Découverte de règles contextuelles pour prédire la présence d'amiante dans les bâtiments

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

[1] LISN, Université Paris Saclay, Orsay, thamer@lri.fr
[2] Centre Scientifique et Technique du bâtiment (CSTB), Champs sur Marne, prenom.nom@cstb.fr
[3] CEDRIC, CNAM - Conservatoire National des Arts et Métiers, Paris, faycal.hamdi@cnam.fr
[4] LIPN, Université Sorbonne Paris-Nord, CNRS UMR 7030, Villetaneuse, prenom.nom@lipn.univ-paris13.fr

18 mai 2021

Résumé

Le Centre Scientifique et Technique du Bâtiment (CSTB) a été sollicité pour développer un outil d'aide à l'identification des matériaux contenant de l'amiante dans les bâtiments. Dans ce contexte, nous avons développé une approche, nommée CRA-Miner, qui utilise des techniques de programmation logique inductive (PLI) pour découvrir des règles à partir d'un graphe de données décrivant des bâtiments et des diagnostics d'amiante. La référence des produits spécifiques utilisés lors de la construction n'étant jamais spécifiée, CRA-Miner considère les données temporelles, les types de produits et les informations contextuelles pour rechercher l'ensemble de règles candidates qui pourront être utilisées pour prédire la présence d'amiante dans les éléments de construction. Les expériences menées sur le graphe de connaissances fourni par le CSTB montrent qu'une F-Mesure prometteuse peut être obtenue.

Mots-clés

Découverte de règles, graphe de connaissances, données temporelles, amiante.

Abstract

The Scientific and Technical Center for Building (CSTB) was asked to develop a tool to help identify materials containing asbestos in buildings. In this context, we have developed an approach, named CRA-Miner, which uses inductive logic programming (ILP) techniques to discover logic rules from a data graph describing buildings and asbestos diagnostics. Since the reference of the specific products used during construction is never specified, CRA-Miner considers the temporal data, the types of products and the contextual information to find the set of candidate rules which can then be used to deduce the presence of asbestos in construction elements. The experiments carried out on the knowledge graph provided by CSTB show that a promising F-Measure can be obtained.

Keywords

Rule mining, knowledge graph, temporal data, asbestos.

1 Introduction

La nocivité de l'amiante est identifiée depuis le début du 20ème siècle. L'inhalation d'air contenant des fibres d'amiante peut entraîner des maladies telles que le cancer des poumons et de la muqueuse thoracique. Cependant, en raison de ses qualités ignifuges, de nombreux pays ont largement utilisé l'amiante dans les bâtiments, en particulier de 1950 à 1970. Même si il est maintenant illégal d'utiliser cette dangereuse fibre minérale, celle-ci est toujours présente dans de nombreux bâtiments. Aussi, l'identification des parties de construction contenant de l'amiante est une tâche importante afin de procéder à un désamiantage. Les professionnels inspectent régulièrement les bâtiments et prélèvent des échantillons pour détecter la présence d'amiante dans les composants des bâtiments mais il est nécessaire de hiérarchiser le trop grand nombre de tests possibles.

Dans le cadre du PRDA [1], le CSTB (Centre Scientifique et Technique du Bâtiment) a été sollicité pour développer un outil en ligne d'aide à l'identification de matériaux contenant potentiellement de l'amiante dans les bâtiments afin de guider l'opérateur dans la préparation de son programme de suivi (Projet ORIGAMI). La difficulté réside dans le fait que les descriptions de bâtiments disponibles ne précisent que les classes de produits utilisés, sans donner leurs références exactes ou toute autre information à leur sujet (e.g. fournisseurs, etc.). Dans [6], une approche basée sur l'ontologie ASBESTOS a été définie, qui permet d'estimer la probabilité de présence de produits amiantés dans un bâtiment. Cette première approche hybride combine des méthodes statistiques et des méthodes basées sur des règles pour générer cette probabilité en se basant sur l'année de construction du bâtiment et des ressources externes fiables mais incomplètes décrivant les produits amiantés existant sur le marché à la même période. Cependant, les

1. Plan de recherche et de développement amiante lancé par la Direction de l'Habitat, de l'Urbanisme et des Paysages (DHUP), rattachée à la Direction Générale de l'Amiante, de l'Habitat et de la Nature (Ministre du Logement et de l'Habitat Durable)

experts supposent que si le type de produit et l'année de construction peuvent être exploités pour prédire la présence d'amiante, le contexte dans lequel le produit est utilisé peut également être pertinent (i.e. la région dans laquelle se situe le bâtiment, les éléments de construction dans lesquels apparaît le produit, ainsi que les autres produits utilisés). Récemment, le CSTB a mis à disposition un ensemble de diagnostics réalisés sur un grand nombre de bâtiments. Ces données ont été représentées à l'aide de l'ontologie ASBESTOS proposée dans [6]. L'objectif est de définir une approche qui vise à apprendre des règles contextuelles à partir de ces données sémantiques en utilisant des prédicats permettant de représenter l'appartenance d'une date de construction à un intervalle temporel. De nombreuses approches de fouille de règles ont été proposées qui peuvent apprendre à classer les données en fonction de leur description RDF [5, 4, 10]. Cependant, aucune d'entre elles n'utilise le type de prédicat numérique dont nous devons disposer pour prendre en compte cet aspect temporel.

Dans cet article, nous proposons une approche basée sur l'ontologie qui découvre des règles qui peuvent être utilisées pour estimer la probabilité de l'existence de produits amiantés dans un bâtiment. L'approche proposée s'inspire des techniques de Programmation Logique Inductive (PLI) de type *générer et tester*, mais se concentre sur la découverte de règles qui décrivent le produit et son contexte par un ensemble de prédicats déclarés comme potentiellement pertinents par l'expert. Sur la base des relations de subsomption et des connaissances générales sur l'évolution de l'utilisation de l'amiante au fil des années, l'algorithme découvre un ensemble de règles qui prédisent la présence d'amiante dans les produits d'un composant de bâtiment. L'originalité de l'approche CRA-Miner est de se baser sur un contexte sémantique, des heuristiques dédiées aux propriétés de type part-of omniprésentes dans les descriptions des bâtiments et des contraintes temporelles utilisant des seuils calculés.

Dans la section 2, nous présentons des travaux connexes. En section 3, nous décrivons l'ontologie Asbestos, puis en section 4, l'approche de classification. La section 5 présente les résultats obtenus sur des données réelles du CSTB en les comparant à une *baseline* et aux résultats obtenus avec AMIE3. Enfin, la section 6 tire un ensemble de conclusions et propose des orientations de recherches futures.

2 Travaux connexes

Dans le contexte des graphes de connaissances (KG), l'exploration de règles peut être utilisée pour enrichir les graphes (prédiction de liens ou de types, ajout de nouveaux axiomes, liaison d'entités), ou pour détecter les triples RDF erronés. Motivées par le besoin de passage à l'échelle, la plupart des approches récentes de prédiction de liens/types sont basées sur des méthodes d'apprentissage profond et de plongement de graphes qui permettent de traduire des vecteurs de grande dimension en espaces de dimension relativement faible [9]. Néanmoins, d'autres applications pour lesquelles des règles interprétables sont nécessaires pour comprendre et maintenir une certaine connaissance du domaine sont toujours intéressées par la découverte de règles logiques. C'est le cas de l'approche que nous proposons dans cet article.

De nombreuses approches se sont intéressées à l'apprentissage de règles et de concepts dans les graphes de connaissance. Les approches d'apprentissage de concepts telles que DL-Foil [3] ou DL-FOCL [10] permettent d'apprendre des définitions de concepts représentées en logique de description. Ces approches s'appuient sur des stratégies de type *separate-and-conquer* qui permettent de construire une disjonction de solutions partielles, pouvant être spécialisées à l'aide d'opérateurs de raffinement basés sur la subsomption, de façon à couvrir autant d'exemples positifs que possible tout en excluant (presque) tous les exemples négatifs. Cependant, ces approches, si elles permettent de générer des définitions de concepts dans des logiques de description expressives, ne recherchent pas toutes les solutions partielles, et donc toutes les définitions. De plus, elles ne permettent pas d'utiliser des prédicats instanciés par des constantes ou de rechercher des valeurs seuils (e.g. $X \leq 17$ pour définir un mineur). Les approches de classification FOLDT (first-order logical decision tree) telles que TILDE (Top-down induction of logical decision trees) ([1]) sont basées sur des arbres de décision dans lesquels les noeuds peuvent partager des variables et impliquer des prédicats numériques comportant des valeurs seuils. Cependant, ces dernières approches n'utilisent pas la sémantique de l'ontologie dans l'exploration de l'espace de recherche.

Les graphes de connaissances sont de plus en plus nombreux et volumineux mais les données ne sont pas nécessairement complètes. Les approches telles que AMIE3 [5] et RUDIK [8] s'intéressent à la découverte d'ensembles de règles exprimées en logique du premier ordre (clauses de Horn) dans des données RDF volumineuses. Pour disposer de contre-exemples, ces approches se basent sur l'hypothèse de complétude partielle (PCA) qui suppose que lorsqu'un objet est représenté pour une entité et une propriété spécifique, tous les objets sont représentés (i.e. les autres étant considérés comme contre-exemples). Pour mieux contrôler l'espace de recherche, AMIE3 se base les mesures de qualité et limite le nombre d'atomes qui apparaissent dans la règle. RUDIK permet de découvrir des règles qui utilisent la négation pour identifier des contradictions et qui comportent des prédicats permettant de comparer des valeurs numériques ou littérales. Cependant, ces constantes doivent être définies dans le graphe de connaissance et associées à deux variables de la règle, l'approche ne permettant pas de découvrir une constante de référence comme "$\hat{a}ge(X, a), a \geq 18 \rightarrow adulte(X)$", ce qui est l'un des objectifs dans notre application. D'autres approches telles que [2] peuvent être guidées par la sémantique de l'ontologie pour éviter de construire des règles sémantiquement redondantes. Cependant, l'auteur a montré que l'exploitation des capacités de raisonnement pendant le processus d'apprentissage ne permet pas d'exploiter les règles sur

les grands graphes. L'approche [6] est une première approche permettant d'exploiter l'ontologie et des données temporelles pour estimer la probabilité qu'un produit soit amianté mais ces connaissances temporelles sont issues de ressources externes incomplètes (ANDEVA et INRS).
Dans ce travail, nous cherchons à découvrir des règles de classification à partir d'exemples positifs et négatifs décrits dans le graphe de connaissances (GC) fourni par le CSTB, afin d'estimer la probabilité de présence d'amiante (faible ou élevée) d'un produit utilisé dans un bâtiment. Les produits commercialisés utilisés étant inconnus, nous nous focalisons sur des règles contextuelles qui suivent des modèles spécifiques définis par des experts du domaine. Comme l'année de construction a un impact important sur la possible présence d'amiante, des opérateurs de comparaison SWRL [2] sont exploités pour comparer une année de construction à une année de référence qui maximise la confiance de la règle pour un type de produit (ex. *SWRL :lowerThanOrEqual(YEAR, ref_year)*). Aucune des approches sémantiques mentionnées précédemment ne permet d'exploiter un tel prédicat numérique dans les règles générées.

3 Ontologie Asbestos

Dans cette section, nous présentons la partie haute de l'ontologie Asbestos (c.f. Figure 1) qui a été construite en exploitant les documents du CSTB, les connaissances experts, et les besoins de prédiction dans le projet ORIGAMI ([6]).

- Building : construction caractérisée par un code CSTB qui correspond à un type de bâtiment donné, le type de bâtiment (e.g. école, maison, etc.), l'année de construction et l'adresse du bâtiment.
- Structure : espace faisant partie du bâtiment (e.g. balcon, toit, escalier, etc.).
- Location : localisation composant une structure (e.g. porte, fenêtre, mur, etc.).
- Product : produit utilisé dans une localisation (e.g. colle, enduit, etc.).
- Diagnostic Characteristic : représente le résultat des diagnostics amiante. La valeur de $has\_diagnostic$ est "positive" lorsque le produit contient de l'amiante ou "negative" sinon.

L'ontologie Asbestos décrit 8 sous-classes de structures, 19 sous-classes de localisation et 38 sous-classes de produit.

4 L'approche CRA-Miner

Dans cette section, nous décrivons tout d'abord les règles logiques contextuelles que nous voulons fournir aux experts pour les aider à détecter les matériaux contenant de l'amiante dans le bâtiment. Nous présentons ensuite l'algorithme CRA-Miner qui permet de générer ces règles à partir de l'ontologie ASBESTOS peuplée.

4.1 Règles contextuelles pour la prédiction de l'amiante

Une règle contextuelle pour la prédiction de l'amiante (CRA) est une conjonction de prédicats qui conclut sur la présence ou l'absence d'amiante dans un produit P. Nous considérons la borne supérieure hors-contexte de l'espace de recherche $\top$ suivante : $product(P),\ has\_diagnostic\_characteristic(P, D) \rightarrow has\_diagnostic(D, Value)$. L'ensemble des règles contextuelles qui peuvent être construites à partir de cette borne supérieure est défini en utilisant un contexte conceptuel. Ce contexte est utilisé par les experts pour sélectionner les éléments de l'ontologie décrivant le produit P pouvant avoir un impact sur la présence d'amiante. Ces prédicats utilisés pour spécialiser la règle représentent un biais de langage tel que défini en Programmation Logique Inductive (PLI) [7].

Définition 1 (Contexte conceptuel) Un contexte conceptuel CO est défini par un sous-graphe de l'ontologie, c'est-à-dire un ensemble de classes et de propriétés, qui déterminera les prédicats utilisables dans le corps de la règle.

Exemple 1 $CO = \{product, location, structure, contain, has\_location, has\_region, has\_year, has\_structure, has\_diagnostic\_characteristic\}$ est un exemple de contexte conceptuel.

Une règle contextuelle est basée sur le vocabulaire de l'ontologie sélectionné dans le contexte conceptuel et les spécialisations du prédicat $SWRL : CompareTo$ qui peut être ajouté pour introduire des contraintes sur l'année de construction du bâtiment (i.e. intervalles ouverts) :

Définition 2 (Règle contextuelle) Soit CO un contexte conceptuel, une règle contextuelle $\overrightarrow{B} \rightarrow h$, où $\overrightarrow{B} = \{B_1, B_2, ..., B_n\}$, est telle que $\forall B_i \in \overrightarrow{B}, \exists B_j \in CO \cup \{SWRL : CompareTo\}$ s.t. $B_i \sqsubseteq B_j$ et h est le prédicat $has\_diagnostic$ qui est instancié par la valeur "positive" ou "negative".

Une règle contextuelle doit également respecter les propriétés de fermeture et de connectivité définies dans les approches de fouille de règles telles que [5].

Exemple 2 La règle suivante est une règle contextuelle connexe et fermée qui peut être formée avec le contexte CO défini dans l'exemple 1 :

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")

Cette règle exprime qu'une colle présente dans un bâtiment parisien construit avant 1950, qui est utilisée dans la même localisation qu'une peinture, est potentiellement amiantée.

Des contraintes supplémentaires sont définies pour réduire la complexité du contexte et limiter la taille de l'espace

2. https://www.w3.org/Submission/SWRL/

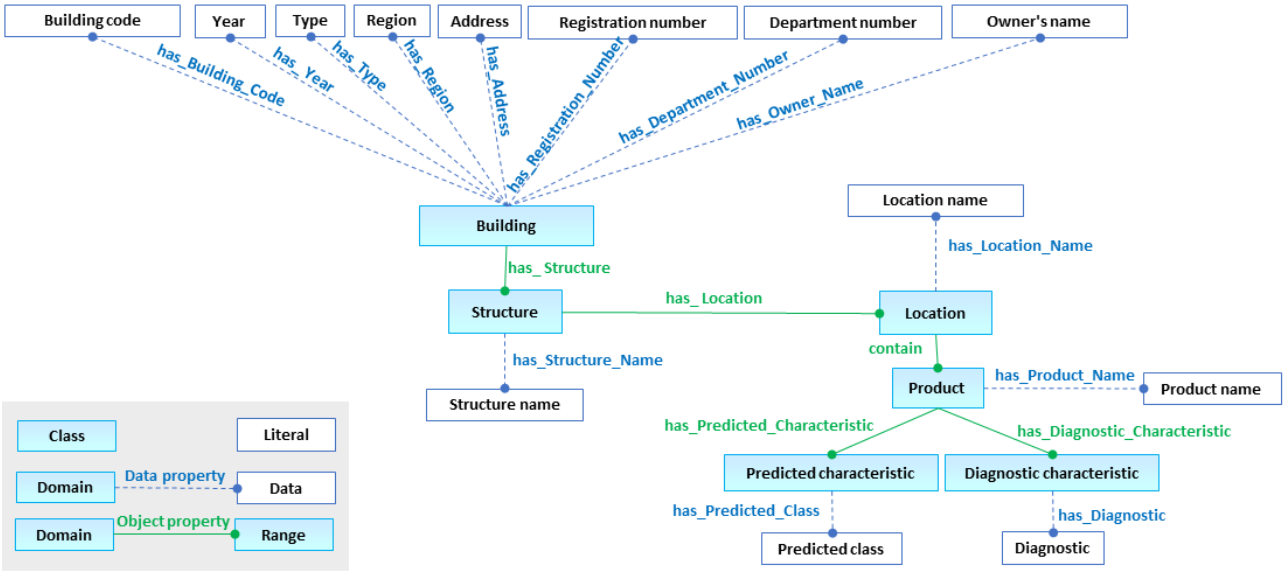

FIGURE *1 – Concepts principaux de l'ontologie Asbestos*

de recherche pour les propriétés multi-valuées décrivant les parties de bâtiments (i.e. $contain$, $has\_location$, et $has\_structure$).
L'expert peut tout d'abord définir le nombre maximum d'occurences des autres composants du bâtiment qui peuvent apparaître dans le corps de la règle : $maxSibS$ est utilisé pour définir le nombre de structures frères de la structure qui contient le produit P, $maxSibL$ est le nombre maximum de localisation frères, et $maxSibP$ représente le nombre maximum de produits frères.

Example 3 Si l'expert considère que le type des autres structures présentes dans le bâtiment ne peut influencer la présence d'amiante dans P, alors $maxSibS = 0$ et l'approche ne pourra construire la règle suivante :

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")

En effet, la structure $S2$ ne devrait pas être considérée (frère de $S1$ par la propriété $has\_structure$, $S1$ contenant le produit cible).

Enfin, les experts du CSTB considèrent que seule la prise en compte des types de produits les plus spécifiques peut impacter le choix du produit cible commercialisé utilisé et donc la présence d'amiante. Par exemple, la présence d'un revêtement (i.e. $coating$) dans la même localisation qu'un produit cible de type colle n'est pas significative tandis que la présence d'un revêtement de sol peut impacter le choix de la colle commercialisée utilisée. Une hypothèse similaire est réalisée pour les localisations et les structures. Aussi, seules les classes les plus spécifiques sont ajoutées dans les relations de type part-of considérées.

Pour mesurer la qualité des règles, nous utilisons les mesures de qualité classiques de *head coverage* (hc) [5] et de confiance ($conf$) qui ont été définies pour les règles relationnelles.
Le *head coverage* (hc) représente la ratio entre le support, i.e. le nombre de prédictions correctes de $has\_diagnostic(D, \text{"positive"})$ (resp. $has\_diagnostic(D, \text{"negative"})$) généré par la règle, et le nombre de diagnostics $has\_diagnostic(D, \text{"positive"})$ (resp. $has\_diagnostic(D, \text{"negative"})$) qui sont présents dans le graphe de connaissance :

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

La confiance ($conf$) est définie par le ratio entre le support de la règle et le nombre de diagnostics différents qui participent à une instanciation du corps de la règle.

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

L'objectif est de découvrir toutes les règles les plus générales qui sont conformes aux contraintes du biais de langage et qui sont telles que $hc \geq minHc$ et $conf \geq minConf$.

4.2 Evolution de la présence d'amiante au fil du du temps

Il a été montré dans [6] que le nombre de produits commercialisés amiantés sur le marché reste stable jusqu'en 1972, puis diminue pour atteindre 0 en 1997, lorsque l'usage de

l'amiante est interdit en France. En effet, soit les produits ont été désamiantés, soit ils ont été abandonnés. Ainsi, même si la probabilité d'amiante diffère d'une classe de produit à une autre (e.g. les adhésifs ont perdu leur amiante plus tôt et plus rapidement que d'autres classes de produits), nous savons que cette probabilité diminue avec le temps. Aussi, si une règle contextuelle conclut sur l'absence d'amiante pour les produits utilisés dans les bâtiments construits après une année donnée Y_1, la confiance ne pourra qu'être égale ou augmenter pour $Y_2 \geq Y_1$. Cette caractéristique est exploitée pour élaguer l'espace de recherche lorsque le prédicat $greaterThanOrEqual$ ou $lessThanOrEqual$ est généralisé.

4.3 Algorithme CRA-Miner

Le but de l'algorithme CRA-Miner est de générer toutes les règles contextuelles permettant de prédire la présence d'amiante dans les produits à partir des exemples positifs et négatifs décrits dans le graphe de connaissances (GC) et telles que $hc \geq minHC$ et $conf \geq minConf$. L'algorithme de type *descendant générer et tester*, spécialise la borne supérieure de l'espace de recherche T en considérant la hiérarchie des classes de produit, en ajoutant des contraintes sur la localisation et la structure de ce produit, sur la présence de produits, localisations ou structures apparaissant dans le même composant, ainsi que des contraintes temporelles sur l'année de construction.

L'algorithme a comme entrées le graphe de connaissances, le biais de langage, un seuil $minConf$ sur la confiance, un seuil $minHC$ sur le *head coverage* de la règle, ainsi que les valeurs de $maxSibP$, $maxSibL$ et $maxSibS$ qui limitent le nombre de frères de produits, de localisations et de structures à ajouter à la règle.Le résultat est un ensemble $\mathcal{CR}$ de règles contextuelles.
L'exploration de l'espace de recherche est guidée par les relations de subsomption de l'ontologie (exploration top-down des produits cibles, de leurs localisations et de leurs structures) et exploite le fait que le nombre de produits amiantés est décroissant au fur et à mesure des années. A chaque étape de spécialisation, les règles construites qui possèdent dans une valeur de confiance et une valeur de *head coverage* plus grande que les seuils sont stockées dans l'ensemble $\mathcal{CR}$. Pour toutes les règles telles que $conf = 1$ ou $hc < minHc$, la spécialisation s'arrête.

Nous décrivons les étapes de l'algorithme pour le contexte le plus général qui a été défini par les experts du CSTB, i.e. le contexte CO défini dans l'exemple 1. L'algorithme comporte les 5 étapes suivantes :

1- **Spécialisation de $\top$ en utilisant des sous-classes de produit** :
Dans cette phase, nous remplaçons dans le $\top$ la classe $product$ par toutes les classes plus spécifiques (e.g. enduit, peinture, etc.) tant que $hc \geq minHC$ et générons donc toutes les règles "*hors-contexte*" qui peuvent être trouvées pour chaque classe de produit sans tenir compte des autres composants ou de la date de construction.

2- **Spécialisation par ajout d'une contrainte temporelle**. Pour chaque règle hors-contexte générée par l'étape précédente, nous ajoutons le chemin de propriété nécessaire pour atteindre l'année de construction. à partir du produit cible P : $has\_location(S, L), contain(L, P), has\_structure(B, S)$, $has\_year(B, Y)$. Le prédicat *SWRL :lowerThanOrEqual(Y, y)* (pour une règle qui conclut sur "positive") ou *SWRL :greaterThanOrEqual(Y, y)* (pour "negative"), est également ajouté pour comparer l'année de construction Y à une année de référence y qui maximise la confiance et préserve $hc \geq minHc$.
Par exemple, si la règle R1 suivante est générée par la première étape :
R1 : *coating(P), has_diagnostic_characteristic(P, D)* $\rightarrow$ *has_diagnostic(D, "positive")*
Cette règle peut être spécialisée de la façon suivante :
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")*

Pour découvrir la meilleure année de référence, CRA-Miner explore les valeurs d'année possibles de la plus récente à la plus ancienne, et considère différemment les règles qui concluent sur "$negative$" et "$positive$".
La Figure 2 montre comment la confiance évolue de 1946 à 1997 pour une règle qui conclut sur "$positive$" et pour une une classe de produit. Quand l'année de référence diminue, le *head coverage* hc diminue et la confiance $conf$ augmente. Pour couvrir le nombre maximum de diagnostics en maximisant la confiance, l'exploration s'arrête quand $hc < minHc$ (i.e. 1966 sur la figure 2). La dernière année explorée telle que $hc \geq minHc$ et telle que la confiance reste maximum (i.e. 1970 sur la figure 2) est choisie. Un processus similaire mais symétrique est appliqué pour choisir y pour les règles concluant sur "$negative$".

3- **Spécialisation par localisation et/ou par structure** (prédicat 'Location' et prédicat 'Structure').
Les hiérarchies de localisations et structures sont explorées pour spécialiser les règles générées en étape 1 et 2 avec des composants de bâtiment spécifiques qui contiennent le produit cible P.
Par exemple, la règle R1 peut être spécialisée en spécifiant que la localisation est un mur et que la structure est un balcon.
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- **Enrichissement par la région** Toutes les règles générées peuvent être enrichies par la propriété 'has_region' qui représente la région dans lequel le bâtiment est situé.

5- **Spécialisation en ajoutant d'autres composants**. Dans cette étape, de nouvelles propriétés sont ajoutées qui représentent des produits spécifiques frères, des localisations spécifiques frères ou des structures spécifiques frères : $contain(L, P_i)$ et $C_p(P_i)$ où i varie de 0 à $maxSiblingP$ et C_p est une feuille de la hiérarchie de produits, puis $has\_location(S, L_j), C_l(L_j)$ où j varie de 0 à $maxSiblingL$ et C_L est une feuille de la hiérarchie des localisations), et $has\_structure(S, L_j), C_l(L_j)$ où j varie de 0 à $maxSiblingS$ et C_S et une feuille de la hiérarchie des structures.

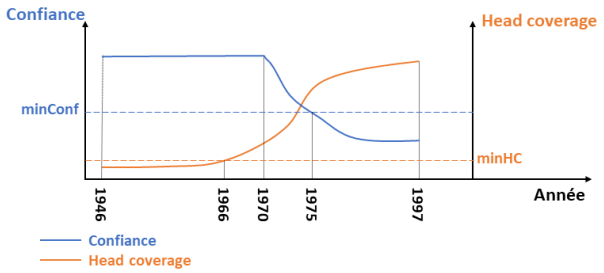

FIGURE 2 – *Évolution de la confiance et du head coverage d'une règle concluant sur "positive"*

5 Expérimentations

Nous avons évalué notre approche sur un GC peuplé à partir d'un ensemble de diagnostics fournis par le CSTB. Il comporte 51970 triplets qui décrivent 2998 instances de produit, 341 localisations, 214 structures et 94 bâtiments. L'année de construction de ces bâtiments varie entre 1948 et 1997. Nous avons 1525 produits contenant de l'amiante et 1473 produits sont sans amiante (les données sont disponibles dans le GitHub [3]).
Le but de l'expérimentation est (1) d'apprendre des règles sur un sous-ensemble de diagnostics et d'étudier la qualité de la prédiction qui peut être faite sur les produits restants (2) de comparer ces résultats à une approche naïve qui n'utilise que les classes de produits (3) de comparer les résultats de notre approche avec AMIE3 [5]. Pour évaluer notre approche, nous avons divisé les données du GC en 3 tiers, et nous avons réalisé une validation croisée. Comme nous disposons de nombreuses classes de produits de tailles différentes, nous avons fixé un seuil de *head-coverage* à $minHC = 0,001$ pour observer le plus de règles possible, puis nous avons évalué les résultats lorsque $minConf$ varie de $0,6$ à 1 en utilisant les mesures classiques de précision, rappel, F-Mesure et exactitude (accuracy). Le nombre maximum de frères a été fixé à 0 pour les structures et à 3 pour les localisations et les produits.

La table 1 montre que CRA-miner découvre 75 règles en moyenne. Les résultats montrent que des composants frères sont effectivement exploités pour prédire la présence d'amiante : 29 règles comportent au moins un produit frère (au maximum 2) et 16 règles comportent une localisation frère. Parmi les 75 règles, 14 d'entre elles exploitent une contrainte temporelle.

Nous avons utilisé une approche pessimiste qui consiste à choisir de classer un produit comme positif si au moins une règle découverte conclut à la présence d'amiante.

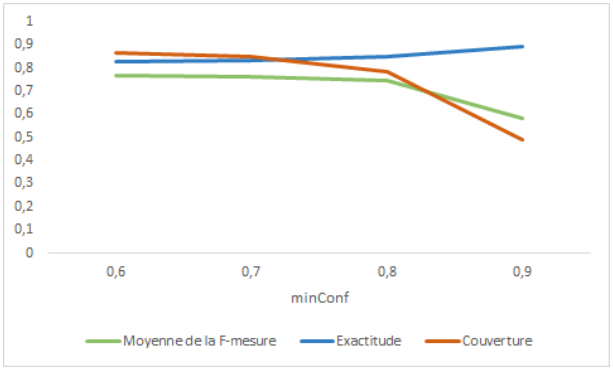

FIGURE 3 – *Résultats de CRA-Miner selon différents seuils de minConf*

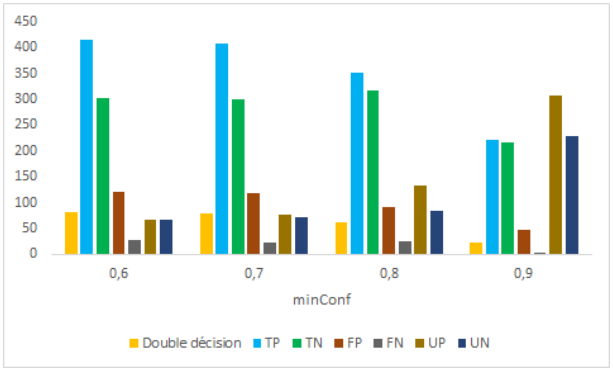

FIGURE 4 – *Résultats détaillés de CRA-Miner selon différents seuils de minConf*

La Figure 3 présente la moyenne de la F-mesure, l'exactitude et la couverture des données (i.e. ratio des produits qui peuvent être classés dans l'ensemble des produits de test), lorsque le seuil $minConf$ varie. Quand le seuil $minConf$ augmente, la couverture des données diminue mais l'exactitude augmente. La meilleure moyenne de F-mesure, $0,77$, (moyenne entre la F-mesure positive et négative) est obtenue pour un $minConf$ fixé à $0,6$. Avec ce seuil, nous pouvons décider pour 87% de l'échantillon de test. La Figure 4 présente les résultats détaillés (TN : vrais négatifs, TP : vrais positifs, FN : faux négatifs, FP : faux positifs, UN : négatifs non classifiés, UP : positifs non classifiés) et le nombre de produits qui ont été classés à la fois comme positifs et négatifs par différentes règles (double décision). Ce chiffre montre que le seuil de $0,6$ entraîne seulement 82 décisions contradictoires parmi les mille produits.
Plus précisément, les vrais positifs TP (resp. Vrais négatifs TN) sont les produits contenant de l'amiante (resp. ne

3. https://github.com/ThamerMECHARNIA/DATA-IC2021

Statistiques	CRA-Miner	$l = 4$	$l = 6$	Baseline
# règles	75	45	91	24
Double décision	82	50	277	0
TP	415	381	473	146
TN	303	288	264	257
FP	121	146	226	30
FN	28	74	32	24
UP	66	54	3	338
UN	67	58	0	204
Pos. précision	**77%**	72%	68%	83%
Pos. rappel	82%	75%	**93%**	29%
Pos. F-mesure	**0,79**	0,73	**0,79**	0,43
Neg. précision	**92%**	80%	89%	91%
Neg. rappel	**62%**	59%	54%	52%
Neg. F-mesure	**0,74**	0,68	0,67	0,66
Moy. F-mesure	**0,77**	0,71	0,73	0,55
Exactitude	**0,83**	0,75	0,74	0,88
Couverture	87%	89%	**100%**	46%

TABLE *1 – Comparaison de CRA-Miner avec une approche non contextuelle (baseline) et AMIE3 avec* $l = 4$ *et* $l = 6$ *(minHC=0.001, minConf=0.6)*

contiennent pas de l'amiante) classés par les règles découvertes comme positifs (resp. négatifs). Les faux positifs FP (resp. Faux négatifs FN) sont les produits sans amiante (resp. avec amiante) classés par les règles comme positifs (resp. négatifs), tandis que les produits non classés sont soit positifs (UP), soit sans amiante (UN) dans le KG.
Nous avons comparé l'approche contextuelle CRA-Miner avec une baseline exploitant uniquement la classe de produit. Cela nous permet d'estimer l'intérêt de considérer la hiérarchie des produits et le contexte dans lequel ils ont été utilisés. La Figure 5 montre que la F-mesure et la couverture sont beaucoup plus faibles quel que soit le seuil $minConf$. Ainsi, pour $minConf = 0.6$, le tableau 1 montre que la baseline ne permet de classer que 46% des échantillons de test et obtient une moyenne de F-mesure de $0,55$. En effet, CRA-miner permet de découvrir des règles contextuelles complexes telles que :

"*plaster-based_plaster_or_smooth_sprayed_cement_under_floats(?P), has_location(?S, ?L), contain(?L, ?P), underlays_of_wall_fabrics(?P2), contain(?L, ?P2), has_structure(?B, ?S), has_year(?B, ?Y), has_diagnostic_characteristc(?P, ?D), lessThanOrEqual(?Y, "1997-01-01T00 :00 :00") → has_diagnostic(?D, "positive")*"

ou

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")

Nous avons également comparé ces résultats avec ceux pouvant être obtenus avec AMIE3 [5] en utilisant les mêmes seuils de $minConf$ et $minHC$, et en fixant le nombre de prédicats des règles recherchées à $l = 4$ et $l = 6$ (cf. table 1)[4]. Notre approche permet d'atteindre une meilleure F-mesure que celle obtenue avec [5] (0,77 contre 0,73 pour $l = 6$, $l = 6$ étant le nombre de prédicats permettant à AMIE3 d'obtenir les meilleurs résultats en termes de F-Mesure et d'exactitude). AMIE3 a pu découvrir 91 règles (75 avec notre approche) ce qui lui permet de couvrir 100% des données test (87% avec notre approche). En revanche, AMIE3 obtient une exactitude plus faible (0,74 contre 0,83 avec CRA-Miner). Cette importante couverture est accompagnée de nombreuses doubles décisions (277). L'approche suivie étant pessimiste (i.e. si un produit est associé à deux décisions différentes, on le considère comme amianté), AMIE3 trouve plus de TP (473 contre 415 avec CRA-Miner) mais également deux fois plus de FP (226 contre 121 avec CRA-Miner), et les TN sont aussi moins nombreux (seulement 264 contre 303 avec CRA-Miner). De manière générale, le fait de disposer d'un contexte sémantique et de pouvoir représenter des intervalles de temps permet de découvrir des règles comportant plus de prédicats tout en améliorant la lisibilité des contraintes temporelles pour un expert du domaine (i.e. une règle sera définie pour un intervalle de temps, ce qui n'est pas possible avec AMIE3 où une année ne pourra apparaître que sous forme d'une constante).

Ces expérimentations ont tout d'abord montré que tous les prédicats du contexte qui ont été sélectionnés par l'expert sont pertinents pour classer le produit. En effet, la baseline

4. Même si AMIE3 est utilisée pour chercher uniquement des règles concluant sur $has\_Diagnosis$, une longueur > 6 ne permet pas d'obtenir de résultats en moins d'une semaine

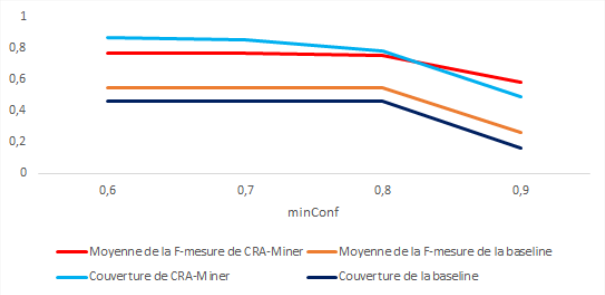

FIGURE 5 – *Comparaison entre CRA-Miner et la baseline non contextuelle selon différents seuils de minConf*

obtient un rappel très faible, et les résultats montrent que tous les prédicats ont été utilisés dans au moins une règle. La comparaison avec un autre système d'exploration de règles [5] montre que CRA-Miner obtient les meilleures valeurs de précision, F-mesure et exactitude, avec une valeur de couverture plus faible mais élevée (87%). Puisqu'il est plus important de détecter les exemples positifs que négatifs, nous avons choisi d'appliquer une stratégie pessimiste, et les résultats montrent que nous obtenons un meilleur rappel pour les exemples positifs que pour les exemples négatifs. Cependant, ce choix affecte la précision des positifs et d'autres stratégies pourraient être envisagées (ex : stratégies de vote, règles ordonnées en fonction de leur sémantique et/ou de leur confiance).

6 Conclusion

Dans ce papier, nous avons présenté l'approche de découverte de règles CRA-Miner qui prédit la présence d'amiante dans les produits en se basant sur un contexte sémantique, des heuristiques dédiées aux propriétés de type part-of et des contraintes temporelles utilisant des seuils calculés par le système. Les expérimentations montrent qu'une bonne F-mesure et une bonne couverture peuvent être obtenus et que ces résultats sont meilleurs que ceux pouvant être obtenus par un autre système de type générer et tester tel que AMIE3.

Dans des travaux futurs, nous envisageons d'explorer la possibilité de combiner cette approche avec une approche hybride telle que définie dans [6] afin d'en améliorer la couverture. Comme les résultats de cette approche doivent être utilisés par les experts amiante du CSTB pour prioriser les produits à diagnostiquer, nous devons également ordonner les produits positifs, non classifiés et négatifs en fonction des règles appliquées, et définir une interface qui permet de présenter et expliquer cet ordre aux experts.

Références

[1] Hendrik Blockeel and Luc De Raedt. Top-down induction of first-order logical decision trees. *Artificial intelligence*, 101(1-2) :285–297, 1998.

[2] Claudia d'Amato, Andrea G. B. Tettamanzi, and Duc Minh Tran. Evolutionary discovery of multi-relational association rules from ontological knowledge bases. In Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, and Fabio Vitali, editors, *EKAW 2016, Italy, November 19-23, 2016, Proceedings*, volume 10024 of *Lecture Notes in Computer Science*, pages 113–128, 2016.

[3] Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito. DL-FOIL concept learning in description logics. In *Inductive Logic Programming, ILP 2008, Czech Republic, September 10-12, 2008, Proceedings*, volume 5194 of *Lecture Notes in Computer Science*, pages 107–121, 2008.

[4] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs : Methods and applications. *IEEE Data Eng. Bull.*, 40(3) :52–74, 2017.

[5] Jonathan Lajus, Luis Galárraga, and Fabian Suchanek. Fast and exact rule mining with amie 3. In *Extended Semantic Web Conference (ESWC)*, volume 12123 of *Lecture Notes in Computer Science*, pages 36–52. Springer, 2020.

[6] Thamer Mecharnia, Lydia Chibout Khelifa, Nathalie Pernelle, and Fayçal Hamdi. An approach toward a prediction of the presence of asbestos in buildings based on incomplete temporal descriptions of marketed products. In Mayank Kejriwal, Pedro A. Szekely, and Raphaël Troncy, editors, *K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*, pages 239–242. ACM, 2019.

[7] Stephen Muggleton and Luc De Raedt. Inductive logic programming : Theory and methods. *J. Log. Program.*, 19/20 :629–679, 1994.

[8] S. Ortona, V. V. Meduri, and P. Papotti. Robust discovery of positive and negative rules in knowledge bases. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1168–1179, 2018.

[9] Heiko Paulheim, Volker Tresp, and Zhiyuan Liu. Representation learning for the semantic web. *J. Web Semant.*, 61-62 :100570, 2020.

[10] Giuseppe Rizzo, Nicola Fanizzi, and Claudia d'Amato. Class expression induction as concept space exploration : From dl-foil to dl-focl. *Future Gener. Comput. Syst.*, 108 :256–272, 2020.