Ontologies in support of data mining based on associated rules: a case study in a diagnostic medicine company

Lucélia P. Branquinho[1], Maurício B. Almeida[1], Renata M.A. Baracho[1]

[1]School of Information Science, Federal University of Minas Gerais (UFMG) - Belo Horizonte - MG - Brazil

luceliabranquinho@gmail.com, mba@eci.ufmg.br, renatabaracho@eci.ufmg.br

Abstract. A well-known alternative to identify hidden standards is the use of data mining techniques. In order to obtain more efficiency in data mining, ontologies have been used to improve the representation in specialized knowledge domains. Here, we apply ontologies in a dataset of a diagnostic medicine company, which concerns to viral human hepatitis, with the aim of obtaining the best correlations between the laboratory tests prescribed by physicians and the real occurrences of diseases. Our preliminary findings show that the use of ontologies provides reduction in the number of attributes in the pre-processing phase, then improving the performance of data mining process as a whole.

1. 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]. Thus, there is a need for adopting approaches that are able to analyze masses of data with the ultimate 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).

So, many approaches in the literature have made reference to ontologies and their semantic descriptors as a way to improve the performance of data mining based on association rules [Ferraz 2008; Vavpetic 2012; Manda 2012].

This paper aims to make an effort towards the improvement of the KDD process through the introduction of domain knowledge in the pre-processing phase. The experiment was limited to the universe of laboratory tests required for clinical analyses. In particular, we focus on diagnostics to identify viral human hepatitis. We use LOINC[1] as a reference for laboratory exam codification.

The diagnosis for viral hepatitis is based on protocols that guide the prescription of laboratory tests by doctors over the course of the disease or at an initial trial for confirmation of infections. Considering these protocols, LOINC and the research conducted in a diagnostic medicine laboratory, it was possible to map knowledge to laboratory tests viral hepatitis and reuse the OGMS[2]. In this process, we also reuse biomedical ontologies as IDO[3] , FMA[4] and DOID[5]. This mapping enabled the

[1] Available: <https://loinc.org/>.
[2] Available: <http://www.obofoundry.org/cgi-bin/detail.cgi?id=OGMS>.
[3] Available: <http://infectiousdiseaseontology.org/page/Main_Page>.
[4] Available: <http://sig.biostr.washington.edu/projects/fm/>.
[5] Available: <http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology>.

generalization and the consequent reduction of the number of attributes to be mined via the identification of similarities between laboratories tests, considering the relations mapped in the viral hepatitis ontology called HVO.

The next sections will be organized as follows: section 2.1 provides a brief description on the use of ontologies in data mining; section 2.2, describes the construction of prototype of a viral hepatitis ontology; section 2.3 explains how generalization of terms will be applied in the data mining pre-processing phase as per association rules; section 3 details the results obtained with the proposed model. Finally, section 4 showcases final considerations.

2. Method

In some fields, such as Biomedicine, specialized communities have been developing and publishing, since the 1990s, a series of ontologies to aid in representing and retrieving informational [Perez-Rey et al. 2004].

2.1. Ontologies and data mining

Knowledge extraction, generally referenced in literature as Knowledge Discovery in Database (KDD) should be grouped into three phases: pre-processing, DM and post-processing. Pre-processing, which is relevant for our goals in this paper, comprises the collection, organization and treatment of data, while DM involves algorithms and techniques to search for knowledge.

Ontologies have been used to increase the relevance of the patterns discovered through the mining techniques. One of the techniques in which ontologies are being utilized is mining through association rules, which display the correlation between sets of items in series of data or transactions. [Ferraz 2008].

The advantage of pruning restrictions is to exclude information in which users are not interested in since the beginning. Every general rule should be able to replace a number of specific rules by means of generalization processes. Whenever this approach is feasible, a semantic improvement of the mined association rules and a reduction in the cardinality of the set rules will simultaneously take place.

2.2. Viral Hepatitis Ontology Construction

The development of the viral hepatitis ontology was based mapped clinical analysis laboratory tests for diagnosing human viral hepatitis considering LOINC, OGMS ontology [Scheuermann et al 2009], IDO ontology [Cowell and Smith 2006], DOID ontology [Lynn Schriml, 2009] and FMA ontology [FMA 2012]. It describes the clinical picture throughout the disease cycle by mapping terminological items that encompass diseases, their causes, their manifestations and diagnosis.

We associate the laboratory tests with the viral infectious disease to enable the generalization of the attributes to be mined, as proposed in Figure 1. In the triage phase of Hepatitis C, for example, four specific tests may be requested for the virus identification and another eight unspecific tests may be ordered for monitoring liver functions. This situation may be generalized without having denominated each laboratory tests as an attribute for data mining.

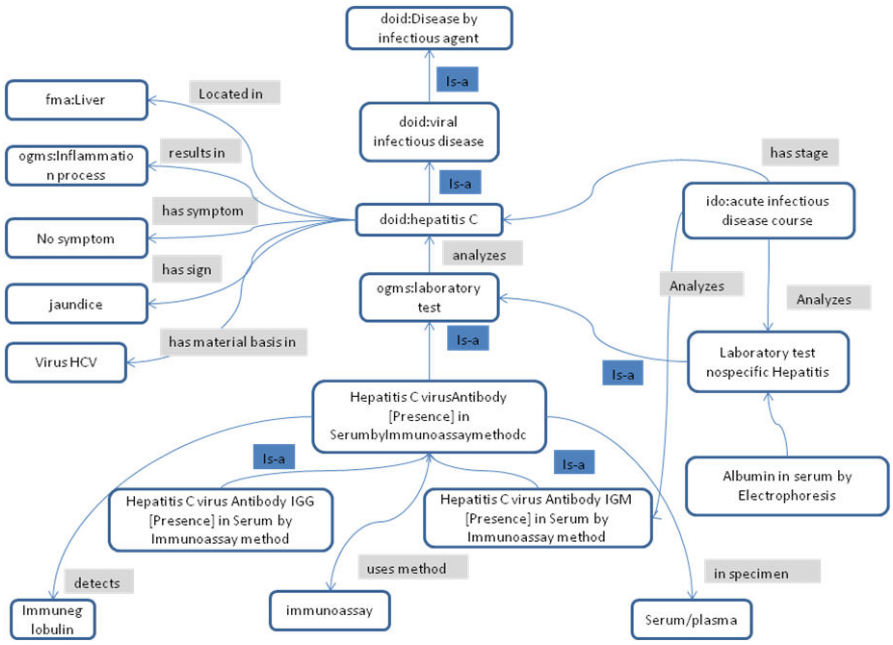

Figure 1. Example of hepatitis C classification in acute stage

The ontology for hepatitis was created at this moment of our experiment so that we could test its application in our computational architecture. We are aware that some improvements in modeling are in order, for example: i) "has symptom" and "is observed" are not ontological relations; ii) "An axioms like Hepatitis C subclass Of hasSymptom some Jaundice can be falsified by one single patient who has Hepatitis C but no jaundice"; iii) instead of "no symptom, and following OGMS, we should think in use "healthy organism". Such improvements which will be part of our future work in the following of the research.

2.3. Generalization of terms in the data mining pre-processing phase

Our study makes use of ontologies, reasoners and Jena software to promote pruning and filtering (generalization) of data from the list of laboratory tests collected from the diagnostic medicine company's database. When analyzing the relationships shared by the terms, one might identify which laboratory tests are related to which disease and stage. Therefore, the similarity between terms is considered as a means to generalize the attributes in the pre-processing phase. Figure 2 depicts the proposed model.

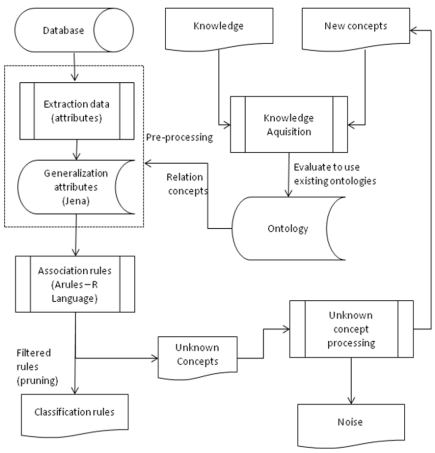

Figure 2. Model for extracting patterns of rules of association with ontologies

Based on HVO ontology, we consider relationships with the disease and with its features and also utilizing the Jena[6] tool, as well as inference rules. So, we were able to obtain more general terms to represent laboratory test groups associated with viral hepatitis diagnosis.

3. Results

The development of ontologies along with the use of inference mechanisms during the pre-mining phase has reduced the number of attributes to be mined by the association rules algorithm, namely, the Apriori algorithm. It reduced the amount of laboratory tests related to the direct diagnosis of the hepatitis virus, and also the number of unspecific tests for assessment over the course of diseases.

Based on knowledge obtained from the development of HVO, a list of laboratory test orders was collected from the company's database containing at least one test directly related to a hypothetical hepatitis virus diagnosis. Laboratory test orders that complied with the previously described rule were selected during three months, January, February and March 2015, totaling 34440 occurrences (Table 1). Test applications, which are complementary to the diagnosis, are distributed in collections made on the organization's service units, conveyed by laboratory partners throughout Brazil. In this sample, the occurrences featured 465 different laboratory tests. Considering the data of service units (support = 0.2), laboratory patterns (support = 0.02) and confidence 0.75 was executed the ARules package [Hasher 2007] in R Language to extract the association rules, we obtained the results presented below.

Table 1. Services per unit

Units	Services featuring viral hepatitis exams			Qty association rules
	Jan	Feb	Mar	Apriori
Service units	927	819	13	573
Laboratory partners	7652	7698	17331	221

Considering the same database obtained by reduction of the number of attributes, with the use of ontology, applied to 439 different laboratory tests and with the same support value and confidence was executed again the algorithm Apriori . For base units were obtained 258 and 115 rules for laboratory partners, we reached a reduction of 50% of the resulting association rules, as shown in table 2.

Table 2. Association rules generalized

Units	Services featuring viral hepatitis exams			Qty association rules
	Jan	Feb	Mar	Apriori
Service units	927	819	13	258
Laboratory partners	7652	7698	17331	115

[6] Available: < https://jena.apache.org/>.

The attributes were categorized considering the relationship between the modeled tests through equivalence axioms as showed in the example below:

hvo:laboratory_diagnostic_process_ hepatitis_A equivalent to: hvo:laboratory_testing_encounter
And (**is_composed_of** some (‘Laboratory test'
and (**diagnoses** only 'hepatitis A'))) and (**is_composed_of** min 0 ('laboratory test'
and (**diagnostic_evaluation** some Liver)))

In this equivalence axiom (described by existential restrictions), a part of the detailed diagnosis of the disease process is comprised of at least one medical application (in this case Class HVO: laboratory_testing_encounter), which is composed of complementary examinations (OGMS : laboratory test) for disease diagnostic (doid: hepatitis a) and can also be a laboratory test for evaluating the state of the health (HVO: diagnostic evaluation) of a liver (FMA: liver).

Considering the limitation of further tests and the disease is possible to identify relationships between them and, thus, promote the generalization and its representation in single attribute, in this case, a type of viral hepatitis, as shown in Table 3.

Table 3. Example of generalization

	Request – Medical prescription					Lab tests after generalization	
Patient	L.T[7]. 1	…	L.T. 8	L.T. 9	L.T. N	L.T.1	L.T. N
Patient 1	A.FETO[8]	---	TGP[9]	AU[10]	...	Hepatitis C	...
Patient 2	ALB-D[11]	---	HAV-G[12]	HAV-M[13]	...	Hepatitis A	...

Table 3 shows the laboratory tests (LT) prescribed to patients by the doctor and sent to the medical diagnostic laboratory. With the generalization attributes through the method was represented axiom disease in which some complementary tests are associated, in this case, a type of hepatitis. The other complementary tests were maintained and makes up the list of attributes analyzed by mining technique Apriori which extracted association rules related to viral hepatitis.

Therefore, our findings suggested that it is possible to reduce the rules resulting from data mining by reducing the possibilities of combining attributes. As a second, we found that the generalization of the terms enables results with a greater significance, since it can guide the post-mining phase analysis process.

4. Discussion and conclusions

The application ontology developed here strongly represents the LOINC tests for viral hepatitis, as we understand that this classification is sufficient for assessing the relationship of association rules. The extension of OGMS, DOID, FMA and IDO brings

[7]Identification of the laboratory test

[8]LOINC 1834-1 - Alpha-1 Fetoprotein – Laboratory test unspecified viral hepatitis

[9] LOINC 61151-7 - Albumin – Laboratory test unspecified viral hepatitis

[10] LOINC 5196-1 - Hepatitis B virus surface Ag

[11] LOINC 1742-6 - Alanine aminotransferase – Laboratory test unspecified viral hepatitis

[12] LOINC 5179-7 - Hepatitis A virus Ab.IgG

[13] LOINC 13950-1 - Hepatitis A virus Ab.IgM

laboratory tests closer to the diagnosis cycle and, consequently, promotes the identification of the correlations between lab tests.

It is relevant to highlight two points. Firstly, the relevance of patterns extracted by means of techniques that identify semantic similarity between terms is highly dependent of the ontology construction and validation. Therefore, it is fundamental that the domain ontologies being used be validated by specialists. Secondly, the KDD approach requires greater reach of the algorithms and cannot be restricted to “is-a” and “part-of" relations, which reinforces the use of formal semantics of ontologies.

In future work, it is intended to promote the enrichment of the ontology with new concepts and equivalence between complementary tests and disease for greater generalization of attributes.

References

Cowell, L.G, and Smith, B. (2006). Infectious Disease Ontology. In: Infectious Disease Informatics. Sintchenko V, editor. New York: Springer; 2010. pp. 373–395.

Dalfovo, O., Juarez, P., R. Alencar A., M., Palo R.D., M., J. , Otto, R., K. Silva, K. B. B. Available: <http://campeche.inf.furb.br/siic/>.

Ferraz, I. Ontology in association rules. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786067/>.

Foundational model of anatomy. Available: http://sig.biostr.washington.edu/projects/fm/.

Hasher,M., Hornik, K., Grun, B., Buchta, C., Introduction to arules – A computational environment for mining association rules and frequent item sets, 2007. Available: <http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf>.

LOINC. Available: <https://loinc.org/>.

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

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.

Piatetsky-Shapiro, G., Fayyad, U. M., Smyth, P. From data mining to knowledge discovery: an overview. Available: <http://dl.acm.org/citation.cfm?id=257942>.

Scheuermann, R. H., Werner, C., Smith, B. Toward an Ontological Treatment of Disease and Diagnosis.2009. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041577/>.

Schriml, L. (2008). Available:< http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page>

Vavpetic, A., Lavrac, N. Semantic Subgroup Discovery Systems and Workflows in the SDM. Available: <http://comjnl.oxfordjournals.org/>.