Annotation sémantique floue de tableaux guidée par une ontologie

Gaëlle Hignette*, Patrice Buche*
Juliette Dibie-Barthélemy*, Ollivier Haemmerlé**

*UMR INA P-G/INRA MIA - INRA Unité Mét@risk
INA P-G, 16 rue Claude Bernard, F-75231 Paris Cedex 5, France
{hignette, buche, dibie}@inapg.fr
**GRIMM-ISYCOM, Univ. Toulouse le Mirail, Dpt. Mathématiques-Informatique
5 allées Antonio Machado, F-31058 Toulouse Cedex
ollivier.haemmerle@univ-tlse2.fr

Résumé. Nous présentons dans cet article différentes étapes de l'annotation de tableaux de données à l'aide d'une ontologie. Tout d'abord, nous distinguons les colonnes de données numériques et symboliques. Les données symboliques sont ensuite annotées de manière floue à l'aide des termes de l'ontologie. Cette annotation nous permet de déduire le type des colonnes de données symboliques. Pour trouver le type des colonnes de données numériques, nous utilisons à la fois le titre de la colonne et les valeurs numériques et unités présentes dans la colonne. Chaque étape de notre annotation est validée expérimentalement.

1 Introduction

Dans le monde scientifique, de nombreuses données sont produites en continu : il est difficile de se maintenir à jour avec le flot d'informations, et de synthétiser les données venant de sources diverses au moment où on en a besoin. Notre but est la construction d'un entrepôt de données XML sur un domaine d'application précis, où différentes données collectées sur le Web seront annotées avec une ontologie du domaine, de manière à être facilement interrogeables. Notre travail se concentre sur l'annotation des tableaux de données, qui sont un moyen de présenter l'information de façon synthétique, très utilisé dans les domaines scientifiques et économiques.

La structure des tableaux de données que l'on trouve dans les rapports et publications scientifiques collectés sur le Web est très hétérogène : elle varie d'un auteur à l'autre, et on observe même souvent différentes formes de tableaux dans un même article scientifique. De plus, le fait que l'on s'intéresse à des tableaux nous prive de l'utilisation d'un contexte linguistique : les techniques de *wrapper induction* basées sur la structure (Baumgartner et al., 2001) ou le contexte linguistique (Freitag et Kushmerick, 2000) ne sont donc pas adaptées à notre problème d'annotation. Notre but est de construire un outil d'annotation sans phase d'apprentissage, reposant uniquement sur une ontologie. Nous ne cherchons pas, comme présenté par Pivk et al. (2004), à découvrir des relations à partir de tableaux de données et d'outils linguis-