Extraction de métadonnées sur les prototypes issus de la classification d'objets

Abdourahmane Baldé[1], Yves Lechevallier[1], Marie-Aude Aufaure[2]

[(1)]*INRIA Rocquencourt*
Domaine de Voluceau
Rocquencourt - B.P. 105
78153 Le Chesnay Cedex
[(2)]*Supélec Plateau du Moulon*
3, rue Joliot-Curie
91192 Gif-sur-Yvette cedex

RÉSUMÉ. Nous présentons ici une méthodologie d'extraction des métadonnées sur des prototypes issus de la classification d'objets (pouvant être élargie aux objets symboliques). Les métadonnées, utilisées généralement pour une meilleure gestion de l'information, seront créées dans le but d'archiver des informations jugées pertinentes lors du processus de classification. Cette étude a été validée en s'appuyant sur des données issues d'une enquête. L'objectif étant de pouvoir décrire un ensemble de données en fournissant toute l'information les concernant et pas seulement l'information d'ordre bibliographique.

MOTS-CLÉS : classification, métadonnées, objets symboliques, ontologies, rdf, dublin core

1 Introduction

La maîtrise sur des grands ensembles d'information devient de plus en plus complexe et de plus en plus fastidieux. Les métadonnées constituent une voie pour aider l'utilisateur ou le gestionnaire d'information à comprendre, retrouver, comparer des informations sans forcément avoir recours directement au contenu de celles-ci.

En effet, les métadonnées peuvent être vues comme étant des données structurées qui décrivent les données et qui peuvent s'appliquer à tous types de données.

L'objectif dans ce travail aura été de construire des métadonnées issues depuis la génération des données jusqu'à leur classification. Celles-ci devant rendre compte du contenu des classes créées, des méthodes de classification utilisées et des informations d'ordre général (l'auteur, l'éditeur, la date de création etc.).

Ce travail pourrait ainsi être élargi au domaine de l'analyse des données symboliques. En effet, les données qui seront manipulées peuvent être des objets symboliques (objets qui constituent les individus de l'analyse des données symboliques, permettant de représenter des individus complexes ou des classes d'individus par des conjonctions de propriétés ou des descripteurs) [Diday, 1998].

Dans le paragraphe 1, nous présentons une idée sur les notions de métadonnées, de classification et d'objet symbolique[1]. Dans le paragraphe 2, nous précisons nos idées en décrivant notre approche. Une simulation est fournie dans le paragraphe 3. Et nous concluons.

2 Notions générales

Dans cette étude, nous utiliserons les définitions données par [Bui thi et al. , 2001], [Diday, 2003] et [Diday, 1998] en matière de métadonnées, de classification et d'objet symbolique respectivement.
Ainsi, les *métadonnées* sont définies comme des informations émises à un niveau d'abstraction supérieur et relatives à un niveau d'abstraction inférieur. Ce qui fait intervenir les notions de réflexivité et d'abstraction.
La *classification automatique* quant à elle, est définie comme un ensemble de méthodes et algorithmes consistant à découper une population d'objets en plusieurs classes, en tenant compte des variables qui les caractérisent et de la mesure de ressemblance choisie.
L'*objet symbolique* est quant à lui défini par une description notée 'd' ; une relation binaire 'R' sur D permettant de comparer d à une autre description de D ; une fonction 'a' permettant d'évaluer le résultat de la comparaison (à l'aide de R) de la description d'un individu du monde réel par rapport à la description donnée 'd'.
Notre approche tiendra compte aussi des normes déjà existantes. Ainsi, les éléments du *Dublin Core* ont largement été utilisés pour constituer les métadonnées (de type bibliographique) de nos fichiers de données/métadonnées.
En effet, le *Dublin Core* est un ensemble de 15 éléments simples qui définissent les catégories d'information à enregistrer à propos d'une ressource (page Web, document ou image) pour que celle-ci puisse être trouvée.
Le schéma *RDF* (composé de trois éléments : *ressource, propriété, déclaration*) sera utilisé pour fournir une description des éléments de métadonnées extraits. *RDF* définit la signification, les caractéristiques et les relations d'un ensemble de propriétés.
Résultant du travail du W3C (*Consortium du world Wide Web:* créateur des standards pour le Web), RDF définit une structure de métadonnées pour décrire le contenu du Web à l'aide du langage *XML* [Gardarin, 2002] ainsi que les relations entre ressources.
Voici un exemple faisant usage de la norme RDF, dans cet exemple on veut expliquer que l'auteur de la ressource *«Web et ontologies»* est *«Marie-Aude» :*

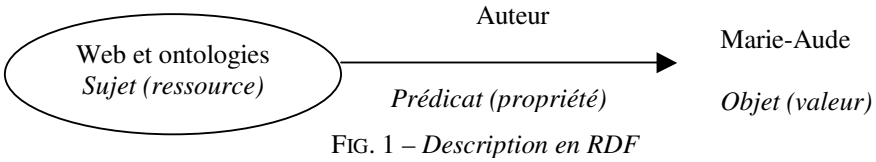

FIG. 1 – *Description en RDF*

3 Idées de base

Pour réaliser ce processus d'extraction, nous partons des données recueillies lors d'une enquête. L'idée étant de définir un ensemble d'éléments de métadonnées pouvant rendre compte d'informations portant sur les données recueillies. Ensuite, définir de nouveaux éléments de métadonnées lors de la phase de classification automatique.
Pour ce faire, nous nous sommes inspirés des éléments de [Csernel, 2002] et du Dublin Core. L'objectif poursuivi étant, bien évidemment, la définition d'éléments pouvant rendre compte des informations sur, d'une part, les données originales et agrégées, et sur les classes obtenues d'autres parts.

4 Méthodologie d'extraction des métadonnées

Les métadonnées que nous avons extraites ont la spécificité d'être intégrées avec les données qu'elles décrivent. Ce qui a l'avantage d'être simple, clair et facile à comprendre pour les utilisateurs.

[1] Nous utiliserons le terme «objet symbolique» pour parler de données agrégées et ce dans un souci de généralisation de ce travail à tous types de données agrégées y compris les objets dits symboliques

Ce travail consiste donc à extraire de manière automatique des informations jugées pertinentes au cours du processus de classification et de collecte des données. Cette étape d'extraction de métadonnées s'est déroulée en plusieurs phases que nous développons ci-après.

4.1 Méthodologie

La méthodologie d'extraction utilisée est celle qui consiste à réaliser l'extraction en 3 phases :

1. Renseigner les éléments de l'entête. Ces éléments sont constitués essentiellement par les éléments du Dublin Core.
2. Renseigner les éléments de métadonnées spécifiques à la classification automatique. Au cours de cette phase particulièrement périlleuse et ce, compte tenu du nombre important d'éléments à renseigner, notre travail a été tout d'abord de donner une sémantique claire à chacun de ces éléments qui étaient déjà définis par [Csernel, 2002].

Extraire, lors de l'application des différentes méthodes de classification, les informations afin de créer un historique des prototypes d'objet. En effet, nous devons être en mesure, lorsque nous disposons d'une classe d'objets, de connaître des informations sur l'origine de ces données, sur le critère de classification choisi, etc.

4.2 Création et maintenance des métadonnées

Ici, nous tentons de faire une analyse sur le processus de création et de maintenance des métadonnées dans le cadre général.

Ce processus de création de métadonnées peut se diviser en quatre (4) étapes :

- La définition des besoins : définir les besoins de l'organisation qui souhaite l'intégration des métadonnées. Cette phase doit tenir compte des normes déjà en vigueur afin de faciliter l'interopérabilité avec d'autres organismes.
- L'extraction et l'intégration des métadonnées : intervient après la définition de l'ensemble des éléments de métadonnées à renseigner.
- La promotion des métadonnées : rien ne sert à créer des métadonnées si l'on ne peut les faire découvrir à d'autres gens. Ainsi, le meilleur moyen de faire connaître «ses » métadonnées, c'est de les publier par le Web.
- La maintenance des métadonnées : mettre à jour les métadonnées dès que les données qu'elles décrivent changent. La réussite de cette étape dépendra de deux facteurs importants : le taux d'obsolescence des données décrites et les moyens mis à disposition par les organisations concernées.

5 Simulation

Dans cette section, nous allons illustrer nos propos (des sections précédentes) en fournissant les résultats issus d'extraction automatique de métadonnées dans le cadre de la classification d'objets.

Pour ce faire, nous allons partir d'une base de données qui décrit 150 iris. Dans cette base, les iris sont décrits par quatre variables (longueur et largeur du sépale, longueur et largeur du pétale). A ces individus, on applique une méthode de classification non supervisée. Cette dernière nous fourni trois classes d'iris avec pour chacune la description des individus qui la composent.

A la suite de cette classification, nous avons obtenu trois classes homogènes. Les métadonnées extraites et relatives aux individus ont cette structure :

```
<Entête>
   <dc : title>Iris</dc : title>
   <dc : author>Yves Lechevallier</dc :author>
   <dc : date>11/03/04</dc : date>
   <dc : language>Français</dc : language>
</Entête>
<OrigInfo>
   <NbOrigVar>4</NbOrigVar>
   <NbOrigMat>1</NbOrigMat>
   <PopSampSize>150</PopSampSize>
</OrigInfo>
```

```
<OrigVar>
        <Num>1</Num>
        <Name>Iris</Name>
        <Label>longueur du sépale</Label>
        <Computed>select * from IRIS</Computed>
</OrigVar>
<MetaInd>
        <Num>1</Num>
        <Name>Setosa</Name>
        <Operator>Native Data</Operator>
</MetaInd>
```

Nos fichiers de métadonnées se compose de trois parties : la première est relative aux informations d'ordre général (titre, auteur, etc.), la seconde est relative à la description des variables (qui décrivent nos individus), la dernière est relative aux objets qui sont agrégées (nous avons ainsi des informations sur l'opérateur d'agrégation, le nombre d'individus qui ont été agrégés pour former notre objet, etc.). Ainsi, pour chacune des classes, nous avons, en plus de l'entête, la description suivante :

<*MetaHistory*> <*SqlQuery*>**select * from IRIS_Classe**</*SqlQuery*> <*Source*>**c:\user\yves\asso\iris**</*Source*> <*OdbcSource*>**11/03/04**</*OdbcSource*> </*MetaHistory*>	<*MetaInd*> <*Num*>**2**</*Num*> <*Name*>**classe3/3**</*Name*> <*NbInObj*>**42**</*NbInObj*> <*GNbInObj*>**42**</*GNbInObj*> < *Operator* >**agregrated**</*Operator*> </*MetaInd*>

Nous voyons, à partir de ces deux bouts de résultats, que la construction d'un historique est relativement facile si nous prenons les métadonnées comme base à cette opération.
A la suite de cette étape, nous avons obtenu des métadonnées relatives aux individus à classifier et celles qui sont relatives aux classes d'individus. Pour exemple, nous avons la description de la 3ème classe (classe3/3) qui nous donne le nombre d'individus de la classe (42) ainsi que le nombre d'individus agrégés pour créer celle-ci (42).

6 Conclusion

Les métadonnées sont un instrument qui transforme les données brutes en connaissances. Elles représentent une valeur ajoutée à l'information en permettant leur compilation et leur repérage. Malgré la différence de structure, tous les types de métadonnées poursuivent un objectif commun : offrir des éléments de description pour faciliter l'accès à des ressources données en fournissant toute l'information les concernant. Le W3C travaille énormément dans le but de donner une dimension supplémentaire à l'utilisation des métadonnées. Elles constituent un véritable moyen de capitalisation des connaissances et du savoir-faire. C'est d'ailleurs la perspective qui paraît la plus prometteuse.
En effet, on pourrait faire intervenir les ontologies dans ce processus de capitalisation et de représentation des connaissances (cf. [Kassel, 2002] et de [Kassel et al, 2000]). Les ontologies peuvent apporter une dimension sémantique aux métadonnées et permettre surtout de faire face à la complexité d'organisations taxonomiques.
Une autre perspective serait d'utiliser RDF pour représenter les liens entre les différentes ressources de métadonnées obtenues au cours d'un processus de classification (cf. au document traduit par Karl Dubost*(http://www.la-grange.net/w3c/REC-rdf-syntax/),* relatif aux spécificités de RDF).
Pour faciliter la définition des métadonnées, RDF aura un système de classe comme dans tout environnement de programmation orienté objet et de modélisation. Ces classes, organisées en hiérarchie, offrent une extensibilité grâce à la subtilité des sous-classes. De cette façon, pour créer un schéma légèrement différent d'un autre déjà existant, il n'est pas nécessaire de "réinventer la roue" mais il faut juste fournir des modifications incrémentales au schéma de base.

7 Bibliographie

[Bui thi et al. , 2001] M.P. Bui thi, P. Joly, P. Faudemay, La description du contenu de la vidéo selon les points de vue de la production audiovisuelle, CIDE 01 (Conférence Internationale du Document Electronique), Toulouse, octobre 2001.

[Csernel, 2002] M. Csernel, Meta Data Specification for ASSO (Analysis System of Symbolic Official data) Library, 20-01-2002

[Diday, 2003] E. Diday, Cours de DEA-Dauphine, année universitaire 2002-2003

[Diday, 1998] E. Diday. L'analyse des Données symboliques : un cadre théorique et des outils. Rapport du CEREMADE n°9821, 1998.

[Gardarin, 2002] G. Gardarin, XML : des bases de données aux services Web, Dunod, 2002

[Kassel, 2002] G. Kassel, OntoSpec : une méthode de spécification semi-informelle d'ontologies, in Actes des journées francophones d'Ingénierie des Connaissances (IC'2002), pages 75-87, 2002.

[Kassel et al. , 2000] C. Barry, C. Irastorza, G. Kassel, M. Abel, P. Boulitreau et S. Perpette, Construction et exploitation d'une ontologie pour la gestion des connaissances d'une équipe de recherche, in Actes des journées francophones d'Ingénierie des Connaissances (IC'2000), 2000.