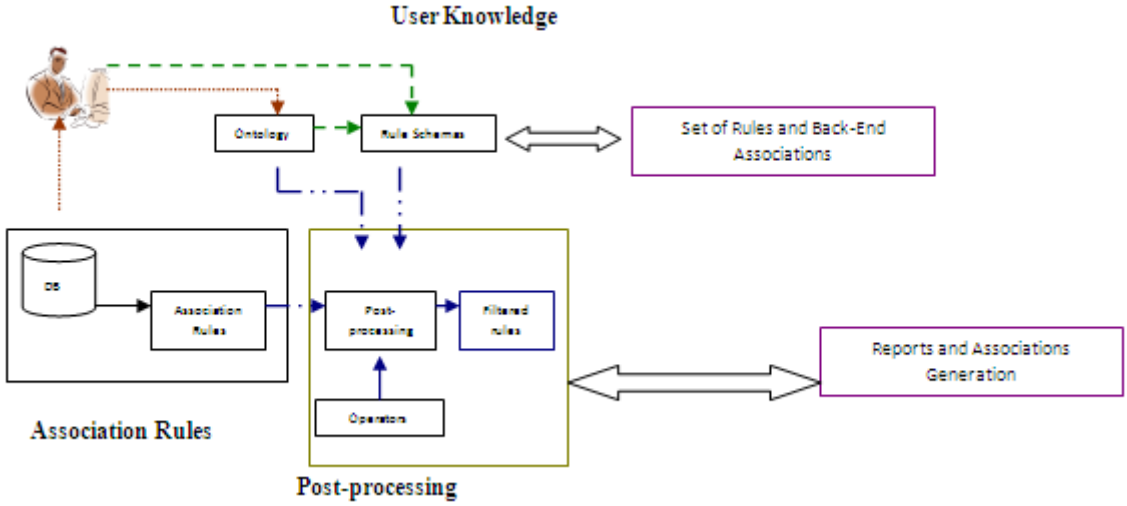

Figure 1. Proposed Framework

The novelty of this approach resides in supervising the knowledge discovery process using different conceptual structures for user knowledge representation: one or several ontologies and several rule schemas.

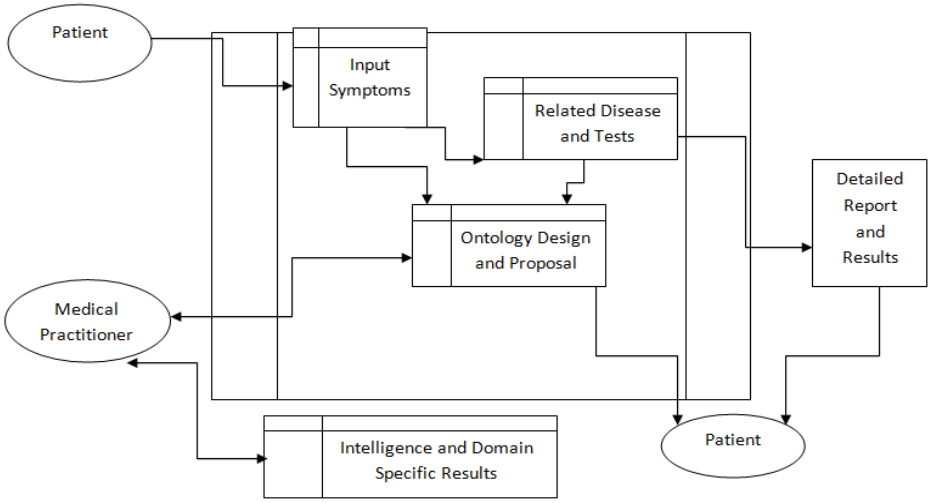

Figure 2. Proposed Flow Diagram

1.1 Database and Association Rule Mining

The association rules mining techniques are applied over databases described as D = {I, T}. Let I ={ I_1, I_2,... ,I_p } be the set of attributes (called items) and T = { t_1, t_2, … t_n } be the transaction set. Each transaction t_i = { I_1, I_2, ... ,I_{mi} } is a set of items, such as $t_i \subset I$ and each subset of items, X, is called item set.

An association rule is an implication $X \rightarrow Y$, where X and Y are two item sets and $X \cap Y = \phi$. This rule holds on D with the confidence c if c% of transactions in T that contain X, also contain Y. The rule has support s in transaction set T if s% of transactions contain $X \cup Y$.

Since their early definition, association rules are mined using Apriori algorithm proposed for the first time in Agarwal et al., 1993.

3.2 User Knowledge

In association rule mining process, user knowledge can be divided into two main types: domain knowledge, mainly related to database items, and user beliefs expressing user expectations according to the discovered knowledge. In addition, we propose a third user-based element described by the actions that a user can realize among his/her different beliefs. Thus, the operators are introduced in order to guide the post-processing step. This element will be discussed in the following section.

Ontology is described as a formal explicit specification of a shared conceptualization for a domain of interest [16].

Definition 1. Formally, ontology is a 3-tuple O = {C, R, H}. C = { C_1, C_2, ... ,C_o } is a set of concepts and R = { R_1, R_2, ... ,R_r } is a set of relations defined over concepts. H is a directed acyclic graph (DAG) over concepts defined by the subsumption relation (is-a relation, $\leq$) between concepts. We say that C_2 is-a C_1, $C_2 \leq C_1$, if the concept C_1 subsumes the concept C_2.

In this approach, we propose a domain knowledge model based on ontologies connecting ontology concepts to a set of database items. Consequently, domain ontologies over database extend the notion of Generalized Association Rules based on taxonomies as a result of the generalization of the subsumption relation by the set R of ontology relations. Besides, ontologies are used as filters over items, generating item families.

The following algorithm is used for Ontology –Database (OntoDB) Mapping .

1. Activate the Super Data Set M
2. Initialize and Activate the Data Item Set DI with Tuples T_j (j<=n) where n refers to the max. tuples or records in the test medical data items set. It refers to the cardinality of the data set.
3. Set multiple parameters S->Symptom | T-> Tests | D-> Diseases
4. K <-> IBox (Input Box) for Search
5. RSi => Recordset fetched based on the Keyword K in the Input Box for Ontology Determination
6. RS [n] <= T [i] (All Recordsets are subset of Key Database Relation or DataSet)
7. D_m -> TS_w -> RM_i

a. DataSet D is associated with respective Tests Set TS in the Medical Data Set Ti
b. Tests associated with the Remedies RM in the Medical Data Set

8. FO (Final Ontology) -> Generated from the associativity of multiple dimensions as aspects -> D -> TS -> RM
9. Verify and Authenticate the Results
10. If (Results -> Successful)

a. Terminate and Stop

11. Else

Go To Step 1

In this scenario, it is fundamental to connect the ontology to the database, each concept and each instance being instantiated in one/several items.

Considering that the set of concepts C is defined as the union of three concepts subsets $C = C_0 \cup C_1 \cup C_2$:

- C_0 is defined as the set of leaf-concepts of the ontology connected in the easiest way to database.

$$C_0 = \left\{ c_0 \in C \middle| \nexists c' \in C,\ c' \leq c_0 \right\}$$

In this manner, each concept from C_0 is associated to an item in the database.

$$f_0 : C_0 \rightarrow I$$

$$\forall c_0 \in C_0, i \in I, i = f_0(c_0)$$

- C_1 is described as the set of generalized concepts in the ontology. A generalized concept is connected to database through its subsumed concepts. That means that, recursively, only the leaf-concepts subsumed by a generalized concept contribute to its database connection.

$$f : C_1 \rightarrow 2^I$$

$$\forall c \in C_1,\ f(c) = \left\{ i = f_0(c_0) \middle| c_0 \in C_0, c_0 \leq c \right\}$$

- More generally, we propose the definition of ontology concepts by logical expressions defined over items, organized in the C_2 subset. In a first attempt, we base the description of the logical expression on the OR logical operator. Thus the defined concept associated could be connected to a disjunction of items.

$$f : C_2 \to 2^I, \forall c \in C_2$$
$$c \to E(c)$$
$$f(c) = \{ f(c') | c' \in E(c) \}$$

To improve association rule selection, we propose a rule filtering model, called Rule Schemas. In other words, a rule schema describes, in a rule-like formalism, the user expectations in terms of interesting/obvious rules. As a result, Rule Schemas act as a rule grouping, defining rule families.

The base of Rule Schema formalism is the user representation model introduced by Liu et al. in [17] composed of: General Impressions, Reasonably Precise Concepts and Precise Knowledge. The proposed model is described using elements from an attribute taxonomy allowing an is-a organization of database attributes.

A Rule Schema is a semantic extension of the Liu model since it is described using concepts from the domain ontology. We propose to develop two of the three representations introduced in [17].General Impressions and Reasonably Precise Concepts. Thus, rule schemas bring the complexity of ontologies in rule mining combining not only item constraints, but also ontology concept constraints.

Definition 2. A rule schema is defined as:

$$\langle X_1, X_2, \ldots, X_{s1} \ (\to) \ Y_1, Y_2, \ldots, Y_{s2} \rangle$$

where X_i and Y_j are ontology concepts and the implication "$\to$" is optional. In other words, we can note that the proposed formalism combines General Impressions and Reasonably Precise Concepts. Consequently, if we use the formalism as an implication, an implicative rule schema is defined extending the Reasonably Precise Concepts. Meanwhile, if we do not keep the implication, we define non implicative rules schemas, generalizing General Impressions.

For example, a rule schema $C_2, \overline{C_3} \to C_4$ corresponds to "all association rules whose condition verifies C_2 and doesn't verify the concept C_3, and whom conclusion verifies C_4".

3.3 Operations in Post-Processing Step

The post-processing task that we design is based on operators applied over rule schemas allowing to user to perform several actions over the discovered rules. We propose two important operators: pruning and filtering association rules. The filtering operator is composed by three operators: conforming, unexpectedness and exception.

These four operators will be presented along this section. To this end, let us consider an implicative rule schema $RS_1 : \langle X \to Y \rangle$, a non implicative rule schema $RS_2 : \langle U, V \rangle$ and an association rule $AR_1 : A \to B$ where X, Y, U, V are ontology concepts, and A, B are item sets.

The pruning operator allows to user to remove families of rules that he/she considers that are uninteresting. In a database, there exist, in most of cases, relations between items that we consider obvious or that we already know. Thus, it is not useful to find these relations among the discovered associations. The pruning operator applied over a rule schema, P(RS), eliminates all association rules matching the rule schema. To extract all the rules matching a rule schema the conforming operator is used.

The conforming operator applied over a rule schema, C(RS), proposes to confirm an implication or to find the implication between several concepts. As a result, rules matching all the elements of a non-implicative rule schema are filtered. For an implicative rule schema, the condition and the conclusion of the association rule should match those of the schema.

The rule AR_1 is selected by the operator $C(RS_1)$ if both the condition and the conclusion of the rule AR_1 respectively match the condition and the conclusion of RS_1. Translating this description into the ontological definition of concepts means that AR_1 is conforming to RS_1 if:

$$\exists i \in f(X), i \in A \text{ and } \exists i \in f(Y), i \in B$$

Similarly, rule AR_1 is filtered by $C(RS_2)$ if the condition and/or the conclusion of the rule AR_1 match the schema RS_2:

$$\forall i \in f(U), i \in A \cup B, \text{ and } \forall i \in f(V), i \in A \cup B,$$

The unexpectedness operator, U(RS), with a higher interest for the user, proposes to filter a set of rules with a surprise effect for the user. This type of rules interests the user more than the conforming ones since a decision maker generally searches to discover new knowledge with regard to his/her prior knowledge.

Moreover, several types of unexpected rules can be filtered according to the rule schema: rules that do not confirm either or both the condition and the conclusion of a rule schema.

For instance, let us consider that the operator $U(RS_1)$ extracts the rule AR_1 which is unexpected according to the condition of the rule schema RS_1. This is possible if rule conclusion B, matches the schema conclusion Y, while the condition, A, is unexpected according to the schema condition X:

$\forall i \in f(X), i \notin A$ and $\exists i \in f(Y), i \in B$

In a similar way, the two other unexpectedness operator usability are defined.

Finally, the exception operator applied over RS_1, is defined only over implicative rule schemas and extracts conforming rules with respect to the following new implicative rule schema: $X \wedge Z \rightarrow \overline{Y}$, where Z is a set of items.

IV. Experimental Results

We implemented the proposed algorithm for connecting ontology concepts with database items using PHP. The data set of 1000 medical records having different attributes such as symptoms, age, disease, tests, treatment is entered in the back – end database. The development of associations and the ontology is fully dynamic. In case, any symptom is searched, the proposed approach search from the back end database and creates the ontology that is dynamic in execution. It refers that the unsupervised approach of ontology generation is implemented so that the unbiased results can be achieved. The implementation of proposed work shows the dynamic results in terms of rules found and their effectiveness in the real world scenario.

Based on the dependency and association in the mentioned attributes, the PHP script is fetching and generating the rule that makes the ontology as well as classification.

4.1. Evaluation metrics

The proposed OntoDB mapping algorithm uses mainly two evaluation metrics, the accuracy and computation time. The accuracy is measured by overall performance that is calculated using fuzzy mathematical formula which is given below.

$$F(x)= [rand(1, 10) * (m / n)]+k$$

Where rand - Random Number Generator used for Fuzzy Inclusion
n - Number of Rules associated with each search term
m - Number of Rules found with the search term
k - Any Constant Value for the Simulation Run

The other metrics used here is the computation time which is measured based on the ontology formation time.

Using simulation, we are able to get the efficient results in terms of less execution time as compared to the classical approach. In the following table, we have processed and analyzed the execution time of ontology or relationship development execution time. It is found from the results that the association generation time if consistent and normal distribution based.

Table 1. Ontology Formation Time for various sized Medical Data Sets

Data Items in the Medical Data Set	Ontology Formation Time
20	1.2
30	1.422332
40	1.5223
100	2.0435853
1000	3.4946468

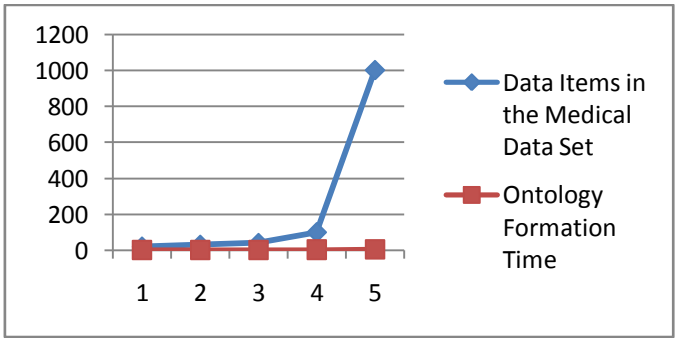

Figure 3. Time Analysis of medical data set based on size of data set

Table 2. Accuracy of Rule Generation

Existing Approach (In Percentage)	Proposed Approach (In Percentage)
30	90
50	95
60	93

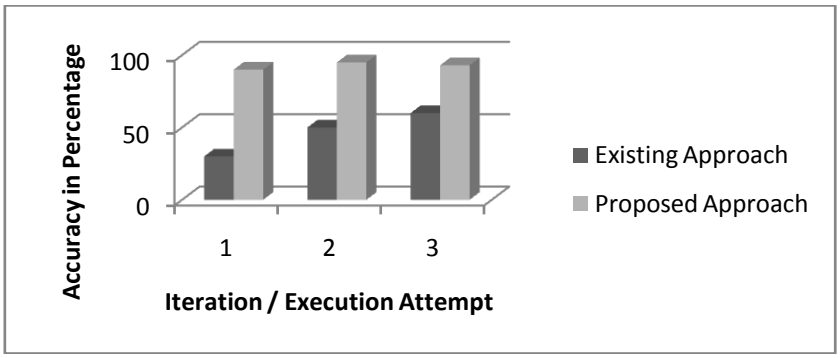

Figure 4. Accuracy of Rule Generation on Medical Dataset

V. Conclusion

This paper propose and implement an effective methodology for the ontology design and discusses the problem of helping the decision maker in the post-processing step of association rule mining. The manuscript is having the specific case study of medical database that is having difference aspects including symptoms, disease, tests and treatment. The manuscript proposes to prune and filter discovered rule by integrating user knowledge and beliefs. User knowledge is modelled in an ontology connected to data. Rule schemas allow user belief representation, and, combined with ontologies, they improve the selection of interesting rules. The manuscript intends to improve this approach in two directions - Developing the rule schema formalism and integrating the approach in the discovery algorithm.

References

[1]. Baesens, B., S. Viaene, and J. Vanthienen. (2000). Post-Processing of Association Rules. The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2000), pages 2 - 8.
[2]. Piatetsky-Shapiro, G. and C.J. Matheus. (1994). The Interestingness of Deviations. In U. M. Fayyad and R. Uthurusamy (eds.), Knowledge Discovery in Databases, Papers from AAAI Workshop (KDD ' 94), pages 25 – 36.
[3]. Klemettinen, M., H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. (1994). Finding Interesting Rules from Large Sets of Discovered Association Rules. International Conference on Information and Knowledge Management (CIKM), pages 401-407.
[4]. Padmanabhan, B. and A. Tuzhuilin. (1999). Unexpectedness as a Measure of Interestingness in Knowledge Discovery. Decision Support Systems, Volume 27, Number 3, Elsevier, pages 303-318.
[5]. Toivonen, H., M. Klementinen, P. Ronkainen, K. Hatonen, and H. Mannila. (1995). Pruning and Grouping of Discovered Association Rules. Mlnet Workshop on Statistics, Machine Learning, and Discovery in Databases, pages 47 - 52.
[6]. Bayardo, R.J. Jr. and R. Agrawal. (1999). Mining the most interesting rules. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 145 – 154.
[7]. Adomavicius, G. and A. Tuzhilin. (2001). Expert-Driven Validation of Rule-Based User Models in Personalization Applications. Data Mining and Knowledge Discovery, pages 33–58.
[8]. An, A., S. Khan, and X. Huang. (2003). Objective and Subjective Algorithms for Grouping Association Rules. Internation Conference in Data Mining, pages 477 - 480.
[9]. Chawla, S., J. Davis, and G. Pandey. (2004). On Local Pruning of Association Rules Using Directed Hypergraphs. Proceedings of the 20th International Conference on Data Engineering, pages 832.
[10]. Berrado, A. and G. C. Runger. (2007). Using metarules to organize and group discovered association rules. Data Mining and Knowledge Discovery, pages 409 – 431.
[11]. Nigro, H.O., S.E. Gonzalez Cisaro, and D.H. Xodo. (2007) Data Mining With Ontologies: Implementations, Findings and Frameworks, Idea Group Reference.
[12]. Srikant, R. and R. Agrawal. (1995). Mining Generalized Association Rules. In U. Dayal, P.M.D. Gray, and S. Nishio, eds, Proceedings of the 21st International Conference on Very Large Databases, pages 407 – 419.
[13]. Češpivová, H., J. Rauch, V. Svátek, M. Kejkula, and M. Tomečková. (2004). Roles of Medical Ontology in Association Mining CRISP-DM Cycle. Workshop Knowledge Discovery and Ontologies in ECML/PKDD
[14]. Euler, T. and M. Scholz. (2004) Using Ontologies in a KDD Workbench. In Workshop on Knowledge Discovery and Ontologies at ECML/PKDD.
[15]. Zhou, X. and J. Geller. (2007). Raising, to enhance rule mining in web marketing with the use of an ontology. Date Mining with Ontologies: Implementations, Findings and Frameworks, pages 18-36
[16]. Gruber, T. (1993). A translation approach to portable ontology specification. Knowledge Acquisition, 5:199-220.
[17]. Liu, B., W. Hsu, K. Wang and S. Chen. (1999). Visually Aided Exploration of Interesting Association Rules. Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining, Lecture Notes In Computer Science, Vol. 1574, Springer-Verlag, pages 26 – 28.