Intégration des méta-données dans la fouille de données

Omar Boussaid – Sabine Loudcher-Rabaseda

Laboratoire ERIC, Université Lyon 2 Campus Porte des Alpes,
69676 Bron Cedex

(Omar.Boussaid, Sabine.Loudcher)@univ-lyon2.fr,
http://eric.univ-lyon2.fr/

RÉSUMÉ. *Dans le cadre de l'extraction de connaissance à partir des données (ECD), les méta-données sont déjà utilisées pour la préparation des données ou pour la validation et/ou visualisation des résultats. En revanche, elles ne sont pas utilisées au cours de la tâche de fouille qui est pourtant le cœur du processus d'ECD. Dans ce papier nous proposons une approche qui permet l'utilisation des méta-données lors de la phase de fouille afin d'enrichir et d'améliorer ses résultats. Nous représentons les méta-données à l'aide du formalisme RDF puis des graphes conceptuels. Elles sont ensuite intégrées aux données sur lesquelles la fouille a lieu. L'approche est testée sur des données et des méta-données du domaine de l'accidentologie avec comme méthode de fouille les arbres de décision. Les résultats montrent une nette amélioration du taux de mauvais classement.*

ABSTRACT. *The usage of metadata is increasing more and more especially in knowledge discovery in databases (KDD) and data mining. In this context, metadata are used either for preparing data, or validating and/or visualizing results. The metadata are not employed in data mining phase which is considered as the principal engine for KDD process. In fact, metadata are used either before or after, but never within the data mining itself. In this article we propose an approach which integrates the use of metadata in data mining. This approach is built on the results of several research areas. Indeed, we utilize RDF/S, introduced within the Web Semantic research field, as a metadata representation tools. We also represent RDF/S metadata through conceptual graphs in order to benefit of their powerful reasoning mechanisms. Throughout this paper, we apply our approach on accidentology data and we show that our approach is an efficient tool to improve data mining results.*

MOTS-CLÉS : *méta-données, RDF/S, graphes conceptuels, fouille de données*

KEYWORDS: *metadata, RDF/S, conceptual graphs, data mining*

1. Introduction

L'extraction des connaissances à partir des données, ECD (knowledge discovery in databases, KDD) (Fayyad *et al.,* 1996),(Fayyad *et al.,* 2001), est un processus constitué de plusieurs étapes, qui couvre la préparation des données, l'application de méthodes de fouille (*data mining*) et enfin la validation (ou la visualisation) des résultats. L'automatisation d'une de ces étapes se révèle une tâche difficile, car l'expertise requise dans le domaine est trop importante. Par conséquent, le besoin d'interactivité est nécessaire. D'autre part, le recours à la connaissance du domaine sous la forme de méta-données peut jouer un rôle considérable dans ce processus. Nous nous intéressons dans cet article aux méta-données et à leur utilisation dans le cadre de l'ECD. Dans plusieurs domaines de recherche, l'intérêt porté aux méta-données est réel. En effet, leur utilisation constitue un moyen efficace dans les traitements complexes nécessitant la prise en compte de la sémantique des données. L'automatisation des traitements qui demandait autrefois une expertise dans le domaine, peut être envisagée. Il suffit pour cela de collecter et de modéliser les informations utiles sur les données. Dans ce contexte et afin d'améliorer le processus d'ECD, de nombreux travaux se sont intéressés aux méta-données. En effet, il est plus que nécessaire de prendre en compte la sémantique des données. L'utilisation des méta-données dans le cadre de l'ECD est réelle quant il s'agit de pré-traitement des données et/ou de validation des connaissances extraites. Les méta-données sont ainsi utilisées en amont de la fouille ou en aval, mais jamais dans la fouille elle-même. Ce constat a motivé notre intérêt pour les méta-données, et nous a amené à réfléchir à une démarche permettant de les utiliser dans la phase de la fouille elle-même. Pour aborder ce problème, nous nous sommes fixés deux objectifs : (1) comment représenter les méta-données afin de les intégrer dans la fouille ? (2) quels traitements peut-on effectuer sur les méta-données pour améliorer les résultats de la fouille de données ?

Pour représenter les méta-données, nous utilisons le formalisme RDF/S[1] (*Resource Description Framework* et les schémas RDF) qui est un standard dans le domaine. RDF utilise des triplets pour décrire les méta-données comme des ressources ayant des propriétés. Il permet en outre de structurer les ressources et les propriétés à l'aide de schémas (RDF/S). Cependant, malgré la souplesse de ce langage, il n'est pas possible d'effectuer un raisonnement sur ces représentations. Pour pallier cette limite, nous avons recours aux graphes conceptuels et à leur capacité d'inférence. Ainsi, nous proposons une approche basée sur la représentation des méta-données en RDF/S puis par les graphes conceptuels pour le raisonnement. A partir de cette représentation , les méta-données pertinentes pour la fouille sont extraites et intégrées dans les données à fouiller. La méthode de fouille de données utilisée dans cet article est celle des arbres de décision avec l'algorithme C4.5. Les expérimentations faites sur le problème de l'accidentologie montrent que les résultats de fouille sont meilleurs avec la prise en compte des méta-données.

1. http ://www.w3.org/RDF/

L'organisation de ce papier est comme suit. Dans la section 2, nous présentons un état de l'art sur l'utilisation des méta-données dans l'ECD. La section 3 expose les formalismes utilisés pour représenter les méta-données. Nous développons notre approche dans la section 4. Enfin, nous illustrons notre démarche à travers une étude de cas dans la section 5. Nous terminons ce papier par une conclusion et des perspectives.

2. Etat de l'art

La fouille de données est définie comme l'art d'extraire des connaissances à partir des données. Les techniques de fouille de données (les algorithmes de segmentation, les règles d'association, les arbres de décision, les réseaux de neurones...), sont proposées selon le problème à résoudre. Il peut s'agir de méthodes de structuration ou de classification (Berkhin, 2002), d'explication ou de prédiction (Zighed *et al.*, 2002). Parmi ces techniques, nous nous intéressons aux arbres de décision dans le cadre de cet article. Dans la fouille de données, et plus spécialement dans les arbres de décision (Breiman *et al.*, 1984, Quinlan, 1993), l'extraction des connaissances se fait sous forme de règles par apprentissage à partir d'un échantillon d'individus (dit échantillon d'apprentissage) de l'ensemble des données de départ. Une autre partie de l'ensemble des individus (échantillon de test) est utilisée pour tester si les règles extraites par l'arbre de décision représentent un bon classement ou non. Plus le pourcentage (coût) de mauvais classement diminue, plus l'arbre est dit meilleur. La fouille de données est l'un des maillons de la chaîne de traitements pour l'extraction de connaissances à partir des données. Elle est le moteur principal dans ce processus. Elle met en oeuvre un ensemble de techniques provenant des bases de données, de la statistique, de l'apprentissage, de l'analyse des données... Les étapes qui précèdent la fouille de données portent sur l'acquisition et le pré-traitement de données. En effet, les données initiales peuvent être incomplètes, bruitées, aberrantes ou incohérentes (Famili *et al.*, 1997), (Soibelman *et al.*, 2002) d'où l'importance d'un travail prépartoire sur les données (Fayyad *et al.*, 1996, Fayyad *et al.*, 2001). Plusieurs techniques sont proposées, elles se basent essentiellement sur l'expertise humaine et utilisent les méta-données. En effet, dans (Engels *et al.*, 1998), les auteurs ont montré qu'il est utile d'avoir à la fois des informations sur les données à traiter et d'autres sur les techniques à utiliser. Afin d'exploiter l'information sur les techniques adéquates aux problèmes étudiés, (Cannataro *et al.*, 2003) proposent dans le cadre des grilles de calcul, une ontologie des différentes techniques de fouille. Dans le cadre de la fouille sur le web (web mining), (Hazman *et al.*, 2005) proposent une architecture d'extraction de connaissances qui fait appel à des règles heuristiques pour catégoriser les pages web (sorte de classification avant la fouille). Ces règles sont en fait des méta-données. Celles-ci sont également utilisées en aval de l'étape de fouille pour valider la connaissance extraite. Dans (Halkidi *et al.*, 2002), les auteurs proposent des indices pour valider les classes obtenues, en se basant sur des lois statistiques. Ces indices peuvent être vus comme des méta-données.

Ainsi, les méta-données sont très utiles dans le processus d'ECD. Souvent elles sont utilisées de manières informelles, sans qu'une modélisation a priori soit faite. Dans nos travaux, nous nous intéressons à la prise en compte des connaissances du

domaine (méta-données) dans la phase de fouille. Nous proposons une approche qui rend effective l'utilisation des méta-données dans la phase de fouille.

3. Représentation des méta-données

Dans de nombreux travaux, les méta-données ont servi à modéliser les attributs nécessaires à la spécification d'un objet donné. Ces travaux visent à déterminer les méta-données susceptibles d'être nécessaires pour la description complète d'un objet (auteur, date de création…). Parmi les premiers travaux sur les méta-données, il faut citer le *Dublin Core Metadata Initiative* (DCMI)[2]. Un autre modèle a été proposé dans le même cadre : le LOM (L*earning Object Metadata*), (Farance, 2003), qui concerne plus spécialement les objets pédagogiques. Cet intérêt pour les méta-données a créé le besoin d'un langage pour mieux les représenter : RDF (Miller, 1998), recommandé par le consortium *W3C* dans le cadre du web sémantique.

3.1. *RDF et RDF/S*

De manière plus générale, RDF permet de voir le Web comme un ensemble de ressources reliées par des liens sémantiques. RDF est basé sur un modèle s'exprimant sous forme de triplets "*ressource-attribut-valeur*", pouvant être lus comme "*sujet-prédicat-objet*" (P. Laublet, 2002), ou comme un graphe (Stuckenschmidt, 2004), (Corby *et al.*, 2000). Les triplets sont définis comme suit :

– Une ressource est une entité accessible par un URI (*Uniform Resource Identifier*) sur le web (un document HTML ou XML).
– Un attribut (ou propriété) définit une relation binaire entre les ressources et/ou les valeurs atomiques. Une propriété permet non seulement de relier l'information aux ressources, mais aussi de fournir des descriptions pour celles-ci.
– Une valeur peut être une chaîne de caractères simple (littéral), comme elle peut être une ressource.

Prenons un exemple simple d'une voiture exposée à l'adresse "http ://www.Peugeot.fr/id19" et de catégorie "break sw". Le triplet correspondant est : (http ://www.véhiculePeugeot.fr/id19, catégorie, "break sw"). Dans RDF, une même ressource peut avoir plusieurs descriptions, en d'autres termes, il est possible d'associer plusieurs triplets à une ressource donnée.

```
<rdf :RDF xmlns :rdf="http ://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf :Description about='http ://www.véhiculePeugeot.fr/id19'>
    <catégorie> break sw </catégorie>
    <nom> 407 </nom>
  </rdf :Description>
</rdf :RDF>
```

Listing 1 : Exemple de RDF exprimé à l'aide de la syntaxe XML.

2. http ://dublincore.org/usage/documents/overview/

RDF utilise aussi la notion de namespaces[3] déjà définie pour les schémas XML. D'autres mécanismes existent pour enrichir les moyens d'expression de RDF tels que le mécanisme de reification ou les containers, (Brickley *et al.*, 2004). Malgré sa simplicité, RDF/S est difficile à gérer en terme de performance. En effet, la vérification des schémas RDF dans les étapes de la recherche d'informations pénalise le temps de réponse. Les travaux de (Stuckenschmidt, 2004) sur le "cache sémantique", tentent d'améliorer les performances. Mais l'une des limites de RDF/S reste tout de même l'absence de mécanisme d'inférence. Bien que le RDF/S soit largement suffisant pour la représentation des méta-données, qui est sa véritable vocation, l'utilisation d'un autre formalisme, tel que les graphes conceptuels, est nécessaire pour disposer d'un moyen d'inférence.

3.2. *Graphes conceptuels*

Les graphes conceptuels (GCs), (Sowa, 1999), sont un mode de représentation des connaissances. Ils utilisent un support qui définit le vocabulaire des connaissances, (Mugnier *et al.*, 1996). Le support est constitué de cinq éléments (TC, TR, σ, I, τ), où :

TC : est un ensemble de types de concepts hiérarchiquement structurés ;

TR : est un ensemble de types de relations hiérarchiquement structurées ;

σ **:** est l'application qui permet d'ordonner les relations en prenant en compte leurs arguments ;

I : est un ensemble de marqueurs individuels (instances de concept) ;

τ **:** une application définie comme : $\forall i \in I, \tau(i) = t \in TC$

L'une des particularités du modèle des GCs est de permettre la représentation des connaissances sous forme graphique. Plus précisément, dans ce modèle un GC simple est défini comme un multigraphe[4] non orienté, biparti[5], non nécessairement connexe. Prenons l'exemple de la figure 1. Elle peut être interprétée intuitivement par : les véhicules créés par le constructeur Peugeot participent aux courses Rallye financées par ce constructeur.

Il est à noter que les concepts (rectangles) sont soit des marqueurs génériques (Véhicule), soit des marqueurs individuels (Constructeur : Peugeot et Course : rallye). Les relations sont quant à elles représentées par une ellipse. Le graphe conceptuel de la figure 1 peut être écrit à l'aide d'une notation que nous adopterons dans cet article et dont le sens de lecture est de gauche à droite :

3. http ://www.w3.org/TR/REC-xml-names/
4. Un multigraphe est un graphe tel qu'il peut exister plusieurs arêtes entre deux sommets, ici entre un sommet relation et un sommet concept.
5. Un graphe où il y a deux types de sommets. Dans notre cas, les deux types de sommets sont : concept (qu'il soit générique ou non) et relation.

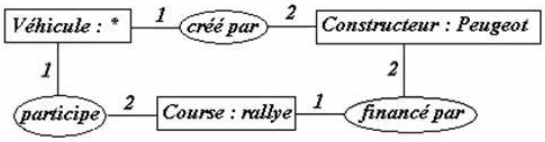

Figure 1. *Un graphe conceptuel simple connexe G.*

```
[véhicule : *] -{
      —> (créé par) —> [Constructeur : Peugeot]
      —> (participe) —> [Course : rallye] —>
         (financé par) — [Constructeur : Peugeot]}                    (G)
```

Listing 2 : Une autre notation du GC de la figure 1.

Le raisonnement sur les GCs repose sur une notion fondamentale dite "opération de subsomption" (ou encore projection) sur l'ensemble des GCs. La subsomption est une suite d'opérations élémentaires de spécialisation/généralisation. Ces opérations élémentaires sont unaires (opérant sur un seul graphe) ou binaires (opérant sur deux graphes). Soit un concept A qui peut être déduit à partir d'un autre concept B (au sens spécialisation) dans le support, alors nous pouvons conclure que le concept B peut être déduit aussi du concept A (au sens généralisation). De ce fait, l'étude d'un seul sens (par exemple spécialisation) est suffisant pour illustrer ce que les GCs apportent comme moyen de raisonnement. Les opérations de spécialisation sont : *la simplification, la restriction de concept , la restriction de relation, le joint interne, la somme disjointe*. Les opérations de généralisation (*duplication, augmentation de relation, augmentation de concept, éclatement, décomposition*) sont toutes unaires et peuvent être déduites des opérations de spécialisation. Pour illustrer le mécanisme de raisonnement (projection), prenons le graphe de la figure 1 (ou du listing 2) comme base de recherche et posons la question suivante : " Est-ce qu'il existe des voitures qui participent aux courses Rallye ? " Cette question est interprétée par le graphe conceptuel H du listing 3.

```
[Voiture : ?x]—>(participe)—>[course : Rallye]                        (H)
```

Listing 3 : Exemple d'une requête sur les graphes conceptuels.

Répondre à cette question revient à chercher s'il existe une projection du graphe G dans le graphe H. En d'autres termes, est-il possible de retrouver le graphe H à partir d'un ensemble d'opérations élémentaires finies de spécialisation appliquées sur le graphe G ? En effet, en appliquant la restriction de ce concept sur Véhicule, nous obtenons le graphe H de la figure 2.

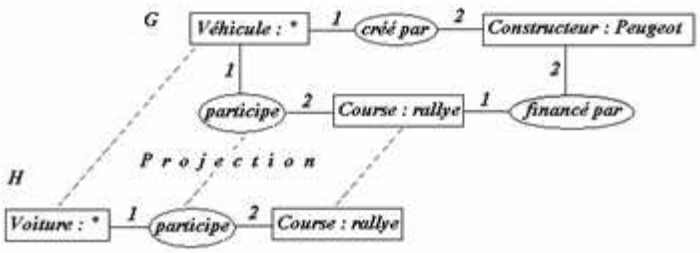

Figure 2. *Exemple d'opération de projection (subsomption).*

3.3. *De RDF aux graphes conceptuels*

D'après les travaux de (Corby *et al.*, 2000), il est possible d'utiliser RDF/S comme standard de représentation et d'autre part de tirer profit des avantages du raisonnement sur les GCs. L'idée est de considérer une description RDF/S comme un GC. Le modèle des GCs décrit dans 3.2 est basé sur (1) un support défini par le quintuplet (TC, TR, σ, I, τ) ; et (2) une base de GCs construits sur ce support. Il faut alors traduire :

1) les descriptions RDF en une base de GCs ;

2) la hiérarchie de classes qui apparaît dans un schéma RDF en une hiérarchie de concepts dans le support ;

3) la hiérarchie de propriétés qui apparaît dans un schéma RDF en une hiérarchie de relations dans le support.

```
[[Ressource :http ://www.véhiculePeugeot.fr/id19]-{
      —>(catégorie)—>[Litéral :break sw]]
      —> (participe) —> [Course : rallye] —>
         —>(nom)—>[Litéral :407 ]}                                    (I)
```

Listing 4 : Exemple de GC de la représentation RDF du listing 1.

En ce qui concerne le schéma RDF dédié à la représentation des connaissances, il suffit d'introduire le type de concept "ressource" au niveau le plus haut dans le support. Ainsi, toutes les classes nécessaires pour la construction d'un schéma RDF sont des spécialisations du concept "ressource" au niveau des GCs. Si le RDF/S permet une représentation facile et efficace des méta-données et que les GCs offrent un mécanisme intéressant de raisonnemnt, la combinaison de ces deux outils permet de bénéficier d'un dispositif utile aux traitement des méta-données, (Corby *et al.*, 2000).

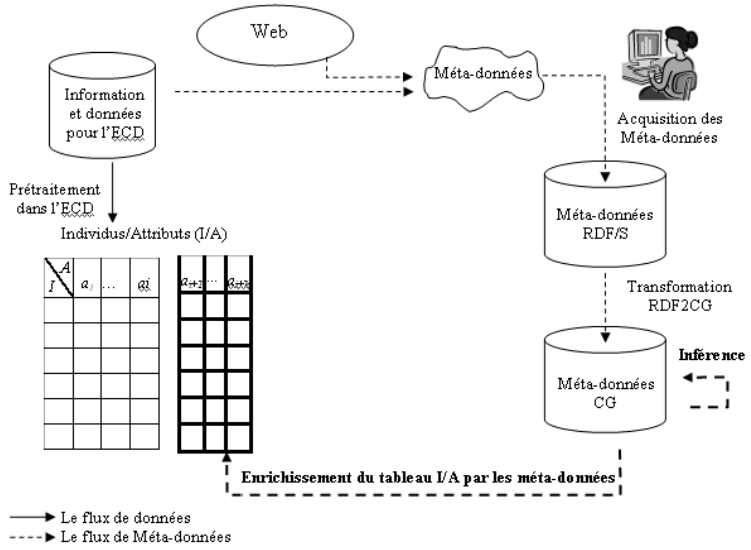

Figure 3. *Architecture de l'approche proposée.*

4. Approche proposée

Notre approche se décompose en quatre étapes (figure 3) : (1) l'acquisition des méta-données en RDF/S ; (2) la transformation des méta-données de RDF/S en GCs ; (3) la recherche des méta-données pertinentes pour la fouille ; (4) l'intégration de ces méta-données dans le tableau Individus/Attributs.

4.1. *Acquisition des méta-données*

A chaque domaine d'étude, est associé un schéma RDF qui permet de définir la manière dont les connaissances (méta-données) vont être spécifiées. Par conséquent, il est impératif de prendre en compte le schéma RDF associé aux méta-données. Nous supposons que les méta-données sont modélisables sous forme de règles (c.f. Listing 6). Les schémas RDF "cos" et "c" utilisés pour représenter ces règles sont ceux utilisés dans CORESE[6] (Conceptual Resource Search Engine), (Corby *et al.*, 2002).

6. http ://www-sop.inria.fr/acacia/corese/

4.2. *Transformation des méta-données en GCS*

Dans certains cas, il est nécessaire d'effectuer des traitements sur les méta-données pour extraire des informations implicites et pertinentes. Afin d'inférer sur les méta-données représentées jusqu'alors en RDF/S, celles-ci sont transformées en GCs. Lors de cette transformation, les schémas RDF doivent être intégrés au support global qui contient tous les concepts d'un ou de plusieurs domaines. Par la suite, une mise en correspondance des balises des schémas avec les types de concepts et les types de relations du support est faite. L'intervention de l'utilisateur est requise dans cette étape. Nous avons fait le choix, à ce stade de nos travaux, de traiter manuellement cette phase. Car l'intégration automatique de ces schémas RDF dans le support relève d'un autre domaine de recherche : l'alignement et la fusion d'ontologies. Cette problématique a pour objectif d'étudier l'utilisation dans un système de plusieurs représentations des connaissances d'un même domaine ou de domaines connexes, (Furst, 2004). Une fois les méta-données RDF transformées en GCs, comment rapprocher celles-ci du tableau Individus/Attributs ?

4.3. *Algorithme de détection des méta-données*

La seule relation existant entre les données et les méta-données est celle qui porte sur les individus ou sur les attributs. Nous disposons des informations portant sur les attributs utilisés dans l'expression des connaissances. Plus précisement, les occurrences des attributs qui apparaisent dans les parties *condition* et *conclusion* des règles. Dans ce cas, une correspondance doit être étabie entre les attributs du tableau Individus/Attributs et les concepts pour pouvoir inférer sur la base des GCs. Partant de cette correspondance, il est alors possible d'inférer sur la base des GCs pour chercher les méta-données pertinentes. Nous présentons l'inférence sous la forme d'une heuristique, nous permettant : (1) de détecter les méta-données pertinentes pour chaque attribut ; (2) de générer de nouveaux attributs ; (3) de valoriser les nouveaux attributs à l'aide des valeurs déjà existantes des individus ou des valeurs inférées.

Les méta-données ne sont pas forcément toutes intéressantes pour être intégrées de manière systématique. C'est pourquoi nous proposons un algorithme basé sur des heuristiques, capable de détecter les méta-données intéressantes et de générer de nouveaux attributs. Nous avons montré un exemple illustratif où les méta-données sont sous forme de règles, il existe également une autre forme de méta-données pouvant être utilisées dans le même contexte. Cette forme peut ne pas avoir une relation directe avec le domaine étudié. Prenant un exemple simple, celui de l'attribut $typedejour$ qui indique le jour de l'accident, et il prend ses valeurs dans l'ensemble $\{lundi, mardi, mercredi, jeudi, vendredi, samedi, dimanche\}$. En faisant une correspondance avec les concepts du support, cet attribut est associé au concept $jour$. Ce dernier est soit un jour de weekend soit un jour de semaine. Par conséquent nous pouvons déduire un nouvel attribut $typedejour$ qui prend ses valeurs dans l'ensemble $\{jour - weekend, jour - semaine\}$. Ce nouvel attribut ajoute une sé-

mantique aux données. La méta-donnée associée à cet attribut est dans le support et non dans la base des GCs. Elle n'appartient donc pas au domaine. Il est donc possible d'utiliser les méta-données qui ne sont pas liées au domaine. En effet, l'intérêt des méta-données dépend des correspondances possibles à trouver d'une part, et de leur forme (règles ou pas) d'autre part. L'algorithme considère ces deux cas : (1) l'attribut est un *concept* (exemple : $jour$ correspond au concept $typedejour$, la méta-donnée cherchée est alors dans le support ; (2) l'attribut est une *relation* associée à un concept (exemple : $poidsduvéhicule$ correspond au couple $(concept, relation) =$ *(véhicule, poids)*, –notons cette correspondance par $(C, R) = (C1, R1)$– la méta-donnée cherchée est alors dans la base des GCs, sous forme de règles.

$1^{ier}Cas$

Pour tous les concepts C faire :
Chercher le concept C dans le support associé à l'attribut ;
Une fois C trouvé, chercher ς ensemble de sous concepts de C
ayant le même niveau hiérarchique dans le support et dont le domaine
de valeurs de l'attribut existant est le même que celui de ς ;
Si $\varsigma \neq vide$:
- Ajouter le concept C comme nouvel attribut et
ς comme domaine de valeurs de cet attribut.
- Valoriser cet attribut dans le tableau Individus/Attributs,
en tenant compte des valeurs de l'attribut correspondant.
Sinon, pas d'enrichissement sémantique pour cet attribut.
Fin Si

$2^{ier}Cas$ **:**

Pour tout attribut A correspondant à (C, R) faire :
(1) Chercher tous les GCs dont :
a) - la paire (C, R) participe à une règle,
b) - le concept C appartient, à la fois, à la condition
et à la conclusion de la règle,
c) - le concept de la conclusion C1 est littéral.
(2) Regrouper les graphes qui ont une même paire (C1, R1) avec R1, une relation appartenant à la conclusion de la règle et attachée à C et C1.
Pour chaque groupe de graphes faire :
a) - sélectionner la paire (C1, R1) comme un nouvel attribut A1
et son domaine de valeurs est le marqueur individuel C1 ;
b) - valoriser A1 en tenant en compte des valeurs
de l'attribut A et de la règle associée à (R, C) ;
Fin pour
Fin pour

Les attributs découverts à partir des méta-données ajoutent une sémantique aux données, et sont représentés comme de nouvelles colonnes dans le tableau Individus/Attributs. Les valeurs de ces colonnes sont déduites à partir des valeurs des attributs déjà existants et des contraintes sur le nouveau domaine de valeurs. Ces nouveaux

attributs peuvent également se substituer aux attributs existants. C'est à l'utilisateur que revient le choix de remplacer ou d'ajouter ces nouveaux attributs.

5. Expérimentation

5.1. *Protocole*

Considérons le cas d'un expert de l'accidentologie qui s'intéresse aux accidents de la route dont il veut connaître le type : grave ou léger. Imaginons qu'il souhaite construire un modèle qui lui permette de comprendre et de prévoir la gravité des accidents. Pour cela, il peut procéder par apprentissage à partir de données. Cela consiste, pour lui, à recueillir des informations sur des accidents dont il sait s'ils sont graves ou légers. Sur la base de ce corpus dit " échantillon d'apprentissage ", il met en œuvre une méthode d'apprentissage supervisé pour construire le modéle d'explication et de prédiction de la gravité des accidents. Plusieurs méthodes de fouille de données peuvent être utilisées dans ce cadre : les arbres de décision, les réseaux de neurones, les méthodes de régression, l'analyse discriminante, les réseaux bayésiens, les régles d'association, etc. Dans le cadre de cet article, nous avons choisi comme méthode de fouille les arbres de décision avec l'algorithme C4.5, (Quinlan, 1993).

Pour expérimenter notre approche nous procédons selon les étapes suivantes :

1) Choisir un jeu de données sur l'accidentologie (un tableau individus/variables) et disposer de quelques méta-données sur ce domaine ;

2) Modéliser les méta-données en RDF/S ;

3) Transformer ces méta-données en GCs (parmi les outils existants, CORESE est retenu) ;

4) Appliquer l'algorithme heuristique pour détecter les méta-données intéressantes et en déduire de nouveaux attributs ;

5) Faire la fouille de données sur le tableau initial Individus/Attributs, en utilisant les arbres de décision ;

6) Faire la fouille de données sur le tableau Individus/Attributs avec les nouveaux attributs issus des méta-données, en utilisant les arbres de décision ;

7) Comparer les deux résultats pour évaluer l'approche proposée.

Les attributs issus des méta-données peuvent être ajoutés ou être substitués.

5.2. *Jeu de données et mise en oeuvre*

Les données considérées portent sur 952 individus (accidents routiers) et 23 attributs. Parmi ces attributs nous avons choisi comme attribut (classe) à expliquer " la gravité de l'accident " qui prend ses valeur dans l'ensemble {tué, blessé grave, blessé

léger, indemne}. Les attributs explicatifs sont de type nominal ou continu et sont par exemple : la vitesse, le poids du véhicule, le type de route …

Le tableau 1 rassemble les 23 attributs du jeu de données avant l'insertion des méta-données.

N°Attribut	Caractéristiques des attributs		
	Nom	Dom. Valeurs	Type (C/N)
1	type de collision	1..9	N
2	type de route	1..4	N
3	tracé en plan	1..2	N
4	mise en circulation du véhicule	1..5	N
5	appartenance	1..3	N
6	tranche d'âge du conducteur	1..3	N
7	sexe	1..2	N
8	année du permis	1..3	N
9	motif	1..4	N
10	csp	1..11	N
11	infraction	1..2	N
12	priorité	1..2	N
13	infraction véhicule	1..2	N
14	responsable	1..2	N
15	alcool	1..2	N
16	mois	1..12	N
17	jour	1..7	N
18	nombre de véhicules	1..25	C
19	nombre de piétons	0..3	C
20	poids du véhicule	560.0...1881.8	C
21	puissance relative	26.0..298.4	C
22	vitesse	106.7..252.8	C
23	km compteur	354.0.. 922103.0	C

Tableau 1. *Les attributs du jeu de données de l'accidentologie considéré.*

D'autre part, nous disposons d'un certain nombre de méta-données à partir desquelles nous allons extraire des données qui seront injectées dans le tableau initial. Dans le listing 5, figure un extrait des méta-données sur l'accidentologie.

1. Les tranches de poids ;
(a) Un véhicule est léger si son poids est moins de 800kg,
(b) Un véhicule est moyen si son poids est entre 800kg et 1000kg,
(b) Un véhicule est lourd si son poids est de 1000kg et plus.

2. L'infraction de vitesse ;
(a) vitesse max pour une route de type autoroute est 130 km/h,
(b) vitesse max pour une route de type RN est 80 km/h,
(c) vitesse max pour une route de type CD est 70 km/h,
(d) vitesse max pour une route de type autre route est 50 km/h,
(e) si la vitesse max est dépassé alors : infraction vitesse,
(f) si la vitesse max n'est pas dépassé alors : pas d'infraction vitesse.

3. Les saisons ;
(a) Un mois est soit un mois (d'hiver, de printemps,d'été, d'automne),
(b) Les mois d'été sont (juin, juillet, août),
(a) Les mois d'automne sont (septembre, octobre, novembre),
(d) Les mois d'hiver sont (décembre, janvier, février),
(e) Les mois de printemps sont (mars, avril, mai).

Linsting 5 : Exemple de méta-données sur le domaine de l'accidentologie.

Un exemple de modélisation des méta-données en RDF/S et en GC est présenté dans les listings 6 et 7.

```
<cos :rule>
  <cos :if>
    <rdf :Description rdf :about='véhicule'>
      < :vitesse>
        <c :valeur rdf :about=' ?v1'>
          <c :supérieur>
            <c :valeur rdf :about=' ?v2'>
              <c :vitesseMaxDe>
                <c :route rdf :about=' ?tr1'>
                  <c :typeRouteDe>
                    <rdf :Description rdf :about='#véhicule'/>
                  </c :typeRouteDe>
                </c :route>
              </c :vitesseMaxDe>
            </c :valeur>
          </c :supérieur>
        </c :valeur>
      </c :vitesse>
    </rdf :Description>
  </cos :if>
  <cos :then>
    <rdf :Description rdf :about='véhicule'>
      <c :infractionVéhicule> infraction vitesse <c :infractionVéhicule>
    </rdf :Description>
  </cos :then>
</cos :rule>
```

Listing 6 : La méta-données 2.(e) en RDF.

```
Véhicule :*]-{
    —>(infractionVéhicule)—>[Litéral :InfractionVitesse]
    —>(vitesse)—>[Valeur : ?v1]-{
      —>(inférieur)—>[Valeur : ?v2]-{
        —>(vitesseMaxDe)—>[route : ?tr]-{
          —>(typeRouteDe)—>[Véhicule :*]}}}}
```

Listing 7 : La méta-donnée 2.(e) transformée en CG.

L'inférence sur les méta-données permet d'obtenir de nouveaux attributs qui se substituent ou s'ajoutent aux attributs existants et qui viennent renforcer la sémantique des données. Ils permettent ainsi une meilleure compréhension des résultats de la fouille de données. Le tableau 2 présente quelqu'uns des nouveaux attributs (en italique) obtenus à partir des méta-données.

Nous réalisons la fouille sur les deux tableaux (initial et après enrichissement) avec l'arbre de décision C4.5.

5.3. *Résultats*

Les résultats obtenus de la fouille sans les méta-données sont montrés dans la matrice de confusion (tableau 3).

Le coût de mauvais classement (ou taux d'erreur) est de 67,32% ce qui est un mauvais résultat avec seulement un tiers des accidents bien classés.

N°Attribut	Caractéristiques des attributs		
	Nom	Dom. Valeurs	Type (C/N)
1	type de collision	1..9	N
2	type de route	1..4	N
3	tracé en plan	1..2	N
10			
11	...	...	...
12			
16	mois	1..12	N
17	*type de jours*	*1..2*	*N*
18	nombre de véhicules	1..25	C
19	nombre de piétons	0..3	C
20	*classe de poids du véhicule*	*1..3*	*N*
21	puissance relative	26.0..298.4	C
22	vitesse	106.7..252.8	C
23	km compteur	354.0.. 922103.0	C
24	*vitesse autorisée*	*1..4*	*C*

Tableau 2. *Les attributs du jeu de données enrichi par les méta-données.*

Prédit observé	Tués	Blessés graves	Blessés légers	Indemnes
Tués	5	25	0	0
Blessés graves	1	34	7	2
Blessés légers	1	33	8	0
Indemnes	0	33	1	3

Tableau 3. *Coût de mauvais classement sur le tableau initial.*

Le résultat de la fouille avec l'intégration des méta-données est présenté dans le tableau 4. Le nombre total d'individus dans les deux tableaux n'est pas le même, cela s'explique par le choix que nous avons fait concernant les données manquantes. En effet, nous avons décidé de ne pas prendre en compte les individus ayant des données manquantes. Par conséquent, comme les attributs dans les deux tableaux (sans et avec les méta-données) ne sont pas forcement les mêmes, le nombre d'individus non pris en compte (supprimés) dans ces deux tableaux est différent.

Le tableau 4 montre que le coût de mauvais classement est de 36,33%, soit une amélioration d'une trentaine de points par rapport à la fouille sans les méta-données.

Prédit observé	Tués	Blessés graves	Blessés légers	Indemnes
Tués	142	15	1	8
Blessés graves	5	36	12	14
Blessés légers	8	14	4	8
Indemnes	2	20	6	16

Tableau 4. *Coût de mauvais classement sur le tableau intégrant les méta-données.*

Ainsi nous avons montré que les méta-données peuvent être intégrées dans la fouille de données et que cette approche d'intégration se révéle être efficace pour améliorer les résultats de la fouille de données.

6. Conclusion et Perspectives

Rappelons, briévement, que notre objectif est d'utiliser les méta-données dans la fouille afin d'enrichir les données et d'améliorer les résultats. L'approche proposée est constituée de quatre étapes : (1) acquisition des méta-données en RDF/S ; (2) transformation des méta-données de RDF/S en graphes conceptuels ; (3) recherche des méta-données pertinentes et construction de données enrichies par les méta-données et (4) intégration de ces données enrichies.

Afin de valider et tester cette approche nous avons utilisé des données et des méta-données sur le domaine d'accidentologie. Les résultats montrent une amélioration d'un tiers du taux de mauvais classement. L'approche proposée s'avére être efficace pour améliorer la fouille.

Bien qu'elle soit incomplète, l'approche proposée est la première tentative dans ce domaine. En effet, il est important de rappeler que dans le cadre de notre problématique, aucune tentative n'était faite auparavant. Pour compléter cette approche, certaines perspectives nous semblent nécessaires pour appliquer l'approche à grande échelle :

– Utiliser le moteur de recherche CORESE comme partie intégrante dans l'approche. Ce qui nous amène à développer une application qui intégre CORESE et s'interface avec les outils de fouille ;

– Valider l'approche par d'autres expérimentations sur des données complexes ;

– Tester l'approche avec d'autres méthodes de fouille de données ;

– Elaborer une maniére automatique ou semi-automatique pour l'acquisition des méta-données.

7. Bibliographie

Berkhin P., *Survey of clustering data mining techniques*, Technical report, Accrue Software, 2002.

Breiman L., Friedman J., Olshen R., Stone C., *Classification and regression trees*, Technical report, Wadsworth International, Monterey, CA, 1984.

Brickley D., Guha R., « RDF Vocabulary Description Language 1.0 : RDF Schema », *W3C Recommendation http ://www.w3.org/TR/2004/REC-rdf-schema-20040210/*, 2004.

Cannataro M., Comito C., « A Data Mining Ontology for Grid Programming », *1st International Workshop on Semantics in Peer-to-Peer and Grid Computing, in conjunction with WWW2003, Budapest*, p. 20-24, 2003.

Corby O., Dieng R., Hebert C., « A conceptual graph model for w3c resource description framework », *International Conference on Conceptual Structures (ICCS00), Darmstadt, Germany*, LNAI, Springer, 2000.

Corby O., Faron-Zucker C., « Corese : a Corporate Semantic Web Engine », *workshop on Real World, RDF and Semantic Web applications at the 11th International World Wide Web Conference (WWW02)*, 2002.

Engels R., Theusinger C., « Using a data metric for offering preprocessing advice in data mining applications », *Thirteenth European Conference on Artifical Intelligence*, p. 430-434, 1998.

Famili A., Shen M., Weber R., E. E. S., « Data Preprocessing and intelligent Data Analysis », *Intelligent Data Analysis*, vol. 1, p. 3-23, 1997.

Farance F., « IEEE LOM Standard Not Yet Ready For 'Prime Time' », *IEEE Computer Society Learning Technology Task Force [LTTF]*, 2003.

Fayyad U., Grinstein G., Wierse A., *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, 2001.

Fayyad U., Piatetsky-Shapiro G., Smyth P., (Eds.) R. U., *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996.

Furst F., Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationnalisation, PhD thesis, Ecole Polytechnique de l'Université de Nantes (EPUN), France, 2004. Thèse de doctorat.

Halkidi M., Batistakis Y., Vazirgiannis M., « Cluster validity methods : Part I », *SIGMOD Record*, vol. 31, n° 2, p. 40-45, 2002.

Hazman M., El-Beltagy S., Rafea A., El-Gamal S., « Knowledge Discovery From The Web », *7th International Conference on Entreprise Information Systems (ICEIS05), Miami U.S.A*, vol. 2, p. 25-28, 2005.

Miller E., « An introduction to the resource description framework », *D-Lib Magazine*, 1998.

Mugnier M., Chein M., « Représenter des connaissances et raisonner avec des graphes », *Revue d'Intelligence Artificielle*, vol. 10, n° 1, p. 7-56, 1996.

P. Laublet C. Reynaud J. C., « Sur quelques aspects du Web Sémantique », *Deuxièmes assises nationales du GdRI3*, p. 59-78, 2002.

Quinlan J., *C4.5 : Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

Soibelman L., Asce M., Hyunjoo K., « Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases », *Journal Of Computing In Civil Engineering*, January, 2002.

Sowa J., « Conceptual Graph Standard », http ://www.bestweb.net/_sowa/cgdpans.htm, 1999.

Stuckenschmidt H., « Similarity-Based Query Caching », *6th international conference on Flexible Query Answering Systems (FQAS), Lyon, France*, 2004.

Zighed D., Rakotomalala R., *Data Mining*, vol. H3 744 of *Techniques de l'ingénieur*, Editions Techniques de l'Ingénieur, p. 1-26, 2002.