USING ONTOLOGIES AND INFERENCE ENGINEs IN ASSOCIATION RULES OF DATA MINING: AN APPLICATION IN A MEDICAL LABORATORY DIAGNOSTIC COMPANY

KNOWLEDGE DISCOVERY USE ONTOLOGIES IN DATA MINING

Lucélia Branquinho
Federal University of Minas
Gerais – B.H. Brazil
renatabaracho@ufmg.br

Renata Abrantes Baracho
Federal University of Minas
Gerais – B.H. Brazil
luceliabranquinho@gmail.com

Maurício Barcellos Almeida
Federal University of Minas
Gerais – B.H. Brazil
mba@eci.ufmg.br

Renato Rocha Souza
Getúlio Vargas Foundation
Rio de Janeiro - Brazil
rsouza.fgv@gmail.com

Abstract: The growing amount of information generated and available nowadays requires studies that add innovations in representation, organization and retrieval of information. A challenge in information retrieval within a specific domain is to create semantic relationships between the terms of a specialized vocabulary. In doing so one can reach efficient representative knowledge models. This can be done, for example, throughartifacts as ontologies, which can improve the decision-making process. The purpose of this paper is to describe how the use of ontologies and their inference engines can enhance the information retrieval regarding the prescriptivebehavior of laboratory tests for viral hepatitis, particularly in Knowledge Discovery in Database (KDD) initiatives. We conducted a literature review in well-kwnon sources using descriptors such as Knowledge Discovery in Databases (KDD), ontologies, data mining, association rules, semantic similarity and the Logical Observation Identifiers Names and Codes (LOINC). Our findings suggest that the combination of knowledge discovery and data mining techniques assisted by ontologies contribute for information retrieval and, consequently, for the efficient knowledge extraction.

Key-Words: Information retrieval, knowledge discovery, ontologies, data mining, information organization, hepatitis virus.

I. INTRODUCTION

The amount of data stored in organizational databases has surpassed the human capacity of analysis, even considering the use of well-established technologies (DALFOVO, 2000). There is a need for adopting approaches that allow one to analyze massive data sets with the aim of improving the medical decision-making to both physicians and managers of healthcare organizations. A well-know alternative is the approach generally referred to as Knowledge Discovery in Databases (KDD) or Data-Mining (DM).

A challenge posed to the research community of data-mining is how to improve accuracy and predictions, when dealing with increasingly large amounts of data. An alternative mentioned in the literature is the so-called “semantic data-mining”, which proposes new approaches to an already existing set of DM techniques (VAVPETIC, 2012). In this context, ontologies have been seen as efficient alternatives for knowledge representation, since one can use them to explicitly describe entities in a domain, as well as the relationships between these entities.

Information systems have been considered inefficient in accurately dealing with the issue of lexical meaning.The search for machines just depends on the effectiveness of the semantic representation adopted. The importance of using computational methods grows to the extent that the amount of information available to support decision making exponentially increases (KASAMA, 2011).

In general, information retrieval systems have difficulty in considering the lexicon of a domain through semantic relations of a specialized vocabulary. An efficient knowledge representation model can be provided by ontological engineering techniques (KASAMA, 2011) .

The application of methods and technologies based on data mining along with semantic basis (ontologies) make possible to discover and transform explicit strategic information of the domain into relevant knowledge, which can be used to understand complex issues in the scope of decision-making process.

In this context, a multidisciplinary view is required, one that considers not only the data, but also information and knowledge issues in organizations. Using such a view, one can establish principles to the settlement of knowledge representation issues, as well as to adopt ontologies for assisting KDD process .

This paper aims to improve the DM process through the introduction of domain knowledge organization in both pre-processing and pos-processing phases. We hope that the use of biomedical ontologies allows us to include subjective measures for classification and generalization, which will be included in the the hierarchy of the ontology. The identification of medical terms that share identical relations can bring effectiveness and relevance to the classification of association rules.

Our experiment was delimited considering the universe of laboratory tests for clinical analyses for diagnotics in the domain of viral human hepatitis. We also use LOINC as a reference for laboratory exams codification and SNOMED-CT for clinical terms defined as interoperability standards for information systems.

II. LITERATURE REVIEW

Our literature review included information retrieval retrieval and knowledge recovery, as well as semantic similarity measuers in ontologies.

INFORMATION RETRIEVAL AND KNOWLEDGE RECOVERY

Information retrieval systems require organizational knowledge systems that allow the representation of knowledge semantic structures in order to facilitate the information retrieval, for example, when eliminating ambiguity and controlling synonyms.

a relevant function of knowledge representation is to create an efficient structure for information retrieval.

Knowledge organization systems (KOS) are defined as a representation of entities (theories, processes and tools) generated by relevant interpretations of a selected parts of the world or domain (SOUZA, TUDHOPE and ALMEIDA, 2010), which can be recordes in information systems and documents. Ontologies are examples of KOS used to solve conceptual domain issues.

Indeed, the term ‘ontologies’ has been recently widely used in fields like Artificial Intelligence, Computer Science, Information Science, in order to promote improvements in information cooperative systems, information intelligent integration, information extraction, information retrieval, knowledge representation, and database management systems (GUARINO, 1998).

Gruber (1998) defines an ontology as a formal and explicit specification of a shared conceptualization. “Conceptualization” refers to an abstract model of some phenomenon in the world, which identifies relevant concepts of that phenomenon. “Explicit” means that the concepts used and the restrictions on their use are clearly defined. “Formal” refers to the fact that ontologies should be readable by machines. “Shared” refers to the idea that ontologies capture consensual knowledge..

Within complex fields, which maitain a constant and large scientific production, data and resources needs to be articulated and described in a standardized way. This can allow one to take advantage ofknowledge scattered within practice communities, where there are professionais with urgent demands for expert information. In some fields, like biomedicine, some communities have developed and released publicly, since the 1990s (PEREZ-REY et al., 2004), a set of ontologies to aid in the description and retrieval expert information.

The extraction of knowledge, which is generally referred to in the literature as Knowledge Discovery in Database (KDD) or data mining (DM) involves " the process of identifying valid standards, new potentially useful and understandable embedded in the data” (FAYYAD et al 1996). KDD process can be grouped into three phases: pre-processing, DM and post-processing. Pre-processing comprises the collection, organization and treatment of data; DM included the creation of algorithms and the use of techniques for knowledge search ; post-processing deals with the knowledge gained in the DM phase and its interpretation.

However, the knowledge discovered in the process, even correctly from the statistical point of view, might not be easily understood by an user. For example, discovered patterns can be many or they can contain a lot of redundancy. In addtion, they can only represent previously known relationships and rule patterns initially insignificant (but one that might be relevant).

In the post-processing phase, among the measures used to evaluate DM resulting patterns, there are objective (data-driven) and subjective (user-driven) measures, whicha are used in the seek of evaluating the novelty of the discovered pattern (Gonçalves, 2005) .

A classic problem in data mining realm is the choice of accuracy, the so-called "gap sieve". In the case of specifications that require little support and confidence values ("screen with close-knit"), the user can be overwhelmed with a large number of very similar patterns and associations. If the gap decreases, through the specification of different values of support and confidence, some interesting associations and patterns might not be identifiable (Ferraz, 2008). Once generated the set of association rules. it is necessary to reduce its cardinality. This can be done through pruning mechanisms. Then, the problem happens to be what rules should be pruned.

The objective measures, like support and confidence, are often used to evaluate the extent in which the association rules are interesting.

Consider $I = \{i1, i2, i3, ..., yn\text{-}1, in\}$ is a set of attributes or items. Transaction is a set of items $T = \{t1, t2, t3, ..., tn\text{-}1, tn\} \subset I$ T. D is a set of transactions or data relevant to the task. An association rule is an implication $P \rightarrow Q$ where $P \subset I$, $Q \subset I$ and $P \cap Q = \emptyset$ (Ferraz, 2008)

The support of rule $P \rightarrow Q$ is defined as the probability that a transaction in both items containing D, P and Q, SD $(P \rightarrow O) = DS (P \cup Q)$. The confidence of rule $P \rightarrow Q$ is defined as the probability that a transaction in D, which already contains the items P and also contains items on Q, CD $(P \rightarrow O) = (SD\ P \cup Q) / SD (P)$ (Ferraz, 2008).

A rule $P \rightarrow Q$ [s, c] is called strong rule, if data values for the parameters minimum support (s) and minimum confidence (c) then (Ferraz, 2008):

sD $(P \rightarrow Q) = s$, $s \geq$ minimum support

cd $(P \rightarrow Q) = c$, $c \geq$ minimum confidence

The minimum support for the item-set guarantee statistically significant sample while avoids the consideration of low frequency combinations. The minimum confidence shows that the result is not occasional, there is a some cohesion between relevant rules (FERRAZ, 2008)

The traditional data mining is based on the frequency of occurrence of cases and the co-occurrence of items in transactions. The meaning of each item or instance is not taken into consideration. The semantic content extracted from ontologies allow the inclusion of more intelligence and knowledge in data mining processes, then improving their quality. (FERRAZ, 2008)

A new paradigm for knowledge discovery has been proposed with application of ontologies in mining, which is guided by domain knowledge (domain-driven data mining, D3M), as well as considered the feasibility of using knowledge discovered and delivered (actionable knowledge discovery and delivery, AKD) (Marinică, 2010; Ribeiro, 2010; CAO, 2010; Coelho, 2012).

Some authors report how as ontologies are used to improve the results of mining association rules methods from the adoption of semantic descriptors (Silva 2007; Ferraz, 2008; Marinică, 2010; KASAMA, 2011; VAVPETIC 2012; HAVE, 2013; GAN, 2013; Hamani, 2014).These authors use semantic treatment with domain ontologies to record prior knowledge and so, in the pre-processing stage or post-processing, to introduce restrictions grouped into two types: pruning restrictions, which intend to filter uninteresting stuff; and abstraction restrictions, which allow the generalization of items in relation to ontologies concepts.

According to Ferraz (2008), the advantage of pruning restrictions is that it eliminates, since the beginning, information that users are not interested on. General rules can replace a number of specific rules through a generalization process. When this is possible, there is both a semantic improvement of the mined association rules set and a future reduction in the cardinality of the set of rules.

Approaches for pruning and generalization rules calculate the degree of semantic similarity between terms that are generally based on graphs hierarchy structure, This is compatible with the ontologies, which can be represented as directed acyclic graphs (DAG).

SEMANTIC SIMILARITY IN ONTOLOGIES

Domain ontologies are used to formalize the semantic relationships between concepts of a particular domain, to recover the best set of semantically similar patterns and to increase the semantics through annotations. If compared with other classification schemes, ontologies allow the construction of more accurate and complete domain models. In the scope of ontologies development, the semantic of data recovered is made explicit. In addition to make explicit standard semantic relationships, ontologies also allow the representation of specific relationships and concept attributes (Breitman et. Al. 2007).

There are several approaches to measure the semantic similarity between terms in ontologies represented by a DAG. In general, two types of comparison are adopted: the edge-based and node-based, as shown in Figure 1 (PESQUITA, 2009).

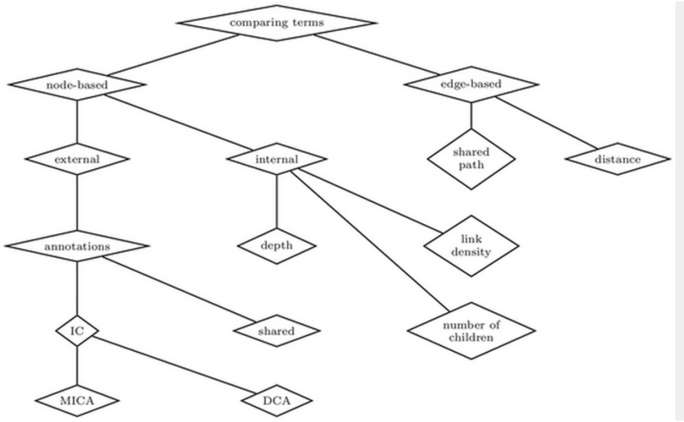

Figure 1 – Semantic Similarity in Biomedical Ontologies
Source: Pesquita (2009)

Edge-based approaches are mainly based on counting the number of edges between two terms in a graph. The most ordinary technique, the distance, either selects the shortest path or the average of all paths. This technique produces a measure of the distance between two terms, which can be easily converted into a measure of similarity. Although these approaches are intuitive, they are based on two assumptions that are rarely true in biological ontologies: (1) nodes and edges are evenly distributed; and (2) edges at the same hierarchical level correspond to the same semantic distance between terms (PESQUITA, (2009). The semantic relationship, in this case, can be obtained using the path length between the terms (in the graph) (RESNIK, 1999 APUD GAN et al, 2013).

Node-Based approaches compare properties of terms (which may be related to themselves), their ancestors, and their descendants terms. A concept ordinarily adopted is information content (IC), which provides a measure of whether a term is more specific and informative.

Performance evaluation studies showing various semantic similarity measures have revealed that the use of information content shared by two concepts is a very effective technique in comparing concepts (BUDANITSKY; HIRST, 2001 APUD Couto, 2011).

According to Ferraz (2008), the interesting rules measures consider the relationship between the antecedent and the consequent. One can said that an instance is covered by a rule, if the rule antecedent is true. Usually, this approach uses the "is-a" relationship (generalization and specialization) and "part-of"relationship (composition) in order to define subclasses and superclasses between concepts in ontologies hierarchy. This can allows a reduction in the number of association rules generated by data mining, which might be more effectively than purely syntactical attempts, while enriching the set of semantic rules under mining.

According to Manda (2013), the methods that use ontologies to calculate semantic similarity are classified into five categories: (1) methods based on semantic distance, (2) information content-based methods, (3) methods based on terms of properties, (4) methods based on the ontologies hierarchy and (5) hybrid methods. The method based on semantic distance produces a measure of the distance between two terms and can be easily converted into a measure of similarity.

The Similarity Resnik (SR) is an example of similarity measure based on MICA (most informative common ancestor), which corresponds to the common ontological ancestor term from two terms that has the largest information content (IC).

IC is defined as: IC = - log (p (c)) (1); where p (c) is the probability of a given word occurrs in a given scenario. The calculation p (c) is done by adding up all occurrences of the term and so dividing the result by the total number of occurrences.

The similarity between two words is found by the following measure: SimTermos (c1, c2) = IC (mica) (2); MICA refers to common ancestry informative, or the ancestor calculated between the parents of c1 and c2.

Wang et al. (2007) cited Gan (2013) developed a hybrid calculation of the SS in which each edge is given a weight according to the type of relationship. Thus, calculates the SS between two terms based on the position of both within the ontologies and the semantic relations of these terms with their ancestors.

Tversky (1977) states that each object is a set of concrete and abstract features and the assessment of weighted selection of product similarity of these common features between objects.

III. METHODOLOGICAL PROCEDURES

Our methodological steps consist of developing ontologies for viral hepatitis, of extratiing association rules and calculatingof the semantic similarity between laboratory tests.

VIRAL HEPATITIS ONTOLOGY CONSTRUCTION

The organization of biomedical terminology is a constant challenge, isofar as there are multiple interpretations of the data and ways to get them (Smith, 2008).

The first part of our study includes the development of an ontologies or viral hepatitis. In order to do this, we adopted the Ontologies Development 101 methodology (NOY and McGuinness, 2001).

In seeking concepts related to diagnosis of viral hepatitis, we used LOINC and SNOMED provided by Bioportal. In addition, the Regenstrief Institute provides a Windows-based mapping utility, called the Regenstrief LOINC Mapping Assistant (RELMA), which facilitates searches through the LOINC database aiding in maping local codes to LOINC codes. This tool is used to query the Portuguese translation of LOINC terms and subsequent validation of the ratios obtained by mapping the codes used in a medical diagnostics company.

We surveyed RELMA and Bioportal in October 2014using as descriptors "hepatitis" and "hepatitis virus". In order to correlate tests with the prescritive behavior of the physician, we identified the possibility of reusing ontologies OGMS[1], IDO[2], DOID[3], OBI[4] e FMA[5]. Symptoms, signs, infectious disease course and disease related the laboratory testing prescription to assessing the patient and confirm diagnosis of viral hepatitis.

Finally, OGMS was adapted to include laboratory tests based on the test proposed by Eilbeck et al. (2013). This proposal features the LOINC structure and considers the identification attributes (ID LOINC, a name, a gender) and four significant relatioships regarding the scope of the search (a system, an analyte6, a method, an organism).

EXTRACTION ASSOCIATION RULES

Based on knowledge gained from the development of the ontologiy of viral hepatitis, we extracted from the laboratory database, the test ordering relationships that contained at least one of the laboratory tests listed when there is a chance of infection by hepatitis viruses.

Therefore, in a laboratory test order prescribed by doctors it is possible to identify the disease stage being investigated and thereby generalize the terms that promote the pruning of non-relevant attributes. In the triage phase of Hepatitis B, for example, 4 specific tests may be requested for the identification of the vírus and another 8 unspecific tests may be ordered for monitoring liver functions. These may be generalized by specifying the disease and the stage being studied since infectologists already know the relationship between these laboratory tests without having to specifically name each one as an attribute for data mining.

We used an algorithm called Apriori and the package Arules, based on R language (Hashler et al, 2007). we considered the support and confidence-building measures with low values for that mining does not seep rules that can be considered relevant.

CALCULATION OF SIMILARITY SEMANTICS BETWEEN LABORATORY TESTS

According to Kasama (2011), ontologies involve the formalities required for the description of expert knowledge, allowing the use of logic and inferences in the structured information.

To build ontologies in Web Ontology Language (OWL) allows one to use OWL reasoners , or inference engines, . The inference mechanisms are responsible for the search of rules in the evaluated knowledge base, in addition to: to direct the process of inference, to enter new information, to classify instances and to check both rules and consistency.

Our study proposes the utilization of ontologies, reasoners and Jena software to promote pruning and filtering (generalization) of data from the list of laboratory tests collected from the medical diagnosis company's database. When analysing the relationships among terms, one can identify which laboratory tests are related to which disease. Therefore, the similarity between terms is considered as a mean to generalize the attributes in the pre-processing phase. The aim is to adopt small support and confidence values ("screen with close-knit") in the post-processing phase. In doing so, one can obtain more rules classified with similarity semantic measure calculation proposed by Tversky (??????). So, we used theontology created to represent the knowledge of laboratory tests. Figure 2 depicts the proposed model.

[1] OGMS (Ontology for general medical science)

[2] IDO (Infectious disease ontology)

[3] DOID (Disease ontology)

[4] OBI (Ontology for Biomedical Investigations)

[5] FMA (Foundation Modelo of Anatomy ontology)

[6] Analyte – chemical species present in the sample whose concentration is to be determined in an analysis.

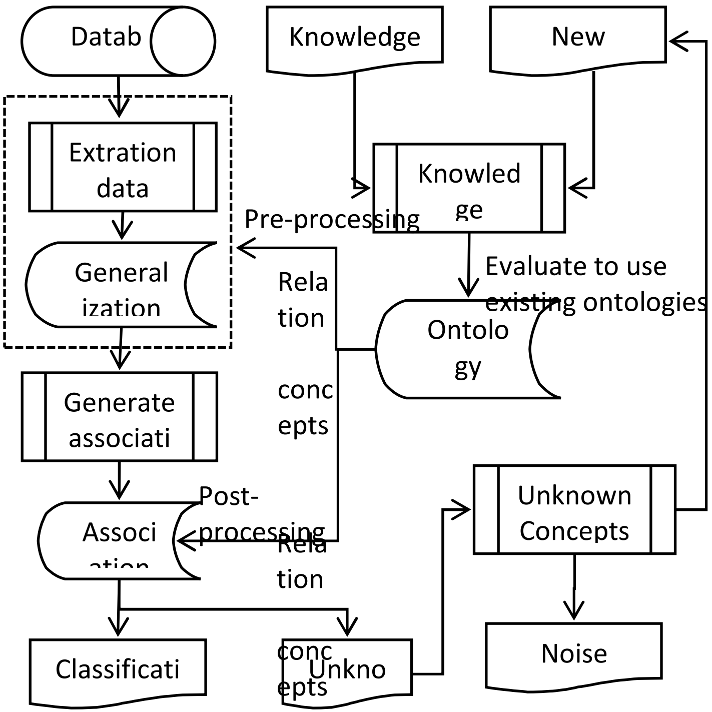

Figure 2 – Knowledge extraction model with association rules and ontologies. Source: adapted of Hamani (2014) and Coelho (2012).

,

The use of inference engines allows one to retrieve information and to classify rules with greater assertiveness in decision support.

IV. RESULTS

The diagnosis for viral hepatitis is based on protocols that guide the prescription of laboratory tests by doctors, either in the course of the disease or at its initial triage, in order to confirm the infection. Considering these protocols, the relation between LOINC laboratory tests and research conducted in a diagnostic medicine laboratory was possible with the reutilization of OGMS, IDO, DOID, OBI e FMA ontologies to map knowledge on viral hepatitis. This mapping enabled the generalization and the consequent reduction of the number of attributes to be mined via the identification of similarities between laboratory tests, considering the relations mapped regarding viral hepatitis ontology, named HVO.

Based on the knowledge obtained in the development of HVO ontology, a list of laboratory test orders was collected from the company's database. This list contains at least one of the tests directly related to a hypothesis of hepatitis vírus diagnosis. Laboratory test orders that complied with the previously described rule were selected from patients service centers in three months, suming up 1760 occurrences. These occurrences featured 458 different laboratory tests. The selected sample was used for the entire process of database selection and transformation. This enabled the subsequent analysis of the relationship between laboratory tests for term generalization.The support value of 0.2 and confidence of 0.75 were used to run the Apriori algorithm.

Based on HVO ontology, considering the relationships with the disease and with its stage and also utilizing the Jena tool and inference rules , it was possible to obtain more general terms to represent laboratory test groups associated with viral hepatitis diagnosis. The generalization possible to reduce the number of attributes to 439. The same support and confidence values were used for the implementation of the Apriori algorithm. So we found the results shown in Table 1.

Services featuring viral hepatitis exams	Number of distinct rules (no generalization)	Number of distinct rules (with generalization)
1760	573	75

Table 1. Rules associaction

The format of an association rule can be represented as an implication LHS -> RHS, wherein LHS (antecedent) and RHS (consequent) are respectively the left side (Left Hand Side) and the right side (Right Hand Side) the rule defined by disjoint sets of items. Considering the extraction of association rules with ontologies rhs list (consequent rule) with the information of the disease (hepatitis A, B, C, D, E and G) the number of association rules is reduced to 75.

Based on this approach, the tests and simulations it was possible to reduce the number of attributes by 15%, according to table 2 and consequently, the generation rules of interest, as exemplified in Table 2.

	Request – Medical prescription					Lab tests after generalization	
Physician Exams	L.T. 1	L.T. 2 ao L.T. 7	L.T. 8	L.T. 9	L.T. N	L.T.1	L.T. N
Patient 1	A.FETO	---	TGP	AU	...	Hepatitis C	...
Patient 2	ALB-D	---	HAV-G	HAV-M	...	Hepatitis A	...
Patient 3	A.FETO	---	HG	PTRHCV		Hepatite C	...
Patient 4	PAL	---	PTRHCV	PTRHCV	...	Hepatite C	...

A.FETO	LOINC 1834-1 - Alpha-1 Fetoprotein – Laboratory test unspecified viral hepatitis
TGP	LOINC 1742-6 - Alanine aminotransferase – Laboratory test unspecified viral hepatitis
ALB-D	LOINC 61151-7 - Albumin – Laboratory test unspecified viral hepatitis
PAL	LOINC 6768-6 - Alkaline phosphatase
AU	LOINC 5196-1 - Hepatitis B virus surface Ag
HAV-G	LOINC 5179-7 - Hepatitis A virus Ab.IgG
HAV-M	LOINC 13950-1 - Hepatitis A virus Ab.IgM
PTRHCV	LOINC 49758 -1 - Hepatitis C virus Ab.IgM
L.T.	Generalized Attribute

Table 2. Example of generalization

The results obtained in the data mining step using the Apriori Algorithm, as well as the construction of ontologies, identify the most adequate technique for calculating the semantic similarity between terms.

The traditional Apriori algorithm uses two pieces of information for inference with association rules: support and confidence. However, these metrics are not always sufficient to determine whether the data patterns are found significant. So, in addition to support and confidence, we used the metric Lift, which have its value recalculated based on the similarity between the complementary antecedent and consequent examinations, as exemplified in Table 3.

Antecedent	Consequent	Lift
Glucose,HIV 1 and 2 Ab Ag, Thyrotropin	Hepatitis C	1,24
Complete blood count, Urynalises, HIV 1 and 2 Ab Ag	Hepatitis C	1,16
Glucose, Complete Blood Count, Thyrotropin, Urinalyses	Hepatitis C	1,13
Glucose, Urinalyses, Reagin Ab	Hepatitis B	1,04

Table 3. Rules associaction Lift

Since these calculated values were added to the amount of lift of the rule, the value is so obtained by calculating the dissimilarity between the antecedent and consequent.

Weather was not identified semantic similarity between some item of antecedent and consequent was regarded as standard to 1, as this precedent is not directly related to the diagnosis of viral hepatitis or routine tests, so it is an interesting item to be evaluated.

The rules were reclassified considering the amount of lift reconfigured after calculating the semantic similarity.From the end of data collection already treated, we determined the most relevant association rules.

Using this information, we could evaluate the similarity between antecedent and consequent using the function of the Relational Model (Model Ratio) of Tversky (1). From the psychology model for the assessment of the similarity, considers the common and uncommon characteristics between stimuli (concepts) and the context.

$$s(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha * f(A - B) + \beta * f(B - A)} \qquad (1)$$

As α, β ≤ b 0. stimuli (concepts), A and B sets the characteristics of the stimuli.

As the experiment purpose is to find interesting association rules was added to the amount of lift of each rule the similarity (SSM) value between concepts (antecendete and consequent rule), as shown in Table 4.

Antecedent	Consequent	Lift	SSM	Result
Glucose,HIV 1 and 2 Ab Ag, Thyrotropin	Hepatitis C	1,24	0,238	1,478
Glucose, Urinalyses, Reagin Ab	Hepatitis B	1,04	0,285	1,325
Complete blood count, Urynalises, HIV 1 and 2 Ab Ag	Hepatitis C	1,13	0,143	1,273
Glucose, Complete Blood Count, Thyrotropin, Urinalyses	Hepatitis C	1,13	0,109	1,239

Table 4 Rules associaction Lift + SSM

V. CONCLUSION

The developed application ontology is the LOINC® tests for viral hepatitis at a high level understanding that this classification is sufficient to evaluation the relationship of association rules. This paper did not intend to exhaust all terms and concepts related to disease diagnosis and laboratory tests.. It is not intended to be a reference ontology for other purposes. It aims to extend the class of laboratory testing for OGMS enabling one to associate the prescription of diagnosing the disease cycle, and therefore allowing to identify better correlations between association rules. This approach is extended to other tests that are extracted in the mining process by association rules and that initially is not related in primary form the diagnosis of viral hepatitis fitting assessment and rehabilitation of ontology. Describe each laboratory test using differentiating principles allows the classification in a hierarchy from general to specific, providing an ideal source of information for pruning rules and also for analysis of relevance to the context. Although laboratory tests in LOINC® are classified into six different axes, in the scope of this model, it was not necessary to use all of them. We maintain the the focus only on components, specimen and method.

It is worth to highlighting two aspects in the use of semantic similarity and ontologies: first, the relevance of patterns extracted by semantic similarity between terms identification techniques is highly dependent on the construction and validation of ontologies; so it is important that the used domain ontologies have been validated by business experts; second, the approach in the KDD processes needs further clarification in algorithms that are not restricted only to relations 'is-a' and 'part of', so improving the use of formal semantics proposed in ontologies, which can be applied to the computational methods of data mining.

The application of methods and technologies based on data mining along with semantic basis allows one to gain strategic information, which are not explicit, to be discovered and transformed into relevant knowledge to understand complex issues and decision-making processes.

We conclude through the literature review and our experiment that the use of ontologies in KDD with association rules has contributed to the enrichment of an unexpected knowledge extraction, which is relevant to business and decision making. The choice of appropriate algorithms for calculating the semantic similarity between terms in ontological basis can aind the obtaining of patterns without redundancy, then improving recall and precision..

When creating generic referenced item-set of the disease occurrence, it is possible to reduce the attributes in the pre-mining phase, then improving the performance of the all mining. This is reached by reducing combinations and bringing meaning to the results in the post-processing phase, which make easy the analysis and classification required. Regarind to this, our method exerts a subjective metric function of the results obtained.

The increased demand for computational methods to solve the problem of knowledge extraction in large data sets, gig data, web, etc, can facilitate by the use of ontologies. This can allows the organization and retrieval of information helping to solve the problems of volume, variety and veracity. Therefore, the merger of the data mining and domain ontologies fields can be a great opportunity to meet the challenge of dealing with Big Data.

VI. REFERENCES

[1] CAO, Longbing. Domain-Driven Data Mining: Challenges and Prospects. **Transaction on Knowledge And Data Engineering**, Vo. 22, n° 6, June, 2010, pp. 755-769. CAO, Longbing. Introduction to Domain Driven Data Mining. Chapter 1. Available from:<http://staff.it.uts.edu.au/~lbcao/publication/dmba-dddm.pdf>.

[2] COELHO,E.M.P., Fuzzy Ontologies in Support Data Mining applications in the Finance Department of the City of Belo Horizonte, Belo Horizonte, 2012. School Information Science Federal University of Minas Gerais, Belo Horizonte, 2012.

[3] COUTO,F., SILVA, M., Disjunctive shared information between ontologies concepts: application to Gene Ontologies, Journal of Biomedical Semantics, vol. 2, no. 5, 2011. Available from:< http://dx.doi.org/10.1186/2041-1480-2-5> .

[4] DALFOVO, O., Who has information is more competitive: the use of information by managers and entrepreneurs to onter competitive advantage. Blumenau. Acadêmica. 2000.

[5] EILBECK, K. ; JACOBS, J. ; STAES, C.J., Exploring the use of ontologies and automated reasoning to manage selection of reportable condition lab tests from LOINC, 2013. Available from: < http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?arnumber=6480135> .

[6] FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R.. Advances in Knowledge Discovery and Data Mining. 1996.

[7] FERRAZ, I., Knowledge of the world as a tool to enrich the results obtained in data mining. UFF, Rio de Janeiro, 2008

[8] GAN, M., DOU, X., JIANG, R., From ontologies to semantic similarity: calculation of ONTOLOGIES-based semantic similarity. Available from: <http://www.hindawi.com/journals/tswj/2013/793091/>.

[9] GONÇALVES,E.C. Objective and Subjective Measures for Association Rules. INFOCOMP Journal of ComputerScience, 2005. Available from: <http://www.dcc.ufla.br/infocomp/artigos/v4.1/art04.pdf>.

[10] GRUBER, T. What is an ontologies? (1993). Available from: <http://wwwksl. stanford.edu/kst/what- is-an-ontologies.html>.

[11] GUARINO, N. ,Formal Ontologies and Information Systems,1998. in: N. Guarino, (Ed.) Formal Ontologies in Information Systems. pp. 3-15, IOS Press, Amsterdam, Netherlands.

[12] HAMANI, M., MAAMRI, R., KISSOUM, Y., SEDRATI, M., Unexpected rules using a conceptual distance based on fuzzy ontologies, 2014. Available from: < http://www.sciencedirect.com/science/article/pii/S1319157813000141>

[13] KASAMA, D., ZAVAGLIA, C., BARCELLOS, G., The term to the semantic structure: ontological representation of the field of nanoscience and nanotechnology using the Quali Structure. Available from: < http://www.erevistas.csic.es/ficha_articulo.php?url=oai:linguamatica.com:article/73&oai_iden=oai_revista909 >.

[14] HASHER,M., HORNIK, K., GRUN, B., BUCHTA, C., Introduction to arules – A computational environment for mining association rules and frequent item sets, 2007. Available from:<http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf>.

[15] MANDA, P., McCarthy, F., BRIDGES, S., Interestingness measures and strategies for mining multi-ontologies multi-level association rules from gene ontologies annotations for the discovery of new GO relationships. Available from: < http://www.ncbi.nlm.nih.gov/pubmed/23850840 >

[16] MARINICA, C. Título: Association Rule Interactive Post-processing using Rule Schemas and Ontologies - ARIPSO. 2010. Tese – Department of Computer Science, Ecole poltechnique de l'Universite de Nantes, 2010. Disponível em: Available from: <http://www.claudiamarinica.com/publications.html>.

[17] NOY, N.; McGUINESS, D. Ontologies Development 101: A Guide to Creating Your First Ontologies. Stanford University, Stanford, CA, 94305, 2001. Available from: http://www.ksl.stanford.edu/KSL_Abstracts/KSL-01-05.html .

[18] PEREZ-REY, D; MAOJO, V; GARCIA-REMESAL, M; ALONSO-CALVO, R. Biomedical Ontologies.In: Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering, p.207, 2004.

[19] PESQUITA, C., FARIA, D., FALCÃO, A.O., LORD, P., COUTO, F.M., Semantic similarity in biomedical ontologies, 2009. Available from: < http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000443>

[20] SILVA, D., Study of semantic similarity functions of terms applied to a domain. Available from:< http://www.cin.ufpe.br/~tg/2007-2/dfs3.pdf>.

[21] SMITH, Barry. Chapter 4: New Desiderata for Biomedical Terminologies .In: MUNN, Katherine; SMITH ,Barry. (Eds.)Applied Ontologies :An Introduction.Frankfurt: Ontos Verlag,2008. p.84-107.

[22] SOUZA, R. R. TUDHOPE, D.; ALMEIDA, M.B. The KOS spectra: a tentative faceted typology of Knowledge Organization Systems. Anais do ISKO, 2010. Available from: <http://mba.eci.ufmg.br/downloads/ISKO%20Rome%202010%20submitted.pdf>.

[23] VAVPETIC, A., LAVRAC, N. Semantic Subgroup Discovery Systems and Workflows in the SDM-Toolkit. Available from: < http://comjnl.oxfordjournals.org/content/early/2012/06/04/comjnl.bxs057>.