Mining Contextual Rules to Predict Asbestos in Buildings

Thamer Mecharnia[1,2], Nathalie Pernelle[3], Celine Rouveirol[3], Fayçal Hamdi[4], Lydia Chibout Khelifa[2]

[1] LISN, Université Paris Saclay, France
thamer@lri.fr
[2] Centre Scientifique et Technique du Bâtiment (CSTB), France
firstname.name@cstb.fr
[3] LIPN, Université Sorbonne Paris-Nord, CNRS UMR 7030, France
firstname.name@lipn.univ-paris13.fr
[4] CEDRIC, Conservatoire National des Arts et Métiers, Paris, France
faycal.hamdi@cnam.fr

Abstract. In the context of the work conducted at CSTB (French Scientific and Technical Center for Building), the need for a tool providing assistance in the identification of asbestos-containing materials in buildings was identified. To this end, we have developed an approach, named CRA-Miner, that mines logical rules from a knowledge graph that describes buildings and asbestos diagnoses. Since the specific product used is not defined, CRA-Miner considers temporal data, product types, and contextual information to find a set of candidate rules that maximizes the confidence. These rules can then be used to identify building elements that may contain asbestos and those that are asbestos-free. The experiments conducted on an RDF graph provided by the CSTB show that the proposed approach is promising and a satisfactory accuracy can be obtained.

Keywords: Rule Mining · Knowledge Graph · Temporal Data · Asbestos.

1 Introduction

Asbestos[5] has been known to be harmful for quite a long time, nevertheless, the dangers associated with it have only been identified since the beginning of the 20^{th} century. Breathing the air that contains asbestos fibers can lead to asbestos-related diseases, such as lung cancer and chest lining. However, for its fireproof qualities, many countries have extensively used asbestos in buildings, especially from 1950 to 1970. The use of asbestos is illegal today in many countries, but several thousand tons have already been used in the past and asbestos is still present in a considerable number of buildings. Thus, the identification of building

[5] https://en.wikipedia.org/wiki/Asbestos

parts containing asbestos is a crucial task. As part of the PRDA[6], the CSTB[7] has been asked to develop an online tool assisting in the identification of asbestos-containing materials in buildings that aims to guide the tracking operator in the preparation of its tracking program (called the ORIGAMI Project). Professionals regularly inspect buildings and collect samples to detect the presence of asbestos in building components. However, information is needed to prioritize a large number of possible tests. The problem is related to the fact that the available building descriptions only contain the classes of used products without giving their accurate references or any other information about them (i.e., valued properties, providers, etc.). In [8], an ontology-based approach that estimates the probability of the existence of asbestos products in a building is defined. To generate this probability, this hybrid approach combines statistical and rule-based methods. However, it is not fully effective since it relies on the construction year of the building, and on reliable but incomplete external resources describing some product references that were used in that period of time. Considering expert feedbacks, we know that the context in which a product is used can also be relevant to predict the presence of asbestos in this product (i.e., the characteristics of the building, the nature of the building components in which the product appears, other products used in the same component, etc.). Recently, the CSTB has made available a set of diagnoses conducted on a large number of buildings. These data have been represented using the Asbestos Ontology proposed in [8] and can be used to learn prediction rules. Many rule mining approaches have been proposed that can learn to classify data based on RDF descriptions [6,10,5]. However, none of them use the ontology semantics, the part-of relations, and the numerical built-in predicates that are needed to represent the context and the temporal constraints.

In this paper, we propose an ontology-based approach to discover rules that can be used to detect the products that contain asbestos in a building or not. The proposed approach focuses on rule premises that describe the product, its context, and the temporal constraints expressed as open intervals. The potentially relevant predicates that can appear in the context are declared by the expert and heuristics are defined to limit the search space when multi-valued part-of relations are exploited. Furthermore, general knowledge about how the asbestos usage evolves through time is exploited to generate the temporal interval.

The rest of this paper is organized as follows. In section 2, we present related works. In section 3, we describe the Asbestos Ontology. Then, in section 4, we present our predictive approach. Section 5 presents the results obtained on a real data set of diagnoses. Finally, Section 6 draws conclusions and outlines the future research directions.

[6] Asbestos Research and Development Plan launched by the Department of Housing, Town Planning and Landscapes (DHUP), attached to the General Directorate of Planning, Housing and Nature (Minister of Housing and Sustainable Habitat)

[7] French Scientific and Technical Center for Building. `http://www.cstb.fr/en/`

2 Related Work

In the context of Knowledge Graphs (KG), rule mining can be used to enrich graphs (by means of link or type prediction, adding new axioms, entity linking), or to detect erroneous RDF triples. To have scalability properties, most of the recent approaches for link or type predictions are based on deep learning methods and embedding that allow translating high-dimensional vectors into relatively low-dimensional spaces [11]. Nevertheless, other applications that need interpretable rules to understand and maintain some domain knowledge, are still in awe of discovering logical rules. Many approaches have addressed the problem of discovering first-order logic (FOL) rules over relational data [4,12] or texts [14]. However, different approaches and hypotheses are needed to discover rules in knowledge graphs. Indeed, data is generally incomplete and counter-examples are not always available (i.e., due to the open-world assumption, it cannot be assumed that a fact that is not in a KG is false). Besides, ontology semantics can be exploited when available. Some unsupervised approaches aim to discover graph patterns in voluminous RDF graphs without taking into account the ontology [6,10]. [6] uses an optimized generate and test strategy while controlling the search space by limiting the number of atoms that appear in the rule. [10] allows to discover FOL rules that may involve SWRL[8] built-in predicates to compare numerical or string values, and negation to identify contradictions. However, these values must be defined in the KG and associated with two variables of the rule. So, the approach does not allow us to discover a reference constant like "*age (X, a), a* $\geq$ *18* $\rightarrow$ *adult (X)*", which is one of the goals in our application. Both approaches are based on a Partial Completeness Assumption (PCA) implying that when at least one object is represented for one entity and one property, all the objects are represented, and others can be considered as counter-examples. The classification approaches based on FOLDT (First-order logical decision tree) such as TILDE [1] (Top-down induction of logical decision trees) are based on decision trees in which nodes can share variables and involve numerical predicates with threshold values. However, these latter approaches do not use the semantics of ontology in the exploration of the search space. Other approaches such as [2] can be guided by the ontology semantics to avoid constructing semantically redundant rules. However, the author has shown that the exploitation of reasoning capabilities during the learning process does not allow mining rules on large KGs.

Some approaches have focused on learning DL concept descriptions such as DL-FOCL [13], an optimization of DL-FOIL [3], or CELOE [7]. Such approaches are generally based on a separate-and-conquer strategy that builds a disjunct of partial solutions, where partial solutions are specialized using refinement operators, so that the description correctly cover as many positive instances as possible while ruling out all the negative ones. To learn expressive DL descriptions, they exploit refinement operators that return a specialization expressed through the $\mathcal{ALCO}$ operators (i.e., universal and existential restrictions on roles, intersec-

[8] `https://www.w3.org/Submission/SWRL/`

tion, complement, union, one-of enumerations and roles with concrete domains). However, they do not allow to take into account class instance properties that change over time. [8] discovers rules that can predict the presence of asbestos considering the construction year of the building, but this statistic and semantic hybrid approach is based on incomplete external web resources that describe how the presence of asbestos in frequently used marketed products has evolved during the last century.

In this work, we aim to discover interpretable classification rules from positive and negative examples described in the Asbestos knowledge graph provided by the CSTB. These horn-clause rules will be used to evaluate the presence of asbestos (negative or positive) in a product. Since the marketed products that have been used in the buildings are unknown, the rules exploit a product's context defined by domain experts to express those elements of this context have a potential impact on the presence of asbestos : (1) part-of properties to take into account the building components and the others used products, (2) the building's construction year. Since the building's construction year can have an important impact on the presence of asbestos, a rule can use SWRL comparison operators to compare a variable year to the reference year (ex. SWRL:lowerThanOrEqual(YEAR, ref_year)) which is the reference year that maximizes the quality of the rule. None of the previously mentioned approaches allow exploiting the ontology and such built-in numerical predicates in the resulting rules. These rules will be transformed in SWRL so that the expert can predict the presence of asbestos in buildings using an existing reasoner.

3 Asbestos Ontology

In this section, we briefly present the upper part of the Asbestos Ontology (see Figure 1) that has been proposed in [8] based on the CSTB documentation resources, expert knowledge and the needs in terms of prediction. The main concepts of this ontology are:

- Building: a construction characterized by a code (CSTB internal code that corresponds to a given type of building: school, housing, etc.), the building type, the construction year, an address.
- Structure: building subspace (e.g. balcony, staircase, roof, etc.).
- Location: indicates a basic element that belongs to a building structure (e.g., door, window, wall, etc.).
- Product: describes a product that can be used in the composition of locations (e.g., glue, coating, etc.). A product is described by its name.
- Diagnostic Characteristic : specifies the results of the existence of asbestos test when it exists. The value of *has_diagnostic* is "positive" when the product contains asbestos and "negative" otherwise.
- Predicted Characteristic : can be used to store that it is predicted that a product is asbestos-free or not.

The Asbestos Ontology describes 8 subclasses of structure, 19 subclasses of location and 38 subclasses of product.

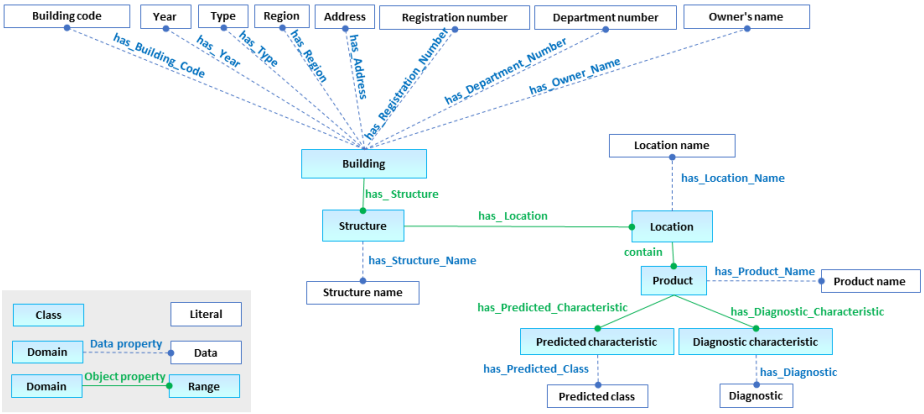

Fig. 1: Main Concepts of the Asbestos Ontology

4 CRA-Miner Approach

In this section, we first describe the contextual rules for the asbestos prediction that we want to provide to experts to help them detect asbestos-containing materials in a building. We then present the CRA-Miner algorithm that generates these rules from the populated Asbestos Ontology.

4.1 Contextual Rules for Asbestos prediction

A Contextual Rule for the prediction of Asbestos (CRA) is a conjunction of predicates that concludes on the presence or absence of Asbestos in a product P. We consider a context-free upper bound of the search space $\top$ that is defined as $product(P), has\_diagnostic\_characteristic(P, D) \rightarrow has\_diagnostic(D, Value)$. The set of contextual rules that can be constructed from $\top$ is defined using a conceptual context that is by default the whole ontology. However, a domain expert can delimit the context by choosing the classes and attributes that can impact the presence of asbestos. This selection can either be used to discard irrelevant ontology elements or to test new hypotheses. The chosen context, that corresponds to the language bias in Inductive Logic Programming (ILP) [9], will then be used to specialize a rule.

Definition 1. *(Conceptual context) A conceptual context CO is a sub-graph of the ontology, i.e. a subset of classes and properties, that specifies the ontology elements that can be used in the body of the rule.*

Example 1 Let's $CO = \{product, location, structure, contain, has\_location, has\_region, has\_year, has\_structure, has\_diagnostic\_characteristic\}$.

A contextual rule is based on the ontology vocabulary defined in the conceptual context and the specializations of the $SWRL{:}CompareTo$ predicate that

can be added to introduce temporal constraints for the building year (i.e. open intervals):

Definition 2. *(Contextual rule) Let CO be a conceptual context, a contextual rule is a rule* $\overrightarrow{B} \rightarrow h$*, where* $\overrightarrow{B} = \{B_1, B_2, ..., B_n\}$*, and* $\forall B_i \in \overrightarrow{B}, \exists B_j \in CO \cup \{SWRL{:}CompareTo\}$ *s.t.* $B_i \sqsubseteq B_j$ *and h is the predicate has_diagnostic instanciated by the value "positive" or "negative".*

A contextual rule must be closed and connected as defined in rule mining approaches such as [6].

Example 2 A closed and connected rule that can be discovered using the context defined in example 1 is:

glue(P), contain(L, P), has_location(S, L), painting(P2), contain(L, P2), has_structure(B, S), has_year(B, Y), has_region(B, "Paris"), lessThanOrEqual(Y, "1950"), has_diagnostic_characteristic(P, D) → has_diagnostic(D, "positive")

This rule expresses that a glue that appears in a building located in Paris and constructed before 1950, when used in the same location as that of a painting, it contains asbestos.

Additional constraints are defined to reduce the search space complexity for multi-valued properties that describe part-of relations: *contain*, *has_location*, and *has_structure*.

First, the expert can define the maximum number of occurrences of co-located building components that can appear in the body of the rule : $maxSibS$ is used to define the maximum number of sibling structures, $maxSibL$ is the maximum number of sibling locations, and $maxSibP$ is the maximum number of sibling products.

Example 3 If the expert considers that the co-located structures cannot affect the presence of asbestos in P, then $maxSibS = 0$ and the approach will not build the following rule since the $S2$ structure should not be considered (sibling of $S1$ that contains the target product for the *has_structure* property):

Coating(P), contain(L, P), Location(L), has_location(S1, L), Vertical_Separator(S1), has_structure(B, S1), has_structure(B, S2), Floor(S2) , has_year(B, Y), has_region(B, "Lyon"), SWRL:lessThanOrEqual(Y, "1963"), has_diagnostic_characteristic(P, D) → has_diagnostic(D, "positive")

Furthermore, the CSTB experts consider that only specific co-located components can affect the marketed product used for P and therefore the presence of asbestos. For instance, the presence of a coating in the same location than a target glue is not significant, while the presence of a floor coating can impact the marketed glue that has been used in this location. A similar hypothesis is assumed for locations and structures. Therefore, only the most specific classes

can be added for sibling products, sibling locations or sibling structures involved in the considered part-of relations.

To evaluate the quality of a rule, we use the classical measures of *head coverage* [6] and *confidence* that has been defined in the relational setting.

The *head coverage* (hc) is the ratio between the support, i.e. the number of correct predictions $has\_diagnostic(D,$"positive") (resp. $has\_diagnostic(D,$"negative")) implied by the rule, and the number of diagnoses $has\_diagnostic(D,$"positive") (resp. $has\_diagnostic(D,$"negative")) that appear in the knowledge graph:

$$hc(\overrightarrow{B} \rightarrow has\_diagnostic(D, val)) = \frac{supp(\overrightarrow{B} \rightarrow has\_diagnostic(D,val)}{\#(D,val):has\_diagnostic(D,val)}$$

The confidence ($conf$) is defined as the ratio between the support of the rule and the number of diagnoses that can participate to an instanciation of the body of the rule.

$$conf(\overrightarrow{B} \rightarrow has\_diagnostic(D, val)) = \frac{supp(\overrightarrow{B} \rightarrow has\_diagnostic(D,val))}{\#D:\exists X_1,...,X_n:\overrightarrow{B}}$$

CRA-Miner aims to discover all the most general rules that conform with the defined language bias and such that $hc \geq minHc$ et $conf \geq minConf$.

4.2 Evolution of the presence of Asbestos over time

It has been shown in [8] that the number of marketed products that contain asbestos remains stable until 1972, and then decreases to zero in 1997 when its usage is prohibited in France. Indeed, these products have either become asbestos-free or their production has been discontinued. Thus even if the probability of asbestos varies depending on the class products, we know that this probability decreases over time. So if a contextual rule concludes on the absence of asbestos for a product used in a building constructed after a given year Y_1, the confidence can only increase for $Y_2 \geq Y_1$. This property is used to prune the search space when the predicate $greaterThanOrEqual$ or $lessThanOrEqual$ is generalized.

4.3 Algorithm CRA-Miner

The aim of the algorithm CRA-Miner is to generate from the positive and negative examples described in the KG all contextual rules such that $hc \geq minHc$ and $conf \geq minConf$. These rules will be used to predict the presence or absence of asbestos in products that have not been tested.

CRA-Miner is a top-down generate and test algorithm that specializes in the upper bound T of the search space by considering the hierarchy of product classes, and by adding temporal constraints, and constraints constructed using the conceptual context. The algorithm takes as input the knowledge graph,

the language bias, the thresholds $minConf$ and $minHc$, and the thresholds $maxSibP$, $maxSibL$ and $maxSibS$ that are used to limit the number of co-located components that can be considered in the rule. The result is a set $\mathcal{CR}$ of contextual rules. The exploration of the search space is guided by the subsumption relations of the ontology (top-down exploration of the targeted products, their locations, and their structures) and the construction of the temporal constraints exploiting the fact that the number of products containing asbestos decreases over time.

At each specialization step, the generated rules that obtain a confidence value and a head coverage higher than the specified thresholds, and that improve the confidence value of the rule from which it is derived, are stored in $\mathcal{CR}$. For all the rules such that $conf = 1$ or $hc < minHc$, the specialization will stop.

We describe the algorithm's steps for the most general context that has been defined by the CSTB experts, in other words, the context CO defined in example 1. In this context, the rule can predict the presence of asbestos in a product using the building's construction year, the region the building is located in, and all its types of components. The five specialization steps of the algorithm are as follows :

1- **Specialization of T with sub-classes of products** :

During this step, we replace the class *product* by more specific classes (e.g. coating, glue, painting) to generate all the context-free rules that only depends on the type of product used. The top-down exploration stops when $hc < minHc$.

2- **Specialization with the temporal constraint** : For each context-free rule generated by the previous step, we add to the body of the rule the path of properties that is needed to reach the construction year from the target product P: $has\_location(S, L), contain(L, P), has\_structure(B, S), has\_year(B, Y)$. The predicate *SWRL:lowerThanOrEqual(Y, y)* (for a rule that concludes on "positive") or *SWRL:greaterThanOrEqual(Y, y)* (for "negative"), is also added to compare the construction year Y to a reference year y which maximizes the confidence and preserves $hc \geq minHc$.

For example, if the rule R1 is generated in step 1:
R1 : *coating(P), has_ diagnostic_ characteristic(P, D)* $\rightarrow$ *has_ diagnostic(D, "positive")*. This rule can then be specialized as follows:
R2 : *coating(P), has_ location(S, L), contain(L, P), has_ structure(B, S), has_ year(B, Y), SWRL:lowerThanOrEqual(Y, 1980), has_ diagnostic_ characteristic(P, D)* $\rightarrow$ *has_ diagnostic(D, "positive")*

To discover the best year value, CRA-Miner explores the possible values from the most recent year to the oldest one and considers the rules that conclude on "*negative*" and "*positive*" differently.

In figure 2, we show how the confidence evolves between 1946 and 1997 for the rules that conclude on "*positive*" for a product class example. When the reference year decreases, hc decreases and the confidence increases. To cover the maximum number of diagnoses while maximizing the confidence, we stop the exploration when $hc < minHc$ (i.e. 1966 on the figure 2) and choose the previously explored year value such that $hc \geq minHc$ and the confidence stays

maximum (i.e. 1970 on the figure 2). A similar but symmetrical process is applied to choose y for the rules that conclude on "*negative*".
3- **Specialization with location and/or structure subclasses** : The hierarchy of locations and structures is explored to specialize the rules generated with step 1 and 2 with specific building components that contains the target product P.
For example, the rule x can be specialized by specificying that the location is a *wall* and that the structure is a *balcony*:
R3 : *coating(P), wall(L), balcony(S), has_location(S, L), contain(L, P), has_structure(B, S), has_year(B, Y), SWRL:lowerThanOrEqual(Y, 1980), has_diagnostic_characteristic(P, D)* $\rightarrow$ *has_diagnostic(D, "positive")*
4- **Enrichment by the region** : All the generated rules are enriched by the datatype property 'has_region' which represents the region where the building is located.
5- **Specialization by co-located components**. During this step, new object properties are added that represents **sibling specific products, sibling specific locations or sibling specific structures** : $contain(L, P_i)$ and $C_p(P_i)$ where i varies from 0 to $maxSiblingP$ and C_p is a leaf of the product hierarchy, then $has\_location(S, L_j), C_l(L_j)$ where j varies from 0 to $maxSiblingL$ and C_L is a leaf of the product hierarchy), and $has\_structure(S, L_j), C_l(L_j)$ where j varies from 0 to $maxSiblingS$ and C_S .

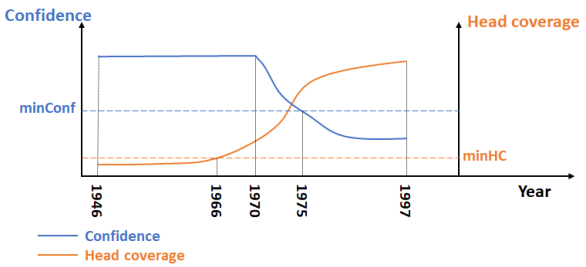

Fig. 2: Evolution of confidence and head coverage over time for the rules that conclude on "positive" for a class product example.

The number of generated and tested rules is mostly impacted by the temporal specialization (in the worst case the whole time interval will be tested) and the addition of co-located components (in the worst case all the possible combinations of co-located components will be checked). Despite this, CRA-Miner can be parallelized since each rule can be specialized independently from the others.

5 Experiments

We have evaluated our approach on a KG that has been populated using a set of diagnostic documents provided by the CSTB. This KG contains 51970

triples that describe 2998 product instances, 341 locations, 214 structures and 94 buildings. The construction year of those buildings varies between 1948 and 1997. We have 1525 products that contain asbestos and 1473 products are asbestos-free. All experiments were performed on a server with 80 physical processors (Intel Xeon E7-4830 2.20 GHz) and 528 GB of RAM.

The aim of the experimentation is (1) to learn rules on a subset of diagnostics and study the quality of the prediction that can be made on the remaining products (2) compare the results of our approach to a naive approach that only uses the product classes and the construction year (baseline) (3) compare the results of our approach with the two rule-mining approaches AMIE3 [6] and TILDE [1] (4) compare our results with an approach [8] that calculates asbestos probability using external resources (ANDEVA[9] and INRS[10]).

To evaluate our approach, we divided the KG data into 3 tiers, and we performed cross-validation. Since we have many different product classes, we set a low head-coverage threshold at $minHC = 0.001$ to observe as many rules as possible. Then we evaluated the results when $minConf$ varies from 0.6 to 0.9 using the classical precision, recall, F-Measure, and accuracy measures. The maximum number of siblings has been set at 0 for structures and at 3 for locations and products by the expert.

Table 1 shows that CRA-miner discovers 75 rules on average. The results show that co-located components are effectively exploited to predict the presence of asbestos: 29 rules involve at least a sibling product (maximum 2 sibling products) and 17 rules involve sibling locations (maximum 3 locations). Furthermore, the results shows that CRA-miner has discovered 14 rules that exploit a temporal constraint.

We have adhered to a pessimistic approach that chooses to classify a product as positive if at least one rule concludes that it contains asbestos.

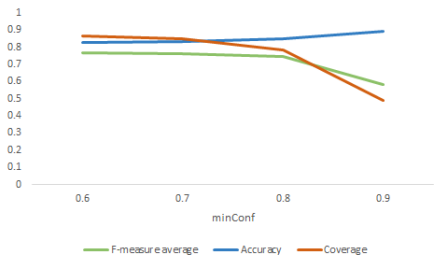

Fig. 3: CRA-Miner results according to $minConf$ threshold

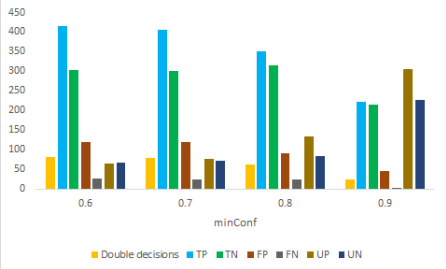

Fig. 4: Detailed CRA-Miner results according to $minConf$ thresholds

[9] National Association for the Defense of Asbestos Victims http://andeva.free.fr/expositions/gt_expos_produits.htm

[10] National Institute of Research and Security http://www.inrs.fr

Figure 3 presents the average F-measure, accuracy, and coverage (i.e. ratio of products that can be classified in the testing set) when $minConf$ varies. These results suggest that when the $minConf$ threshold increases, the accuracy increases but the data coverage decreases. The best F-measure average 0.77 (average between the positive's and negative's F-measure) is obtained for a $minConf$ fixed to 0.6. With such a threshold, we can decide for 87% of the test samples. Figure 4 details results (TN, TP, FN, FP, unclassified negative: UN, unclassified positive: UP) as well as the number of products that have been classified as positive and negative by different rules (double decisions). More precisely, the true positives TP (resp. true negatives TN) are the products that contain asbestos classified by the discovered rules as positive (resp. negative). The false positives FP (resp. false negative FN) are the asbestos-free products classified by the rules as positives, while the unclassified products are either positives (UP) or asbestos-free (UN) in the KG. This figure shows that a threshold fixed to 0.6 leads to an average of only 82 contradictory decisions for test samples that describe a thousand products.

We have compared the contextual CRA-Miner approach with a non-contextual baseline that is only based on the product class to mine rules. This baseline allows us to estimate the benefits of taking into account the hierarchy of products and the context in which they have been used (i.e., the location type and other products used in the same location). Figure 5 shows that the F-measure and the coverage are lower for the baseline approach regardless of the $minConf$ threshold that is varying from 0.6 to 0.9. In particular, Table 1 shows that the baseline only classifies 46% of the test samples and obtains a F-measure average of 0.55. Indeed, CRA-miner allows to discover complex rules such as:
plaster_or_cement_based_coating(?P), has_location(?S, ?L), contain(?L, ?P), smoothing_bubbling_leveling_plasters(?P2), contain(?L, ?P2), has_structure(?B, ?S), has_year(?B, ?Y), has_diagnostic_characteristic(?P, ?D) , lessThanOrEqual(?Y, "1991-01-01T00:00:00") → has_Diagnosis(?D, "positive")

We have compared our obtained results with AMIE3 [6] using the same thresholds of $minConf$ and $minHC$, and setting the number of predicates of the sought rules to $l = 4$ and $l = 6$ (cf. table 1)[11]. Our approach achieves a better F-measure than what is obtained with [6] (0.77 against 0.73 for $l = 6$, the specified l being the number of predicates allowing AMIE3 to obtain the best results in terms of F-Measure and accuracy). AMIE3 was able to discover 91 rules (75 with our approach) which allows it to cover 99% of the test data (87% with our approach). On the other hand, it obtains a lower accuracy (0.74 compared to 0.83 with CRA-Miner). This important coverage is accompanied by numerous double decisions (277). The approach is pessimistic (i.e., if a product is associated with two different decisions, it considered the product as one containing asbestos), AMIE3 finds more TP (473 against 415 with CRA-Miner) but almost twice as much FP (226 against 121 with CRA-Miner), and TNs are also less numerous (only 264 against 303 with CRA-Miner). Having a semantic

[11] Despite the fact that AMIE3 is used to search only for conclusive rules on $has\_Diagnosis$, a length > 6 does not yield results in less than three weeks.

context and being able to represent time intervals makes it possible to discover rules that involve more atoms while improving their readability for an expert in the field (more precisely, a rule can be defined for a time interval while AMIE3 can only generate rules involving a specific year).

Additionally, we tested our language bias using the TILDE system [1] that generates relational decision trees which allow representing complex language bias emulating similar target languages (relational context and maxSibling values) and handling (although not optimally) a hierarchy of types. The relational context used is slightly different, only imposing at least one instantiated type in the context (not necessarily the product). This top-down strategy obtains by definition a coverage of 100% without double decisions but leads to a lower precision for the positive and negative examples. Indeed, it gets more FP and FN since the last general rule classifies all the unclassified remaining individuals as positive or negative whatever its description is. CRA-Miner was not able to classify all examples (87%) but the obtained accuracy is higher (0.83 against 0.51 for TILDE). Given the strategy of TILDE, it was not possible to use the inequalities on years, because introducing this possibility yields the possibility to learn closed intervals on years and an overfitting that is difficult to control.

We have also compared CRA-Miner to the hybrid approach used in [8]. This approach uses two external resources that describe marketed products that contained asbestos during at least one period to compute a probability based on the product class and the construction year. Table 1 shows that [8] obtains a higher F-measure and accuracy in particular for positive products (0.94 against 0.79 for CRA-Miner). This can be explained by the additional information provided by the web resources that focus on positive products. However, CRA-Miner could cover more data samples (87% against 83% for the hybrid). Indeed, the hybrid approach could not decide on a product if its class was not mentioned in the external resources.

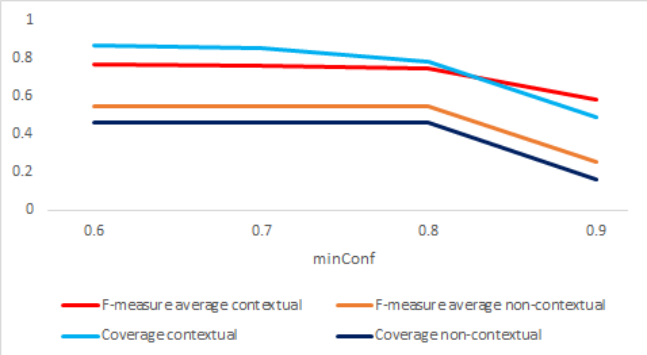

Fig. 5: Comparison between contextual and non-contextual approaches according to minConf thresholds

System classification	Rule mining systems					Systems based on external resources
System	CRA-Miner	AMIE3 $l = 4$	AMIE3 $l = 6$	TILDE	Baseline	Hybrid
# rules	75	45	91	34	24	/
Double decision	82	50	277	0	0	0
TP	415	381	473	424	146	465
TN	303	288	264	88	257	348
FP	121	146	226	359	30	16
FN	28	74	32	127	24	5
UP	66	54	3	0	338	38
UN	67	58	0	0	204	127
Pos. precision	**77%**	72%	68%	54%	83%	97%
Pos. recall	82%	75%	**93%**	77%	29%	92%
Pos. F-measure	**0,79**	0,73	**0,79**	0.63	0,43	0.94
Neg. precision	**92%**	80%	89%	41%	91%	99%
Neg. recall	**62%**	59%	54%	20%	52%	71%
Neg. F-measure	**0,74**	0,68	0,67	0.27	0,66	0.83
Avr. F-mesure	**0,77**	0,71	0,73	0.45	0,55	0.89
Accuracy	**0,83**	0,75	0,74	0.51	0,88	0.97
Coverage	87%	89%	**100%**	**100%**	46%	83%

Table 1: Comparison between CRA-Miner, AMIE3 with $l = 4$ et $l = 6$ (minHC=0.001, minConf=0.6), TILDE, the baseline approach, and the hybrid approach

These experiments have first shown that all the predicates of the context selected by the expert are relevant to classifying the product. Indeed, the baseline obtains a very low recall, and the results show that all the predicates have been used in at least one rule. The comparison with the other two available rule-mining systems illustrates that CRA-Miner obtains the best precision values, with a lower but still high value of coverage (87%). As expected, the experiments also show that the use of external resources about marketed products containing asbestos can lead to more precise decisions. However, this kind of resource is incomplete, and the obtained coverage is lower. Since it is more important to detect positive examples than negative ones, we have chosen to apply a pessimist strategy, and the results show that we obtain a better recall for positive than for negative examples. However, this choice affects the precision for the positives and other strategies could be considered (e.g., voting strategies, rules ordered according to their semantics and/or confidence). Another possibility is to use a higher confidence threshold for the negative. The results have shown that when the confidence value is fixed to 1, 43% of the negatives can still be discovered with only one false negative among 210 decisions (99.52% of precision).

6 Conclusion

In this paper, we presented the CRA-Miner rule discovery approach which can predict the presence of asbestos in products based on a semantic context, heuristics dedicated to part-of relations, and computed constraints on numerical values that represent temporal information. The experiments show that we can obtain a better precision and accuracy than two other rule-mining systems and better coverage than an approach based on external resources.
In the future, we plan to investigate the combination of CRA-Miner with the approach in [8] (that uses external resources) to enable decisions for the undefined individuals and improve the data coverage. Since results of our approach will be used by asbestos experts to select which products have the strongest priority to get tested, we also need to rank the positive, unclassified, and negative products according to the applied rules and define an interface that can present and explain this ranking to the experts. Besides, we plan to generalize CRA-Miner to fit different problems that involve part-of relations and temporal constraints (such as prediction of adverse events for treatments composed of several drugs in pharmacology). The idea is to follow the model of TILDE[1] where the language bias is declarative but the algorithm is generic (i.e., not ontology-based).

References

1. Blockeel, H., De Raedt, L.: Top-down induction of first-order logical decision trees. Artificial intelligence **101**(1-2), 285–297 (1998)
2. d'Amato, C., Tettamanzi, A.G.B., Tran, D.M.: Evolutionary discovery of multi-relational association rules from ontological knowledge bases. In: Blomqvist, E., Ciancarini, P., Poggi, F., Vitali, F. (eds.) EKAW 2016, Italy, November 19-23, 2016, Proceedings. Lecture Notes in Computer Science, vol. 10024, pp. 113–128 (2016)
3. Fanizzi, N., d'Amato, C., Esposito, F.: DL-FOIL concept learning in description logics. In: Inductive Logic Programming, ILP 2008, Czech Republic, September 10-12, 2008, Proceedings. Lecture Notes in Computer Science, vol. 5194, pp. 107–121 (2008)
4. Fürnkranz, J., Kliegr, T.: A brief overview of rule learning. In: Bassiliades, N., Gottlob, G., Sadri, F., Paschke, A., Roman, D. (eds.) Rule Technologies: Foundations, Tools, and Applications - 9th International Symposium, RuleML 2015, Berlin, Germany, August 2-5, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9202, pp. 54–69. Springer (2015)
5. Hamilton, W.L., Ying, R., Leskovec, J.: Representation learning on graphs: Methods and applications. IEEE Data Eng. Bull. **40**(3), 52–74 (2017), `http://sites.computer.org/debull/A17sept/p52.pdf`
6. Lajus, J., Galárraga, L., Suchanek, F.: Fast and exact rule mining with amie 3. In: Extended Semantic Web Conference (ESWC). Lecture Notes in Computer Science, vol. 12123, pp. 36–52. Springer (2020)
7. Lehmann, J., Auer, S., Bühmann, L., Tramp, S.: Class expression learning for ontology engineering. J. Web Semant. **9**(1), 71–81 (2011)

8. Mecharnia, T., Khelifa, L.C., Pernelle, N., Hamdi, F.: An approach toward a prediction of the presence of asbestos in buildings based on incomplete temporal descriptions of marketed products. In: Kejriwal, M., Szekely, P.A., Troncy, R. (eds.) K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019. pp. 239–242. ACM (2019)
9. Muggleton, S., Raedt, L.D.: Inductive logic programming: Theory and methods. J. Log. Program. **19/20**, 629–679 (1994). https://doi.org/10.1016/0743-1066(94)90035-3, https://doi.org/10.1016/0743-1066(94)90035-3
10. Ortona, S., Meduri, V.V., Papotti, P.: Robust discovery of positive and negative rules in knowledge bases. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE). pp. 1168–1179 (2018)
11. Paulheim, H., Tresp, V., Liu, Z.: Representation learning for the semantic web. J. Web Semant. **61-62**, 100570 (2020)
12. Quinlan, J.R.: Learning logical definitions from relations. Mach. Learn. **5**, 239–266 (1990)
13. Rizzo, G., Fanizzi, N., d'Amato, C.: Class expression induction as concept space exploration: From dl-foil to dl-focl. Future Gener. Comput. Syst. **108**, 256–272 (2020)
14. Schoenmackers, S., Davis, J., Etzioni, O., Weld, D.S.: Learning first-order horn clauses from web text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1088–1098. ACL (2010)