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-

tiques généraux tels que WordNet et GoogleSets, mais nous voulons au contraire reconnaître des relations prédéfinies dans une ontologie spécifique au domaine d'application.

Notre approche utilise les idées développées par Gagliardi et al. (2005) concernant l'annotation de tableaux guidée par une ontologie, sur la base d'égalités de mots entre les termes de l'ontologie et ceux du Web. Cependant, nous allons plus loin dans le sens où nous distinguons deux méthodes de traitement selon que les données sont numériques ou symboliques, et que nous proposons une annotation floue pour les données symboliques.

La section 2 décrit l'ontologie, élément central de notre système. Nous présentons ensuite notre système d'annotation dans l'ordre d'application des différentes étapes : distinction entre données numériques et symboliques en section 3, annotation des données symboliques en section 4 et annotation des données numériques en section 5. Chaque étape de ce travail est validée expérimentalement.

2 L'ontologie dans le système MIEL++

Le domaine d'application de notre entrepôt de données est défini dans une ontologie, et tout notre système est guidé par cette ontologie : pour changer de domaine d'application, il suffit de changer d'ontologie.

Le travail présenté ici est appliqué au domaine de la microbiologie alimentaire, et l'entrepôt de données construit s'intègre dans un système existant appelé MIEL[1] (Buche et al., 2005). Dans le système MIEL, les données de microbiologie alimentaire sont entrées manuellement dans une base de données relationnelle, et les utilisateurs interrogent cette base via une interface de requêtes où ils sélectionnent dans l'ontologie les microorganismes, produits alimentaires et facteurs expérimentaux qui les intéressent, avec la possibilité de définir des préférences : la base de données est interrogée selon ces critères, qui sont cependant élargis pour ramener plus de données, et les résultats sont ordonnés suivant leur proximité avec la requête de l'utilisateur. Dans le système MIEL++ (MIEL élargi à l'entrepôt de données XML, voir Buche et al., 2006), on souhaite conserver le même mode d'interrogation, la base de données et l'entrepôt étant interrogés simultanément, de façon transparente pour l'utilisateur. Pour cela, les données de l'entrepôt XML doivent être annotées avec la même ontologie que celle déjà utilisée dans le système MIEL.

L'ontologie utilisée dans le système MIEL++ décrit les relations sémantiques intéressantes pour le domaine de la microbiologie alimentaire, et les types de données impliqués dans ces relations. Par exemple, la relation *Growth kinetics* est composée des types *Food product*, *Microorganism*, *Temperature*, *Time*, *Colony count*. Les types sont décrits dans l'ontologie de deux manières, suivant qu'ils sont symboliques (*Food product*, *Microorganism*) ou numériques (*Temperature*, *Time*, *Colony count*). Les types symboliques sont décrits par une taxonomie des valeurs possibles (par exemple, taxonomie des microorganismes). Les valeurs qui peuvent être prises par un type symbolique sont appelées termes. Les types numériques sont décrits par les unités dans lesquelles on peut les exprimer (par exemple, °C ou °F pour *Temperature*) ainsi que, le cas échéant, par un intervalle de valeurs possibles (par exemple un pH est compris entre 0 et 14). Cette ontologie, construite manuellement lors de la création du système MIEL, est représentée dans un format spécifique. Nous étudions la possibilité de la représenter dans un des langages standards du web sémantique.

[1] Moteur d'Interrogation ELargie

3 Distinction entre colonnes numériques et symboliques

On suppose dans cette section qu'un prétraitement permet de mettre les tableaux issus du Web dans un format standard, avec des entêtes de colonnes, puis des lignes composées d'un ensemble de cellules : chaque ligne est une instance de la relation sémantique présentée par le tableau. Notre objectif est de reconnaître le type des colonnes du tableau, pour en déduire la relation sémantique représentée par le tableau. Un traitement différent est appliqué suivant qu'une colonne contient des données numériques ou symboliques : notre premier travail est donc de faire la distinction entre les colonnes numériques et symboliques.

3.1 Classification par règles des colonnes numériques et symboliques

Faire la différence entre des colonnes numériques et symboliques dans un tableau n'est pas si simple qu'il y paraît, surtout dans le domaine de la microbiologie alimentaire où de nombreuses données symboliques comportent des chiffres (par exemple la souche de microorganisme "E. coli O 157 : H7") alors que les données numériques comportent souvent des mots (unités, précision d'un intervalle de confiance...). La solution que nous proposons pour distinguer les colonnes numériques des colonnes symboliques tient compte des unités définies dans l'ontologie pour les types numériques.

Tout d'abord, les nombres sont reconnus selon l'expression régulière `(digit)+((',' | '.')(digit)+)*`, avec `digit` correspondant à l'un des dix chiffres `('0' | '1' | ... | '9')`. Les nombres en notation scientifique sont reconnus suivant l'expression régulière `digit '.' (digit)+ 'x 10 ' (digit)+`. On recherche également dans le tableau les occurrences des unités définies dans l'ontologie, et des *indicateurs de résultat absent*, qui sont des chaînes de caractères prédéfinies (par exemple, "No result", ou "NS" pour *Not Specified*). Chaque cellule d'une colonne est analysée pour compter le nombre d'occurrences des indicateurs suivants :

- *numérique sûr* : un nombre en notation scientifique, ou un nombre immédiatement suivi d'une unité ;
- *indice de numérique* : une unité seule ou un nombre seul ;
- *indice de symbolique* : tout mot qui n'est ni une unité ni un *indicateur de résultat absent*.

On applique ensuite sur chaque cellule la classification suivante :

- *numérique*, si la cellule contient au moins un *numérique sûr* ou si elle contient plus d'*indices de numérique* que d'*indices de symbolique* ;
- *symbolique*, s'il y a plus d'*indices de symbolique* que d'*indices de numérique* ;
- *type inconnu*, si la cellule ne contient pas d'indices ou s'il y a autant d'*indices de symbolique* que d'*indices de numérique*.

Une colonne est classifiée comme numérique si elle contient autant ou plus de cellules classifiées numériques que de cellules classifiées symboliques. Sinon la colonne est classifiée comme symbolique.

3.2 Résultats expérimentaux

Notre méthode de classification a été expérimentée sur 60 tableaux intéressants pour le domaine de la microbiologie alimentaire. Une classe *symbolique* ou *numérique* a été manuellement assignée à chacune des colonnes de ces tableaux, résultant en 264 colonnes numériques

et 85 colonnes symboliques. Les résultats de notre classification ont été comparés à ceux d'un classifieur « naïf », dans lequel toute cellule contenant un chiffre est considérée comme numérique et une colonne est considérée comme numérique si plus de la moitié de ses cellules sont numériques. Les résultats de cette classification sont donnés dans le tableau 1.

	classification utilisant les unités		classification naïve	
prédit / manuel	numérique	symbolique	numérique	symbolique
numérique	263	1	228	36
symbolique	5	80	13	72

TAB. 1 – *Résultats de classification numérique/symbolique sur 349 colonnes.*

La précision globale de la classification (proportion de colonnes bien classifiées par rapport au nombre total de colonnes classifiées) atteint 98% pour la classification utilisant les unités définies dans l'ontologie, contre 86% pour le classifieur naïf. On voit ici à quel point il est intéressant d'utiliser l'ontologie dès ce stade de l'annotation.

4 Annotation des données symboliques

Lorsqu'on a affaire à une colonne de données symboliques, on cherche d'une part à annoter le contenu de chaque cellule avec les termes de l'ontologie, et d'autre part à reconnaître le type de la colonne. La première étape consiste en l'annotation du contenu des cellules par les termes de l'ontologie qui en sont lexicalement les plus proches, comme présenté en section 4.1. En deuxième étape, les résultats de cette annotation sont utilisés pour déduire le type de la colonne, comme présenté en section 4.2. Enfin, l'annotation des cellules obtenue en première étape est modifiée afin de ne conserver que les termes correspondant bien au type trouvé pour la colonne.

4.1 Annotation des cellules au sein d'une colonne de valeurs symboliques

Dans notre système, le contenu d'une cellule symbolique, ci-après appelé « terme du Web », est annoté non pas uniquement avec un terme de l'ontologie, mais avec plusieurs termes possibles. Contrairement aux travaux de Gagliardi et al. (2005), les différents termes de l'ontologie proposés pour l'annotation n'ont pas tous la même importance, mais sont ordonnés selon leur similarité avec le terme du Web. Nous utilisons pour représenter notre annotation le modèle des sous-ensembles flous.

4.1.1 Les sous-ensembles flous

Le système MIEL, que nous souhaitons étendre pour l'interrogation des données annotées, utilise en effet le formalisme des sous-ensembles flous (Zadeh, 1965, 1978) pour l'expression des requêtes. Nous utilisons ce même formalisme pour représenter nos annotations.

La notion de sous-ensemble flou est un assouplissement de la notion de sous-ensemble classique d'un ensemble de référence X. Dans le cas classique, les éléments de X qui possèdent une certaine propriété constituent un sous-ensemble A de X, les éléments de X qui ne possèdent pas cette propriété appartiennent au complémentaire de A dans X. Dans le cas d'un

sous-ensemble flou, les éléments peuvent appartenir partiellement à un sous-ensemble, avec un degré d'appartenance compris entre 0 (élément n'appartenant pas au sous-ensemble) et 1 (élément appartenant totalement au sous-ensemble).

Définition. Un *sous-ensemble flou* A d'un ensemble de référence X est défini par une fonction d'appartenance μ_A de X dans $[0, 1]$ qui associe à chaque élément x de X le degré $\mu_A(x)$ avec lequel x appartient à A.

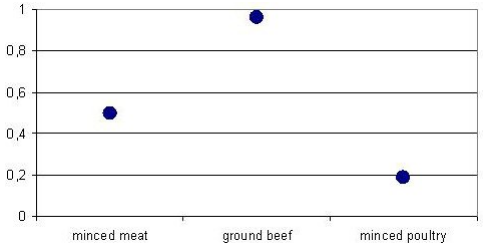

FIG. 1 – *Exemple de sous-ensemble flou sur un ensemble de définition à valeurs symboliques.*

Dans notre système d'annotation, nous utilisons les sous-ensembles flous pour décrire la similarité d'un terme du Web avec différents termes de l'ontologie. L'ensemble de définition du sous-ensemble flou est l'ensemble de tous les termes de l'ontologie, la fonction d'appartenance est une mesure de similarité entre le terme du Web et chacun des termes de l'ontologie. Par exemple, la figure 1 représente l'annotation du terme "minced beef" trouvé sur le Web. Ce terme n'existe pas tel quel dans l'ontologie, mais est similaire à divers termes de l'ontologie : "ground beef", et dans une moindre mesure "minced meat" ou "minced poultry". Les termes de l'ontologie dont le degré de similarité avec "minced beef" est nul ne sont pas représentés.

4.1.2 Degré de similarité d'un terme du Web avec un terme de l'ontologie

Il nous faut maintenant définir quelle mesure de similarité nous utilisons comme fonction d'appartenance pour nos sous-ensembles flous servant à l'annotation des termes du Web. Différentes mesures de similarité sémantique entre deux termes ont été présentées par Lin (1998); Resnik (1999); Seco et al. (2004) : ces mesures ont en commun qu'elles nécessitent l'utilisation d'une ontologie tierce comprenant les deux termes à comparer. Dans notre cas, une telle ontologie n'existe pas. Des essais que nous avons menés avec WordNet ont montré que cette ontologie était trop généraliste pour comprendre tous les noms d'aliments trouvés dans les publications issues du Web ; le thésaurus AgroVoc, utilisé par la FAO[2] et spécialisé dans l'agriculture et l'agro-alimentaire, ne contient pas lui non plus les noms d'aliments répertoriés dans les publications scientifiques en microbiologie alimentaire que nous cherchons à exploiter. Une mesure de similarité lexicale fondée sur les n-grammes est présentée par Lin (1998), mais cette mesure cherche à retrouver des mots de même racine plutôt que de même signification.

Nous proposons une mesure de similarité lexicale entre deux termes, fondée sur des égalités de mots. Chaque terme de l'ontologie, ainsi que chaque terme du Web, est décomposé en un ensemble de mots, qui sont lemmatisés (par exemple, "carrot cuts" et "cut carrots" donnent

[2]Food and Agriculture Organisation

tous deux le même ensemble $\{carrot, cut\}$). Tous les mots d'un terme n'ont pas la même importance dans la signification du terme, et ce de façon différente suivant le domaine d'application : dans le terme "minced poultry", c'est "minced" qui est le plus important si on se concentre sur les procédés alimentaires, tandis que c'est "poultry" si on s'intéresse plutôt à l'origine des produits. La distinction entre mots importants et moins importants est donc une affaire d'experts, qui ajoute de la connaissance sur un domaine. Dans l'ontologie, chaque mot de chaque terme se voit attribuer manuellement un poids entre 0 et 1 correspondant à son importance dans la signification du terme. Pour simplifier le travail d'attribution de poids dans les termes de l'ontologie, on conserve trois niveaux de poids :

- poids de 0 pour les mots n'apportant pas de sens, tels qu'articles et conjonctions, listés dans une *stop-list* ;
- poids de 1 pour les mots d'importance majeure pour la signification du terme ;
- poids de $0,2$ pour les mots d'importance moyenne (la valeur $0,2$ a été définie lors d'expériences préliminaires qui ont montré une annotation de meilleure qualité avec ce poids qu'avec le poids intermédiaire de $0,5$).

Comme on ne dispose pas de connaissances d'expert pour les termes du Web, tous les mots des termes du Web se voient attribuer un poids de 1 (sauf les mots de la *stop-list* qui gardent un poids de 0).

Lors de l'annotation d'un terme du Web, tous les termes de l'ontologie et celui du Web sont représentés comme des vecteurs, dont les coordonnées représentent l'ensemble de tous les mots lemmatisés possibles (i.e. tous les mots présents dans l'ontologie et les mots du terme du Web), les valeurs de ces coordonnées correspondant au poids du mot dans le terme, ou 0 si le mot n'est pas présent dans le terme. Un exemple de représentation vectorielle de termes est donné dans le tableau 2. Dans cet exemple, nous ne montrons que les termes de l'ontologie ayant au moins un mot en commun avec le terme du Web.

termes \ coordonnées		mince	beef	meat	ground	poultry
terme du Web	minced beef	1	1	0	0	0
termes de l'ontologie	minced meat	1	0	1	0	0
	ground beef	0	1	0	0,2	0
	minced poultry	0,2	0	0	0	1

TAB. 2 – *Représentation vectorielle de termes du Web et de l'ontologie.*

Une fois les termes représentés en tant que vecteurs, on définit la similarité entre un terme du Web et un terme de l'ontologie comme la mesure de similarité par cosinus (Van Rijsbergen, 1979) entre les deux vecteurs. Cette mesure a été choisie car c'est l'une des plus répandues pour la comparaison de vecteurs pondérés, et qu'une comparaison expérimentale avec d'autres mesures ne nous a pas apporté de meilleurs résulats.

Définition. La similarité entre un terme w du Web et un terme o de l'ontologie, représentés comme des vecteurs pondérés $\vec{w} = \{w_1, ..., w_n\}$ et $\vec{o} = \{o_1, ..., o_n\}$ est définie par la formule suivante :

$$sim(w, o) = \frac{\sum_{i=1}^{n} w_i o_i}{\sqrt{\sum_{i=1}^{n} w_i^2 \times \sum_{i=1}^{n} o_i^2}} \quad (1)$$

Exemple. Selon les poids donnés dans le tableau 2, on calcule les degrés de similarité utilisés dans la figure 1 :
$sim(minced\ beef, minced\ meat) = \frac{1+0+0+0+0}{\sqrt{(1^2+1^2)\times(1^2+1^2)}} = 0,5$
$sim(minced\ beef, ground\ beef) = \frac{0+1+0+0+0}{\sqrt{(1^2+1^2)\times(1^2+0,2^2)}} = 0,96$
$sim(minced\ beef, minced\ poultry) = \frac{0,2+0+0+0+0}{\sqrt{(1^2+1^2)\times(0,2^2+1^2)}} = 0,19$

4.1.3 Résultats expérimentaux

La validation de notre méthode d'annotation floue des termes du Web a été faite sur la partie aliments de l'ontologie : 185 termes distincts ont été manuellement annotés par leur terme le plus proche (ci-après appelé *best match*) dans la taxonomie des aliments. Pour valider la généralité de notre approche, nous avons également fait des tests d'annotation avec la taxonomie du Codex Alimentarius (taxonomie d'aliments utilisée par l'OMS[3]). Les deux taxonomies ont été retravaillées pour donner des poids aux mots de tous les termes. Chaque terme du Web a été annoté selon la méthode présentée en section 4.1.2 dans chacune des deux taxonomies, une fois avec les poids des mots définis manuellement, une fois avec des poids de mots de 1, comme si tous les mots avaient la même importance (mis à part les mots de la *stop-list* qui conservent un poids de 0). Les termes de la taxonomie proposés pour l'annotation d'un terme du Web sont ordonnés selon leur degré de similarité avec le terme du Web, et l'on regarde en quelle position se trouve le *best match*. Cette position est évaluée « au pire », c'est à dire que s'il y a plusieurs termes ayant le même score de similarité avec le terme du Web, le *best match* est considéré comme étant en dernière position. Cette méthode d'évaluation est due à notre souhait de proposer une annotation semi-automatique, où une liste des n meilleurs termes sera proposée à un utilisateur pour l'annotation du terme du Web : si plusieurs termes ont le même score, on ne maîtrise pas si le « bon » sera affiché dans la liste ou non. Les résultats de l'annotation sont présentés dans le tableau 3.

taxonomie	Codex Alimentarius		aliments dans MIEL++	
termes dont le *best match* a un score non nul	60%		78%	
position du *best match* / poids des mots	1	1 à 5	1	1 à 5
poids de 1 partout	30%	46%	46%	62%
poids définis manuellement	34%	52%	46%	66%

TAB. 3 – *Résultats de l'annotation de 185 noms d'aliments.*

Tout d'abord, on s'aperçoit que les résultats d'annotation sont meilleurs avec l'ontologie de MIEL++ qu'avec le Codex Alimentarius. Ceci s'explique par le fait que l'ontologie de MIEL++ a été construite spécialement pour le domaine de la microbiologie alimentaire, contenant notamment des noms d'aliments transformés, alors que le Codex Alimentarius a été construit à d'autres fins et est essentiellement tourné vers les matières premières (par exemple, on n'y trouve pas "butter" mais "cow milk fat", "goat milk fat",...). Cela plaide en faveur de l'utilisation d'ontologies de domaine vraiment adaptées au centre d'intérêt applicatif.

[3]Organisation Mondiale de la Santé

Pour qu'un terme de l'ontologie ait un score de similarité non nul avec un terme du Web, il faut et il suffit que ces termes aient un mot en commun : utiliser une méthode fondée sur l'égalité de mots n'est pas dénuée de sens, puisque 78% des *best match* ont effectivement un mot commun avec le terme du Web dans le cas de l'ontologie de MIEL++. L'utilisation de poids sur les mots dans les taxonomies ne modifie pas les termes qui seront retenus pour l'annotation, mais modifie l'ordre dans lequel ils sont présentés : on voit que l'utilisation de poids apporte une amélioration, faible mais systématique. En revanche, que l'on utilise ou non les poids sur les mots, l'utilisation d'une annotation floue avec termes ordonnés selon un degré de similarité est un gain important, puisqu'il y a en moyenne 16 termes de la taxonomie des aliments de MIEL++ ayant au moins un mot commun avec le terme du Web (avec un maximum à 94 termes pour le terme du Web "raw milk cheese") : une présentation non ordonnée de tous les termes possibles pour l'annotation est donc à proscrire. En utilisant le score de similarité, on arrive à obtenir 66% des *best match* dans les 5 premières positions pour l'ontologie de MIEL++, ce qui est bien plus intéressant dans le cadre d'une annotation semi-automatique, ou pour une interrogation où les résultats sont ordonnés selon leur similarité à la requête.

4.2 Détermination du type d'une colonne de valeurs symboliques

Une fois les termes des cellules d'une colonne annotés en utilisant le degré de similarité avec les termes de l'ontologie qu'on vient de présenter, on peut déterminer le type de la colonne symbolique.

4.2.1 Utilisation des degrés de similarité avec les termes de l'ontologie

On détermine le type de chaque terme de la colonne d'après ses scores de similarité avec les différents termes de l'ontologie. Les types de tous les termes d'une colonne sont ensuite utilisés pour déterminer le type de la colonne. Soit col une colonne d'un tableau, $type$ un type symbolique défini dans l'ontologie. Soit T_{type} l'ensemble de tous les termes de l'ontologie appartenant au type $type$. Soit t_{ext} le terme du Web contenu dans une cellule de la colonne col. Alors le score du type $type$ pour le terme t_{ext} est le suivant :

$$score(t_{ext}, type) = \sum_{t \in T_{type}} sim(t_{ext}, t) \tag{2}$$

Pour chaque terme de la colonne, on calcule le score de chaque type de l'ontologie. Soit $bestType(t_{ext})$ le type qui a le meilleur score pour le terme t_{ext}. Si ce score est supérieur à un seuil θ défini par l'utilisateur, alors le terme t_{ext} est considéré comme étant du type $bestType(t_{ext})$. Si par contre $score(t_{ext}, bestType(t_{ext})) < \theta$, alors t_{ext} est considéré comme de type inconnu.

Le score du type $type$ pour la colonne col est la proportion de termes de la colonne ayant le type $type$. Soit T_{col} l'ensemble de tous les termes de la colonne et T_{col}^{type} l'ensemble de tous les termes de la colonne ayant le type $type$, alors

$$score(col, type) = \frac{\mid T_{col}^{type} \mid}{\mid T_{col} \mid} \tag{3}$$

Considérons le type $bestType(col)$ qui a le meilleur score pour la colonne col. Si ce score est supérieur à un seuil α défini par l'utilisateur, avec $\alpha \in [0, 1]$, alors la colonne est classifiée comme étant du type $type$. Sinon le type de la colonne est considéré comme non reconnu.

Lorsque le type de la colonne est reconnu, on restreint le domaine de définition des sous-ensembles flous servant à l'annotation des termes au sein de la colonne : le nouveau domaine de définition est l'ensemble de tous les termes de l'ontologie correspondant au type de la colonne.

4.2.2 Résultats expérimentaux

Les 80 colonnes ayant bien été reconnues comme symboliques lors de la classification numérique/symbolique (section 3.2) ont été utilisées pour cette expérience. Les colonnes ont été manuellement classées en trois types : *aliment*(46 colonnes), *microorganisme*(16 colonnes) et *autre*(18 colonnes). Tous les termes de ces colonnes ont ensuite été automatiquement annotés avec les termes de l'ontologie correspondant aux types *aliment* et *microorganisme*. Les types des colonnes ont été calculés selon la méthode présentée ci-dessus, avec pour paramètres $\theta = 0,2$ et $\alpha = 0,5$, les colonnes de type non reconnu étant classées comme de type *autre*.

La qualité de cette classification a été comparée avec une classification automatique par apprentissage : on a utilisé la méthode de classification SMO, une optimisation des SVM (voir Platt, 1999), implémentée dans Weka[4], en conservant les paramètres par défaut, les colonnes étant transformées en vecteurs pondérés de tous les mots qu'elles contiennent. La classification par SMO a été évaluée en validation croisée par *leave one out* (chaque colonne est classifiée en utilisant un classifieur entraîné avec l'ensemble des 79 autres colonnes). Les résultats obtenus sont présentés dans le tableau 4.

manuel \ prédit	utilisation de l'ontologie			SMO		
	aliment	micro.	autre	aliment	micro.	autre
aliment	34	0	12	46	0	0
microorganisme	1	12	3	5	11	0
autre	1	0	17	6	0	12

TAB. 4 – *Résultats de classification sur 80 colonnes symboliques.*

Avec notre méthode sans apprentissage utilisant l'ontologie, on obtient une précision de 94% et une couverture de 74% pour les aliments : la classification par apprentissage suivant la méthode SMO donne certes une couverture de 100%, mais avec une précision plus faible à 81%. Pour les microorganismes, notre méthode donne une précision de 100% et une couverture de 75%, alors que SMO permet aussi une précision de 100% mais avec une couverture plus basse (69%). Notre méthode donne donc des résultats tout à fait comparables (voire meilleurs dans le cadre de notre application où l'on cherche avant tout une bonne précision) aux méthodes classiques de classification par apprentissage. L'avantage de notre méthode est qu'elle ne nécessite pas de phase d'apprentissage, en utilisant une ontologie déjà existante. Nous avons également testé la sensibilité de notre méthode au choix des paramètres. La sensibilité pour θ est faible : on obtient les résultats présentés dans le tableau 4 pour tout θ entre $0,01$ et 0.4 ; cependant pour des valeurs de θ plus élevées, on perd en couverture plus vite qu'on ne gagne en précision. On atteint une précision de 100% pour les aliments et microorganismes avec $\theta = 1$: on a alors une couverture de 65% pour les aliments et 69% pour les microorganismes. Notre méthode est un peu plus sensible pour le choix du paramètre α : plus α est grand, plus

[4]http ://www.cs.waikato.ac.nz/ml/weka

la précision est grande, avec une moindre couverture. Cependant les variations ne sont que de quelques points par tranche de $0,1$ ajoutée ou enlevée à α.

5 Annotation des données numériques

De même que nous avons recherché le type des colonnes symboliques, nous recherchons le type de colonnes numériques afin de pouvoir ultérieurement déterminer la signature de la relation représentée dans chaque tableau.

5.1 Reconnaissance du type d'une colonne numérique

Afin de déterminer le type d'une colonne numérique, on combine deux scores : le score de similarité du titre de la colonne avec les noms des différents types numériques, et un score déduit des unités utilisées dans la colonne.

On considère tout d'abord le titre de la colonne : on ne conserve que les mots qui ne correspondent ni à une unité, ni à un mot «sans intérêt» de la *stop-list* et on leur attribue un poids de 1. On calcule ensuite le score de similarité entre le titre de la colonne et chacun des types numériques de l'ontologie, selon la formule du score de similarité donnée en section 4.1.2, avec comme terme du Web le titre de la colonne et comme terme de l'ontologie le nom du type numérique dans l'ontologie (par exemple, "Samples tested" pour le nombre d'échantillons sur lesquels l'expérience porte). Soit t_{titre} le titre de la colonne col et t_{type} le nom du type $type$, alors le score de similarité du titre de la colonne col avec le type $type$ est :

$$score_{titre}(col, type) = sim(t_{titre}, t_{type}) \tag{4}$$

Examinons maintenant les unités utilisées dans la colonne. Soit u une unité et T_u l'ensemble de tous les types numériques pouvant s'exprimer dans cette unité. Le score du type $type$ pour l'unité u est $score(u, type) = \frac{1}{T_u}$ si u est une unité valable pour $type$, et $score(u, type) = 0$ si le type $type$ ne s'exprime pas dans l'unité u. Soit U_{col} l'ensemble de toutes les unités présentes dans la colonne : on considère également les unités présentes dans le titre de la colonne à condition qu'elles ne fassent pas partie d'un couple nombre-unité, qui représente généralement une précision de condition expérimentale (par exemple "at 37°C"). Le score sur les unités de la colonne col avec le type $type$ est :

$$score_{unit}(col, type) = max_{u \in U_{col}}(score(u, type)) \tag{5}$$

Ainsi le score de la colonne col avec le type $type$ est :

– si toutes les valeurs numériques contenues dans les cellules de la colonne sont compatibles avec l'intervalle de valeurs possibles associé au type $type$, alors

$$score_{final}(col, type) = 1 - (1 - score_{titre}(col, type)) \times (1 - score_{unit}(col, type)) \tag{6}$$

Ce score est inspiré de Yangarber et al. (2002), où une mesure similaire est utilisée pour combiner les confiances que l'on a en différentes règles de reconnaissance d'une entité nommée. Les deux scores se renforcent ainsi mutuellement, mais il suffit que l'un des deux scores soit bon pour que le score final soit bon.

– s'il existe une valeur dans les cellules de la colonne qui est en dehors de l'intervalle de valeurs défini dans l'ontologie, alors $score_{final}(col, type) = 0$

Le type retenu pour la colonne est celui qui a le meilleur $score_{final}$. Si tous les scores sont nuls, le type de la colonne est considéré comme non reconnu.

5.2 Résultats expérimentaux

Les 263 colonnes numériques reconnues lors de la classification numérique/symbolique (section 3.2) ont été utilisées pour la validation de notre approche. Ces colonnes ont été manuellement classées suivant 19 types numériques définis dans l'ontologie. Les résultats de notre classification utilisant l'ontologie ont été comparés à ceux de la méthode de classification par apprentissage SMO, les colonnes étant représentées par des vecteurs pondérés de tous les mots contenus dans les cellules et le titre de colonne, toutes les valeurs numériques étant remplacées par le mot-clef #NUM. La classification par SMO a été évaluée en validation croisée par *leave one out*.

Avec notre méthode de classification sans apprentissage utilisant l'ontologie, on obtient une précision globale de 96% et une couverture de 95%, sur l'ensemble des 19 types. La méthode SMO classe toutes les instances, avec une précision globale et une couverture globale de 96%. Notre méthode de classification, qui ne nécessite pas de données d'entraînement, donne donc des résultats de classification tout à fait comparables à une méthode classique de classification par apprentissage, plus gourmande en temps d'expert si l'on part du postulat que l'ontologie existe de toute manière (ce qui est le cas puisque l'objet de la classification est de reconnaître les types définis dans l'ontologie).

6 Conclusion et perspectives

Nous avons présenté une méthode d'annotation de tableaux de données guidée par une ontologie, sans phase d'apprentissage. Nous distinguons tout d'abord les données numériques et symboliques, pour les traiter différemment. Les données symboliques sont annotées avec les termes de l'ontologie, et ces annotations permettent de déduire le type de chaque colonne symbolique. Pour les colonnes numériques, on utilise à la fois le titre de la colonne et les valeurs et unités contenues dans la colonne pour déterminer le type de la colonne : là encore nous utilisons l'ontologie, dans laquelle sont définis les unités et intervalles de valeur valables pour chaque type numérique. Notre approche donne des résultats comparables à une méthode classique de classification par apprentissage, mais sans nécessiter la construction d'un jeu d'entraînement.

A partir des types de colonnes ainsi identifiés, il nous reste maintenant à reconnaître les relations représentées par le tableau de données. Ensuite nous travaillerons sur les techniques d'interrogation de l'entrepôt de données, en tenant compte du fait que les critères d'interrogation permettent d'exprimer des préférences, que l'annotation des données symboliques est floue et que le type de certaines colonnes n'est pas reconnu.

Références

Baumgartner, R., S. Flesca, et G. Gottlob (2001). Visual web information extraction with Lixto. In *VLDB '01*, pp. 119–128.

Buche, P., C. Dervin, O. Haemmerlé, et R. Thomopoulos (2005). Fuzzy querying of incomplete, imprecise, and heterogeneously structured data in the relational model using ontologies and rules. *IEEE T. Fuzzy Systems 13*(3), 373–383.

Buche, P., J. Dibie-Barthélemy, O. Haemmerlé, et G. Hignette (2006). Fuzzy semantic tagging and flexible querying of xml documents extracted from the web. *Journal of Intelligent Information Systems 26*(1), 25–40.

Freitag, D. et N. Kushmerick (2000). Boosted wrapper induction. In *Proceedings of AAAI-2000*, pp. 577–583.

Gagliardi, H., O. Haemmerlé, N. Pernelle, et F. Saïs (2005). An automatic ontology-based approach to enrich tables semantically. In *AAAI Context and Ontologies Workshop*.

Lin, D. (1998). An information-theoretic definition of similarity. In *ICML '98*, San Francisco, CA, USA, pp. 296–304.

Pivk, A., P. Cimiano, et Y. Sure (2004). From tables to frames. In *ISWC 2004*, Hiroshima, Japan, pp. 116–181.

Platt, J. C. (1999). *Fast training of support vector machines using sequential minimal optimization*, pp. 185–208. Cambridge, MA, USA : MIT Press.

Resnik, P. (1999). Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research 11*, 95–130.

Seco, N., T. Veale, et J. Hayes (2004). An intrinsic information content metric for semantic similarity in wordnet. In *ECAI 2004*, Valencia, Spain.

Van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow.

Yangarber, R., W. Lin, et R. Grishman (2002). Unsupervised learning of generalized names. In *19th International Conference on Computational Linguistics*, Taipei, Taiwan, pp. 1–7.

Zadeh, L. (1965). Fuzzy sets. *Information and control 8*, 338–353.

Zadeh, L. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems 1*, 3–28.

Summary

This paper deals with automatic annotation of data tables on a given application domain, represented by an ontology. First, we discriminate between numeric and symbolic data. Then data are further processed: symbolic data are annotated with terms of the domain ontology, and both symbolic and numeric columns of the table are annotated with the type of data they contain. The annotation of symbolic data is fuzzy, as they are annotated with several terms of the ontology that can be ordered according to a similarity measure. Each step of our approach is experimentally validated.