Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 360–367

Procedia Computer Science

www.elsevier.com/locate/procedia

The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2017)

Enrichment of Association Rules through Exploitation of Ontology Properties – Healthcare Case Study

Eliot Bytyçi[a], Lule Ahmedi[a,*], Francesca A. Lisi[b]

[a]*University of Prishtina "Hasan Prishtina", Rr. "George Bush", p.n.,Prishtina 10000, Republic of Kosovo*
[b]*Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro", Via E. Orabona, 4, Bari 70126, Italy*

Abstract

Association rule mining as descriptive data mining category aims to find interesting patterns on data. The quality of the patterns is measured with two metrics: confidence and support. Especially in fields dealing with sensitive data, such as healthcare, the resulting patterns should be novel and interesting. To achieve that, not only the quality of the data itself should be superior, but also other additional attributes added, do support the results. That should be achieved by using Semantic Web technologies and thus enriching data used with semantic relations between properties. A hypothesis suggests that especially tackling property relations, chain property being part of the current version of the W3C Web Ontology Language (OWL), will yield better rules. To validate the hypothesis, experiments were performed on raw data, then on an older version of OWL, which does not support the chain properties and finally on the current version of language involving chain properties. Results obtained suggest that the latter produces novel rules with strong confidence and support, not encountered in former two experiments.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: association rules, semantic web, property chains

1. Introduction

Data mining is seen as part of a bigger process which includes preprocessing, analyzing and summarizing mined knowledge for getting insight on the data related process[1]. Data mining tasks are divided into two categories:

* Corresponding author. Tel.: +381 38 554 896; fax: +381 38 42 525.
E-mail address: lule.ahmedi@uni-pr.edu

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.345

predictive and descriptive. The predictive category, as the name suggest, aims to predict future feature significances. The descriptive category on the other hand, aims to find interesting patterns describing the data. The best representative of the descriptive category is association rule mining (ARM)[2]. One of the most used algorithms of the ARM is Apriori[3]. It is used for finding rules in different kinds of databases which contain transactions, such as the items bought together, medical records of a patient, etc. Different application domains dealing with data, that are eager to find relations between those data, use the Apriori and its derivatives.

One of those domains, involving usage of massive data, is healthcare. The most important portion of the healthcare data concerns patient medical history and treatment records. By analyzing the data, healthcare industry can unleash tremendous potential and usefulness[4]. Results of the analysis can be used for patient disease diagnoses, patient profiling or even history generation. All of those support medical staff in their everyday decision making. However, even though the technology has been tried and tested and proven right, still there are concerns regarding the data itself. One of the questions that arise is: are databases used unified for all different kind of sources?

Kolias et al.[5] believe that by using Semantic Web technologies the above mentioned concern could be overcome. Furthermore, the community would be able to retrieve hidden knowledge that remains unexploited in vast and diverse pools of medical data. Supportive to that is the fact that Semantic Web technologies provide tools to more accurately and effectively integrate and process data. One of the basic components of the Semantic Web are collections of information, called ontologies[6].

Additionally, ontologies could be used to infer new relations by usage of the advanced properties of the concepts described. On the basis of that fact, a hypothesis could be constructed that supports the claim that usage of advanced properties would led to new stronger and more confident rules. The hypothesis could be verified with experiments involving the application of ARM algorithms on several types of data. First, the ARM would be applied to raw patient data and secondly on ontology populated patient data but leaving aside advanced properties. In the end, experiments would be performed on ontology populated data enriched with advanced properties.

The rest of the paper is organized as follows: Section 2 describes the related work, Section 3 provides background on OWL2 and ARM, Section 4 presents the ontology created and used, Section 5 introduces the setup and the results of the experiments performed and Section 6 discusses concluding remarks and future research directions.

2. Related work

Abedjan and Naumann[7] suggest that ARM benefits from integration and usability of semantic data. One of the mining configurations mentioned is mining predicates. Mining predicates will result in gaining dependencies among elements of the schema or the properties of the subjects. In relation to our case, by introducing further object properties, we gain new relations between elements, which as will be shown, results into better rules.

Nebot and Berlanga[8] use schema level knowledge encoded in ontologies to derive appropriate transactions, which are then fed into traditional ARM algorithms. Experiments were performed on semantic data of biomedical application and showed usefulness and efficiency of the approach. Still, the approach required domain experts to define targets and contexts of the mining process, so that the correct transactions are generated.

Bytyçi et al.[9] use a different approach by enriching association rules with context ontologies. In order to evaluate their approach, they compare results obtained from deploying association rule mining in raw data and in context ontology derived data. Results inclined towards the context data association rule mining, present new rules not otherwise inferred following the other approach.

J´ozefowska et al.[10] presents a hybrid method of combining Semantic Web ontologies expressed in description logic as well rules in disjunctive Datalog. Ontology level association rules are extracted, presented as queries in Datalog. Used description logic SHIF[11] is the DL logic of OWL lite which then with DL-safe rules represents Semantic Web Rule Language (SWRL) rules but restricted to known individuals. Furthermore, the approach ignores mining of rules over predicates.

Lisi[12] performed research on mining the Semantic Web present in the fields of inductive logic programming (ILP) and generalization that make use of the description logic of a knowledge base. They concentrate on mining answer-sets of queries towards a knowledge base. Based on a general reference concept, additional logical relations are considered for refining the entries in an answer set.

Galarraga et al.[13] present AMIE, a rule mining system that extracts logical rules (in particular Horn clauses) based on their support in a knowledge graph. In contrast to other approaches, AMIE can handle the open-world assumption of knowledge graphs. Furthermore, it has been shown to be much faster on the large knowledge graphs. Authors claim that in a number of experiments, AMIE showed to be the most efficient and effective approach to generate new facts.

In our case, compared to the above mentioned approaches, the logic used is SROIQ, i.e. the description logic underlying OWL2[14], which is far more expressive; no other DL-safe rules or Datalog programs are used. Even though, in[13], Horn clauses were used in order to overcome the weakness of OWL1[11] in expressing more complex relations. This is similar to our approach of using property chains of OWL2. In both cases, new facts are generated.

3. Background

In the following sub sections, a brief introduction of the ontologies and association rule mining will be presented. Ontologies are used to model the concepts, while ARM is used to infer new knowledge on the relations of the data.

3.1. Web ontology language

Ontologies as a component of Semantic Web collect information, by using formal specification describing its concepts and relations between them[15]. W3C as the international organization for World Wide Web (WWW) offers a large palette of techniques to describe and define different forms of vocabularies in a standard format. Among those is also Web Ontology Language (OWL) as a semantic language for publishing and sharing ontologies in WWW. The current version of OWL is OWL2[14].

OWL2 Web Ontology Language is an extension and revision of OWL1 Web Ontology Language[11]. In overall OWL2 has a very similar structure to OWL1 but it adds new functionalities. Those functionalities offer new expressivity including: keys, self-restriction, qualified cardinality restrictions and property chains, being few of them.

Keys are used to uniquely identify individuals of a given class by values of key properties. Another functionality is the self-restriction which describes all the individuals connected to themselves by an object property expression.

Object cardinality restrictions are of several kinds describing minimum cardinality, maximum cardinality or exact cardinality. Those restrictions help in forming class expressions. Even though those were allowed in OWL1 for restrictions on the number of instances of a property, still they couldn't provide a means to restrain the class or data range of the instances to be counted. In OWL2, qualified and unqualified cardinality restrictions are possible[14].

In our case, we gave a special attention to the property chains. This property, as an object property, is specific to OWL2 and provides means to define property as a composition of different other properties[14]. An example of the chain property presented in[14] is: if x is located in y and y is part of z then x is located in z, for example a disease located in a part is located in the whole. This helps in deriving new facts by means of reasoning.

3.2. Association rule mining

Recently, ARM is commonly being used as well over ontological derived data, even though it was originally coined to find frequent items over a transactional database. Even though ARM is a simple technique, it provides very useful insights on the data. The goal of the technique is to find rules, presented in the form $X \Rightarrow Y$. Left hand side of the rule is known as antecedent and the right hand side as consequent, otherwise also known as *if-then* rules. Suggested rules have to satisfy two constraints: support and confidence. Support can be defined as the fraction of transactions in a transaction database that satisfies the union of items in the consequent and antecedent of the rule[2]. Further, confidence describes the ratio between the support of the rules and the support of the antecedent. In other words, support describes frequency of the rule, while the confidence the number of times that the rule was found to be true[2].

One of the most used algorithms of ARM is Apriori[3]. In Apriori the mining process consists of two steps: candidate generation and candidate evaluation step. A candidate generation step involves extension of frequent

subsets which meet the condition of minimum support, one item at the time, tested against the data. Candidate evaluation step, in the other hand, is responsible for filtering candidate patterns with insufficient support[10].

4. A proposal of a heart condition ontology

The description of concepts is done through ontologies and their engineering should be done on top of reusable ontologies and maybe ensuring a specialized form. The specific ontology used was found to be most suitable for this stage of the work, since we could easily create different properties and especially chain properties on the objects. For further research, a broader ontology could be taken into account such as Heart Failure Ontology (HFO)†, but for this specific case, the ontology fulfils the criteria for the evaluation of the hypothesis.

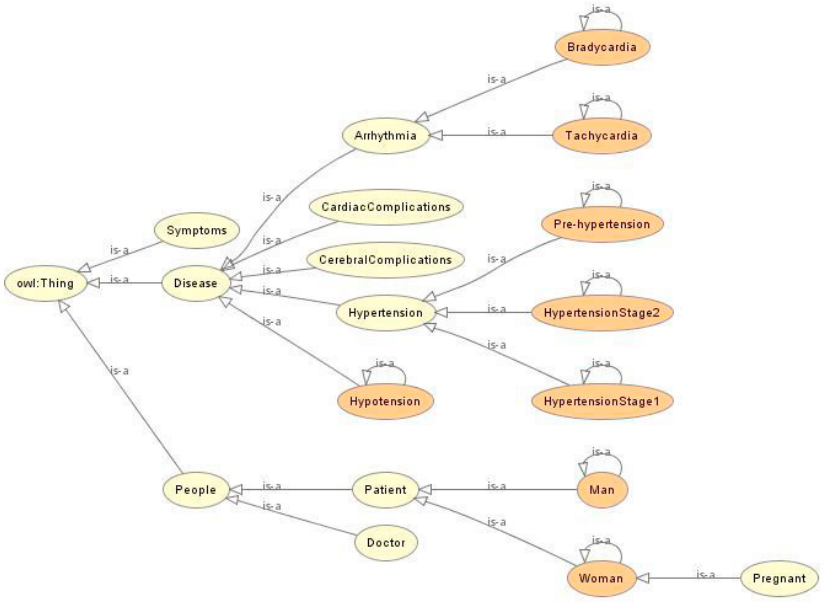

Fig. 1. Heart condition ontology.

Data itself was the biggest problem. Not only due to privacy concerns but also because of quality and quantity. The datasets found had only few hundred patient records and had no explanation about the quality of the data. Therefore, we have simulated data for the use case. We have created 4000 patients records randomly providing the age, gender and heart condition factors such as blood pressure, pulse and as well the symptoms related to heart disease.

The final ontology presented in Figure 1 by showing the taxonomy with its classes and subclasses, describes the people involved in the process – class *People* and their subclasses *Patient* and *Doctor* and the disease itself as well.

The *Disease* class models few of the most common complications related to the heart, which were chosen for the fact of their relation to the patient data. For example, *Arrhythmia* relates to the heart rate and it can be divided into two categories: the class describing that the heart rate is too slow i.e. *Bradycardia,* and the one describing that the heart rate is too fast i.e. *Tachycardia*. The *Hypertension* class, known as high blood pressure, on the other hand, relates to the blood pressure, and is divided into three stages: *Pre-hypertension*, *HypertensionStage1* and *HypertensionStage2*. In all of the cases, restrictions describing the condition are set into the class and subclass, in order to have inferences when the reasoning is performed. In the end, several properties such as *isSickOf*, *treatedBy*

† http://purl.bioontology.org/ontology/HFO

etc. describing the relations between data are modeled into the ontology, but initially none of specific OWL2 properties. The ontology as it is presented was used for the second part of the experiment: ontology data without specific / special object properties.

For the third part of the experiment, we have added several other object properties to the ontology, which are characteristic of OWL2[14]. In order to prove the claims, one object property was paid a special attention: property chains. In our case, the property chain *mayVisit* is composed by two other properties *isSickOf* and *treatedBy*. Object property *isSickOf* gives an overview of the *Disease* of the *Patient*. The other object property *treatedBy* models the relation between the *Doctor* and the *Disease*. With the composition of the two properties, one could suggest to the *Patient* the *Doctor* he or she *mayVisit*.

5. ARM setting and experiment results

The patient data concern patient's vital characteristics such as: age, gender, blood pressure, pulse, diseases that they may have and other symptoms. Each patient can be seen as a transaction, in which every description element can be seen as a particular item. Any particular item has a specific relation with another item, described as an association rule that could be of the form *age --> diabetic*, where age and diabetic are items. This exemplary rule suggesting that people with specific age may suffer from diabetes, can be derived from an itemset: *{age, diabetes}*, with a high confidence. The dataset was generated with the help of an online tool[‡], since we were unable to find a dataset with more than 100 patient records.

But, first in order to perform the ARM task, the measured values describing patient properties have to be binned into several bins. In our case study, we have binned data into three bins, regulated manually by the WEKA[§] tool. That not only helps with removing noisy data, but also suggests stronger relations between attributes. Therefore, the attribute age, was binned according to the data into three categories, which could be coded as young, middle aged or old, allowing stronger rules to be drawn.

After binning the attributes, we have specified the minimum support and confidence thresholds. The minimum support was set to 10% and the minimum confidence was set to 30%. Then we have deployed the data (raw data, ontology derived data) in the Apriori algorithm, which resulted in different sizes of itemsets found. In some of the cases, a manual search of the itemsets was performed in order to find more interesting rules. Interestingness of the rules can be described as a concept emphasizing several criteria, such as novelty and surprisingness[16]. It should be emphasized that the obtained results are based on generated data and therefore are not to be used to describe possible real life rules.

5.1. Results from raw data

The results presented in Table 1, show just a few of the more than 100 hundred rules obtained, when Apriori was applied on raw data. One of the best rules found, is the one describing the relation between a "high" pulse and the female patients. It has a support of around 13%, the rule has been found in 514 transactions from 4000 of them in total. The confidence is higher than the threshold provided by us and it is 55%. Since neither support nor confidence are of a very high value, it cannot be considered a good rule. The other rules presented, each one of them had lower confidence and therefore we had picked just 10 of them to present.

[‡] https://mockaroo.com/
[§] www.cs.waikato.ac.nz/ml/weka/

Table 1. Rules obtained from raw data.

Rule	Best rules found:	Confidence
1	pulse='(106-inf)' 942 ==> gender=Female 514	conf:(0.55)
2	systolic='(164-inf)' 785 ==> diabetic=TRUE 418	conf:(0.53)
3	isSickOf=Disease_3 834 ==> diabetic=FALSE 440	conf:(0.53)
4	smoker=FALSE diabetic=TRUE 1028 ==> gender=Female 542	conf:(0.53)
5	systolic='(-inf-116]' 770 ==> smoker=FALSE 405	conf:(0.53)
6	age='(49-66]' 793 ==> diabetic=TRUE 417	conf:(0.53)
7	systolic='(148-164]' 976 ==> smoker=FALSE 513	conf:(0.53)
8	systolic='(116-132]' 984 ==> smoker=TRUE 516	conf:(0.52)
9	pulse='(78-92]' 1103 ==> gender=Male 578	conf:(0.52)
10	distolic='(108-inf)' 1079 ==> gender=Female 565	conf:(0.52)

5.2. Results from ontology data

The second part of the experiment, with the same settings, was performed on the ontology data. So, first of all the ontology created by following the standard OWL1[11] was populated with the data describing patient's attributes. After the reasoning on the ontology, new relations between data were inferred, which resulted in better rules gained. Those derived attributes and the data were transferred into an appropriate format for further analysis with the Apriori algorithm.

As seen in Table 2, where we have extracted some of more than 1000 rules gained, they all show a high support (the first rules more than 50%) and even higher confidence of 100%. In order to have the same experiment settings for all parts of it, we have let the support and confidence be as low as described in the beginning of this Section. Otherwise, in this specific case we could have raised the support and confidence minima. As for the rules gained from the process, we have chosen just some of them manually from the results pool, to present in Table 2. Almost identical rules can be found in the third part of the experiment.

Table 2. Rules obtained from OWL1 enriched data.

Rule	Best rules found:	Confidence
1	gender/0/@value=Female 2039 ==> @=#Patient 2039	conf:(1)
2	diabetic/0/@value=TRUE 2000 ==> @=#Patient 2000	conf:(1)
3	diabetic/0/@value=FALSE 2000 ==> @=#Patient 2000	conf:(1)
4	gender/0/@value=Male 1961 ==> @=#Patient 1961	conf:(1)
5	hasSymptoms/0/@id=#Dizziness 1175 ==> @=#Patient 1175	conf:(1)
6	pulse/0/@value='(78-92]' 1103 ==> @=#Patient 1103	conf:(1)
7	distolic/0/@value='(108-inf)' 1079 ==> @=#Patient 1079	conf:(1)
8	smoker/0/@value=FALSE #gender/0/@value=Female 1053 ==> @=#Patient 1053	conf:(1)
9	diabetic/0/@value=TRUE #gender/0/@value=Female 1042 ==> @=#Patient 1042	conf:(1)
10	distolic/0/@value='(-inf-72]' 1031 ==> @=#Patient 1031	conf:(1)

5.3. Results from ontology data enriched with object properties

In the third part of the experiment, the ontology was re-created in OWL2. One of the main additions was the usage of the *inverseOf* property and the chain property. The former added a big number of attributes, even though they did not have any impact on the generated rules. To verify the impact of the *inverseOf*, we have performed another experiment without the *inverseOf* property and our claim was validated: there was no impact on the generated rules and furthermore, the experiment concluded much faster. Even though the claim itself should be further investigated, we believe that this resulted in our case as such, since the attributes inferred from the *inverseOf* property were the same and brought nothing new to the data, which resulted in no other important rules obtained.

On the other hand, the property chain added additional attributes to the data bringing the total number of the attributes to 19. Some of the best rules obtained are presented in Table 3. The support was strong and the confidence as well, which can be seen from the first rules presented in Table 3, were support is 1138 and confidence 100%. Most of the rules obtained were similar to the ones found at Table 2, except few additional ones, obtained as a direct result of usage of the chain property. One of the rules obtained suggest that a patient may visit a specific doctor, in relation to his/her disease, presented in Table 3 in row 10.

This usage of pre-defined object properties of OWL2, not only made possible the generation of new facts and making thus attributes more complete but also the property could be used to identify potential errors if the rule is not as one expected. Furthermore, it helps in better understanding the data by describing consistency.

Table 3. Rules obtained from OWL2 enriched data.

Rule	Best rules found:	Confidence
1	Disease @=#People 1138 ==> @=#Patient 1138	conf:(1)
2	People 1473 ==> @=#Patient 1473	conf:(1)
3	Patient 1473 ==> @=#People 1473	conf:(1)
4	Disease 1421 ==> @=#Bradycardia 1421	conf:(1)
5	Bradycardia 1421 ==> @=#Disease 1421	conf:(1)
6	Woman 1383 ==> #gender/0/@value=Female 1383	conf:(1)
7	People 1307 ==> @=#Patient 1307	conf:(1)
8	Patient 1307 ==> @=#People 1307	conf:(1)
9	Tachycardia 1159 ==> @=#Disease 1159	conf:(1)
10	mayVisit/0/@id=#Doctor3 834 ==> #isSickOf/0/@id=#Disease 3 834	conf:(1)

6. Discussion and future work

The association rules mined represent only a fraction of rules that can be discovered using patient heart data. More rules can be found by using different search criteria or by lowering further the support or using another larger database. But, from the dataset used in our experiments, we could obtain new rules by using an ontology to represent patient data and their relations. Furthermore, we believe that by exploiting especially object properties available in OWL2, the attributes will be enhanced and thus the results will offer new rules - as presented in Section 3.

We have demonstrated that claim, with the usage of property chains. Comparing to the results obtained in the same setting environment but on raw data and on ontology data – without usage of properties, the improvement was obvious: in both number of rules and the interestingness of the rules. Some of those novel rules obtained could be seen as “life changing”. This increase on the attributes was supported as well from other authors, which claim that the greater the number of attributes, the greater the combinatorial explosion caused to obtain frequent item sets[17]. Thus in our case, by only using one property restriction offered by OWL2, the number of attributes increased, resulting in an augmented number of rules.

In the future, we tend to go through all the predefined object properties of OWL 2 and find which ones are more convenient and interesting for the creation of new rules. Furthermore, we tend to investigate the correlation between the usage of object properties and conditional functional dependencies.

References

1. J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
2. R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Acm sigmod record*, 1993, vol. 22, no. 2, pp. 207-216: ACM.
3. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, 1994, vol. 1215, pp. 487-499.
4. D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," *International Journal of Bio-Science and Bio-Technology,* vol. 5, no. 5, pp. 241-266, 2013.
5. V. D. Kolias, J. Stoitsis, S. Golemati, and K. S. Nikita, "Utilizing Semantic Web Technologies in Healthcare," in *Concepts and Trends in Healthcare Information Systems*: Springer, 2014, pp. 9-19.
6. T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american,* vol. 284, no. 5, pp. 28-37, 2001.
7. Z. Abedjan and F. Naumann, "Improving rdf data through association rule mining," *Datenbank-Spektrum,* vol. 13, no. 2, pp. 111-120, 2013.
8. V. Nebot and R. Berlanga, "Finding association rules in semantic web data," *Knowledge-Based Systems,* vol. 25, no. 1, pp. 51-62, 2012.
9. E. Bytyçi, L. Ahmedi, and A. Kurti, "Association Rule Mining with Context Ontologies: An Application to Mobile Sensing of Water Quality," in *Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*, 2016, pp. 67-78: Springer.
10. J. JÓzefowska, A. Lawrynowicz, and T. Lukaszewski, "The role of semantics in mining frequent patterns from knowledge bases in description logics with rules," *arXiv preprint arXiv:1003.2700,* 2010.
11. D. L. McGuinness and F. Van Harmelen, "OWL web ontology language overview," *W3C recommendation,* vol. 10, no. 10, p. 2004, 2004.
12. F. A. Lisi, "AL—QUIN: An Onto—Relational Learning System," *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications,* p. 52, 2013.
13. L. Galárraga, C. Teflioudi, K. Hose, and F. M. Suchanek, "Fast rule mining in ontological knowledge bases with AMIE+," *The VLDB Journal,* vol. 24, no. 6, pp. 707-730, 2015.
14. C. Golbreich, E. K. Wallace, and P. F. Patel-Schneider, "OWL 2 Web Ontology Language new features and rationale," *W3C working draft, W3C (June 2009) http://www. w3. org/TR/2009/WD-owl2-new-features-20090611,* 2009.
15. T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition,* vol. 5, no. 2, pp. 199-220, 1993.
16. L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys (CSUR),* vol. 38, no. 3, p. 9, 2006.
17. J. M. Luna, J. R. Romero, and S. Ventura, "Design and behavior study of a grammar-guided genetic programming algorithm for mining association rules," *Knowledge and Information Systems,* vol. 32, no. 1, pp. 53-76, 2012.