Découverte de relations candidates à l'enrichissement d'un entrepôt thématique

Hélène Gagliardi, Ollivier Haemmerlé, Nathalie Pernelle, Fatiha Saïs

LRI (UMR CNRS 8623 - Université Paris-Sud) / INRIA (Futurs)
Bâtiment 490, F-91405 Orsay Cedex, France
{gag,pernelle,sais}@lri.fr ; ollivier.haemmerle@univ-tlse2.fr

Abstract. Ce travail a pour objectif d'enrichir automatiquement un entrepôt thématique de données à partir de documents de format divers provenant du Web et comportant des tableaux. Cet enrichissement est guidé par une ontologie comportant un ensemble de termes et de relations. Le caractère automatique de notre approche nous a conduits à utiliser un format flexible permettant de représenter les données (termes ou relations) dont l'identification est partielle, incertaine ou multiple. Cet article se focalise sur la découverte dans les tableaux de relations non identifiées ainsi que sur la découverte d'attributs qui peuvent enrichir ou modifier l'interprétation de relations identifiées.

Mots clés: extraction de connaissances, entrepôt, ontologie, XML, Web.

1 Introduction

Notre travail concerne la construction automatique d'entrepôts thématiques. Il s'agit plus précisément d'enrichir un entrepôt constitué de différentes bases de données à l'aide de données trouvées sur le Web.

Ce travail a été réalisé dans le cadre du projet e.dot [3] (Entrepôt de Données Ouvert sur la Toile[1]). Le domaine d'application choisi est le risque microbiologique dans les aliments, domaine qui présente des enjeux majeurs. En effet, les lois sur la sécurité alimentaire étant de plus en plus strictes, la quantité d'analyses effectuées tous les jours par l'industrie alimentaire entraîne des dépenses importantes qui peuvent être limitées si des outils sont développés pour mieux connaître le comportement des microorganismes dans les aliments. Il sera alors possible de prévenir le risque de contamination au lieu de constater les épisodes de crises. C'est l'objectif du projet Sym'Previus, lancé par des institutions gouvernementales, et, c'est dans le cadre de ce projet, que le système MIEL [2] a été développé. Ce système permet d'interroger deux bases de données locales contenant des résultats expérimentaux et industriels sur le comportement de germes pathogènes dans des aliments en fonction de différents facteurs tels

[1] partenaires : IASI-Gemo (LRI), Verso-Gemo (INRIA-Futurs), INAP-G/INRA et la société Xyleme

que le pH ou la température. Ces données sont incomplètes : il y a en effet une infinité d'expériences pouvant être menées. Il est donc important de pouvoir alimenter cet entrepôt avec des données trouvées sur le Web.

Nous nous sommes focalisés sur les données présentées sous forme de tableaux. En effet, il s'agit d'un mode de représentation habituel et synthétique de résultats expérimentaux. Ces données sont interrogées à l'aide d'une architecture médiateur fondée sur une ontologie du domaine, ontologie développée lors du projet Sym'Previus. Pour utiliser cette ontologie lors de l'interrogation, nous devons représenter les données à l'aide du vocabulaire présent dans l'ontologie. Plus précisément, notre approche permet de découvrir dans les tableaux des instances de relations dont la description est représentée dans l'ontologie sous forme d'un ensemble d'attributs. Nous avons défini une représentation XML - appelée SML (Semantic Markup Language) - permettant de stocker ces relations dans l'entrepôt avec un maximum de flexibilité [4]. En effet, la transformation des données du tableau en une représentation utilisant le vocabulaire (termes et relations) de l'ontologie est complètement automatique.

En particulier, et c'est sur ce point que nous nous focalisons dans ce papier, il nous semble utile de pouvoir enrichir l'entrepôt avec des instances de relations existantes complétées d'informations imparfaitement identifiées ou même de garder, en cas d'échec lors de l'identification des relations représentées dans le tableau, une relation dite générique qui permettra de conserver l'existence d'un lien (même si il n'est pas identifié) entre certaines données. Ces relations sont stockées et peuvent être exploitées lors de l'interrogation. Elles peuvent également servir de base à un enrichissement de l'ontologie si un expert les identifie par la suite.

Ce papier est structuré de la manière suivante. Nous présentons en section 2 le format d'entrée des tableaux ainsi que le format SML dans lequel ces tableaux sont transformés. En section 3 nous présentons l'alimentation de l'entrepôt avec des relations enrichies et des relations génériques candidates. Nous présentons ensuite comment ces relations peuvent être exploitées lors d'une interrogation, avant de terminer par une étude des premiers résultats obtenus.

2 Notions préliminaires

Dans le projet e.dot, les données sont acquises par un crawler combiné à un outil de filtrage qui ne sélectionne que les documents html ou pdf qui contiennent des tableaux présentant des mots clef de l'ontologie Sym'Previus [1]. Ces tableaux sont tout d'abord transformés dans un format XML utilisant des tags syntaxiques indépendants du domaine (format XTab [3]). Ils sont donc représentés par une liste de lignes, chacune comportant une liste de cellules (Fig. 1). L'ontologie, développée par des experts lors du projet Sym'Previus, concerne le domaine du risque microbiologique. Elle comporte une taxonomie de 428 termes ainsi que 25 relations sémantiques qui sont caractérisées par leur signature. Comme en algèbre relationnelle, une relation est décrite par un ensemble d'attributs. Ces derniers correspondent à des termes définis dans la taxonomie. Par exemple,

dans l'ontologie la relation *foodFactorMicroorganism* est décrite par une signature représentée par l'ensemble d'attributs (*food, factor, microorganism*).

Notre système utilise cette ontologie pour transformer le document XTab en document SML - Semantic Markup Language - où les lignes du tableau ne sont plus représentées par des cellules mais par un ensemble de relations.

Products	**pH values**
Cultivated mushroom	5.00
Crab	6.60

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<table> <title> <table-title>
approximative pH of some food products</table-title>
<column-title>Products</column-title>
<column-title>pH values</column-title> </title>
<nb-col>2</nb-col>
<content>
<line>
<cell>cultivated mushroom</cell>
<cell>5.00</cell>
</line>
<line>
<cell>crab</cell>
<cell>6.60</cell>
</line>
</content> </table>
```

Fig. 1. *Un tableau et sa représentation XTab*

```
<table> <title> <table-title>
approximative pH of some food products </table-title>
<column-title> Products </column-title>
<column-title>pH values</column-title>...
</title> <content>
<rowRel>
<foodPH>
<food> <ontoVal>mushroom</ontoVal>
<originalVal> cultivated mushroom </originalVal>
</food>
<ph> <ontoVal/>
<originalVal>5.00</originalVal> </ph> </foodPH>
</rowRel>
<rowRel>
<foodPH>
<food> <ontoVal>crab</ontoVal>
<originalVal>crab</originalVal> </food>
<ph> <ontoVal/> <originalVal>6.60</originalVal> </ph>
</foodPH> </rowRel>
</content> </table>
```

Fig. 2. *Une représentation SML simplifiée de la représentation XTab de la Figure 1*

Dans l'exemple des figures 1 et 2, la relation *foodPh* qui lie un aliment à sa valeur de pH a été reconnue. Elle est représentée et instanciée par les valeurs apparaissant dans les cellules du tableau présentes dans <originalVal>. L'utilisation d'opérations de mapping nous permet d'associer à chacune de ces

valeurs un ou plusieurs termes de l'ontologie et ces termes peuvent être utilisés lors de l'interrogation du document, interrogation guidée par l'ontologie. La valeur originale, lorsqu'elle diffère du terme de l'ontologie, peut alors être visualisée sur demande de l'utilisateur.

La reconnaissance des relations s'effectue en deux étapes : dans un premier temps le système associe, lorsque cela est possible, un terme de l'ontologie à chaque colonne du tableau, ce qui permet d'obtenir le schéma du tableau. Ensuite, le système représente l'ensemble des relations sémantiques dont la signature est compatible avec un sous-ensemble du schéma du tableau. La transformation étant complètement automatique, le système est également flexible lors du processus de reconnaissance de relations de l'ontologie dans le tableau.

Il est en effet possible de reconnaître des relations même si tous leurs attributs ne sont pas identifiés dans le tableau ; tel est le cas si, par exemple, l'un des attributs est une constante présente dans le texte qui environne le tableau. Il s'agit alors de relations dites partielles. Nous avons pu montrer dans [4] que la possibilité de reconnaître des relations partielles permettait d'augmenter significativement le rappel sans trop faire chuter la précision.

3 Découverte de relations candidates

La représentation de relations partielles permet d'alimenter l'entrepôt avec des informations qui semblent être incomplètes. Il s'agit maintenant de savoir ce que le système doit faire des informations contenues dans les tableaux et qu'il n'a pas pu représenter dans une instance de relation de l'ontologie. Nous avons distingué deux cas de figures : celui où nous utilisons les informations pour compléter une relation reconnue et celui où aucune relation n'a pu être identifiée dans le tableau.

3.1 Capture de l'information non reconnue dans des attributs supplémentaires

A l'étape de l'identification des colonnes du tableau, il arrive souvent que certaines colonnes ne soient associées à aucun attribut ou qu'elles soient identifiées mais ne soient associées à aucune relation reconnue dans le tableau. Nous considérons alors que ces attributs sont des informations additionnelles qui peuvent venir préciser ou modifier l'interprétation d'une relation reconnue. Les informations de ces colonnes sont alors représentées dans le document SML par des attributs supplémentaires à l'aide du tag générique <attribut>. Ce dernier est ajouté à toute relation reconnue dans le tableau.

Ces informations additionnelles peuvent jouer plusieurs rôles :

- recueillir un attribut non identifié dans une relation partielle (ex : attribut numérique) ;
- recueillir un argument supplémentaire de la relation qui n'est pas défini dans l'ontologie (ex : forme ou quantité du produit) afin de préciser l'interprétation de la relation ;

– recueillir une relation supplémentaire qui partage des attributs avec la relation reconnue (ex : le tableau comportant les colonnes *food*, *lipid*, *ph*, *microorganism*).

Le tableau de la figure 3 représente une étude de l'influence de trois facteurs expérimentaux sur la croissance de la *listeria* dans les *aliments fumés*. Dans cet exemple de tableau, le document SML correspondant comporte deux attributs supplémentaires permettant de représenter deux attributs *Factor* représentés dans les 4ème et 5ème colonnes. Le système n'a en effet pas reconnu que les colonnes ayant pour titre *growth rate* et *egrc at 5◦c (log10 cfu/day)* représentaient des attributs *Factor*.

microorganism	**food**	**temperature**	**growth rate**	**egrc at 5◦c (log10 cfu/day)**
listeria	coppa (smoked pork)	4◦c	2.1 logs in 28 days	0.107
listeria	cold-smoked salmon	8◦c	5.4 logs in 21 days	0.116

Fig. 3. *Microorganismes et aliments fumés*

Dans ce deuxième exemple de tableau de la figure 4, l'attribut additionnel, que l'on peut voir dans le document SML de la figure 5, permet de compléter la relation reconnue par une information qui n'a pas été prévue dans l'ontologie. Ainsi, dans le domaine du risque alimentaire, on trouve souvent des tableaux qui précisent des références d'articles pour chaque résultat présenté dans une ligne. Ce paramètre n'a pas été prévu dans la signature de la relation mais il est recueilli dans un attribut additionnel et peut être proposé par la suite à l'utilisateur.

microbe	**vehicle/source**	**reference**
Campylobacter spp.	poultry	Friedman et al 2000 [1]
Clostridium perfringens	meat	Todd 1997; Olsen, S. J. et al. 2000 [1]

Fig. 4. *Aliments et Microorganismes*

3.2 Représentation de tableaux par une relation générique

Nous allons maintenant montrer le rôle des relations génériques dans la représentation des informations des tableaux et leur intérêt lors de l'interrogation. Une relation générique permet de représenter les données d'un tableau dans lequel aucune relation de l'ontologie n'a été reconnue. L'échec de la reconnaissance des relations sémantiques est dû à plusieurs facteurs :

```
<table>
</table-title >
<column-title> microbe </column-title>
<column-title>vehicle/source</column-title>
<column-title>reference</column-title>
<column-nb> 3 </column-nb>
<content>
<rowRel additionalAttr="yes">
...
<foodMicroorganism relType ="completeRel">
<food>¡finalVal¿...</food>
<microorganism> <ontoVal>campylobacter</ontoVal>
<originalVal>campylobacter spp. </originalVal> </microorganism>
< attribute indMatch="attribut"> <ontoVal/> <originalVal> Friedman et al 2000
</originalVal>
</attribute>
</foodMicroorganism>...

</rowRel> ... </content> </table>
```

Fig. 5. *Représentation SML du tableau de la fig.4 - Attribut supplémentaire*

- les relations du domaine représentées dans le tableau ne sont pas décrites dans l'ontologie du domaine;
- des tableaux peuvent représenter des informations concernant deux thématiques à la fois. Par exemple, certains des tableaux sur l'épidémiologie représentent des informations concernant aussi le risque alimentaire. Il s'agit en fait d'une thématique connexe;
- des attributs de relations sémantiques non reconnus dans le tableau empêchent la reconnaissance des relations. C'est en particulier le cas pour des attributs qui représentent des valeurs numériques et pour lesquels le processus de reconnaissance se fonde entièrement sur le titre des colonnes. Dans ce cas, la relation générique permet de représenter les attributs reconnus et non reconnus.

Les relations génériques permettent de conserver ces informations. La figure 6 représente, par exemple, la dose à laquelle on peut dire qu'un germe a infecté un support. Il s'agit d'une relation factorMicroorganism mais la colonne numérique qui a pour titre *infective dose* et qui correspond au facteur n'a pu être reconnue. Ce tableau a pour représentation SML le document présenté en figure 7, document dans lequel les valeurs apparaissant dans les colonnes sont conservées et liées par une relation générique.

Pathogen	**Infective dose**	**reference**
Campylobacter spp.	500-800	Robinson 1981; Black et al 1988
Clostridium perfringens	10^7	Bermith 1988

Fig. 6. *Doses infectieuses*

```
<table>
</table-title >
<column-title> pathogen </column-title>
<column-title>infective dose</column-title>
<column-title>reference</column-title>
<column-nb> 3 </column-nb>
<content>
<rowRel additionalAttr="no">
...
<relation relType ="generic">
<attribute indMatch="microorganisme">¡finalVal¿...</attribute>
<attribute indMatch="attribut"> <ontoVal>campylobacter</ontoVal>
<originalVal>campylobacter spp. </originalVal> </microorganism>
< attribute indMatch="attribut"> </ontoVal> <originalVal> Robinson 1981; Black
et al 1988 </originalVal>
</attribute>
</relation>
...

</rowRel> ...
</content> </table>
```

Fig. 7. *Représentation SML du tableau de la fig.6 - Relation générique*

4 Interrogation des données imparfaitement identifiées

Dans le cadre du projet e.dot, le moteur d'interrogation MIEL++ est capable d'interroger les deux bases existantes, relationnelle et graphe conceptuel, en combinaison avec l'interrogation de la base SML. Il est donc en mesure de traduire une requête MIEL en une requête sur la base des documents SML. Le vocabulaire utilisé pour exprimer les requêtes et les relations sémantiques à interroger est représenté dans l'ontologie Sym'Previus.

Ainsi, l'interrogation de l'entrepôt SML et des bases pré-existantes se fait de manière uniforme et transparente pour l'utilisateur par le biais d'une même interface graphique qui permet à l'utilisateur de sélectionner dans l'ontologie ses attributs de projection et de sélection.

Une extension de ce moteur d'interrogation peut permettre d'interroger les données additionnelles non identifiées ou imparfaitement identifiées qui apparaissent en tant qu'attributs supplémentaires et relations génériques.

4.1 Exploitation des attributs supplémentaires

Les attributs supplémentaires peuvent être exploités pour toute requête dont les réponses proviennent de relations avec attributs supplémentaires. Pour permettre à l'utilisateur d'avoir plus d'informations concernant le contexte des réponses à sa requête, le contenu et les titres des colonnes correspondant aux attributs supplémentaires peuvent être affichés.

Voici un exemple de requête Q1 où l'utilisateur cherche les aliments pouvant véhiculer le microorganisme "*Campylobacter spp.*". La figure 8 représente le résultat de l'évaluation de la requête Q1 sur l'entrepôt de données SML con-

tenant des documents représentant des informations sur "*Campylobacter spp.*". Parmi ces documents se trouve le document SML de la figure 5.

Microorganisme	**Aliment**	**Attributs Supp.**
campylobacter	poultry	Référence: Friedman et al 2000
campylobacter	turkey	

Fig. 8. *La réponse à la requête Q1*

Cet exemple montre l'intérêt de présenter le contenu des attributs supplémentaires. Ici, la complétion de la réponse par la référence bibliographique intéressera vraisemblablement l'utilisateur. La présentation du titre de la colonne devant la valeur de cet attribut facilitera l'interprétation de cette nouvelle information.

4.2 Exploitation des relations génériques

L'exploitation des relations génériques par l'utilisateur consiste entre autres en l'interrogation de la base de données SML par des requêtes par mots clés. Une requête Q2 cherchant toute l'information que la base de données contient au sujet du microorganisme *"Campylobacter"* est un exemple d'une telle requête par mots clés. L'évaluation de cette requête contiendra toutes les relations identifiées et toutes les relations génériques dont une des valeurs du tag *ontoVal* est "*Campylobacter*" ou une spécialisation de "*Campylobacter*".

Attribut1	**Attribut2**	**Attribut3**
Pathogen: campylobacter	Infective dose: 500-800	reference: Robinson 1981; Black et al 1988
Organism : campylobacter	Temp: 5∘C	

Fig. 9. *Réponse à la requête Q2*

Dans la figure 9 est présenté un extrait de la réponse à la requête Q2 qui correspond à une ligne de la relation générique représentant le tableau de la figure 6. Si nous n'avions pas interrogé les relations génériques, nous aurions raté cette réponse et nous aurions peut-être même obtenu aucune réponse dans l'hypothèse où aucune relation de l'ontologie ne contient la valeur *"campylobacter"*.

5 Evaluation de l'approche

Nous avons évalué notre approche d'enrichissement sémantique sur 61 documents XTab représentant des tableaux de données réelles. Ces derniers ont été

extraits de documents collectés sur le Web. Un document XTab est sélectionné pour l'évaluation s'il contient au moins une valeur correspondant à un terme de l'ontologie Sym'Previus.

Dans la section 5.1 nous décrivons la chaîne de traitement des documents. Dans la section 5.2, nous présentons et interprétons les résultats de l'évaluation de la découverte de relations candidates.

5.1 Chaîne de traitement des documents évalués

L'application AQWEB est un service Web qui a été développé dans le but de faciliter la tâche de validation par un expert des différents modules développés dans le projet e.dot. Cette application permet de relier les différentes tâches intervenant dans la chaîne de traitement des documents, de leur collecte jusqu'à leur enrichissement sémantique.

Le service permettant de rechercher et de collecter des documents PDF [2] potentiels à partir du Web considère en entrée un tuple (aliment, microorganisme, expressions exactes, mots exclus, langue du document, table, références) et récupère des documents vérifiant la directive exprimée dans ce tuple. Les fichiers PDF sont convertis automatiquement en format Word avec l'outil Omnipage de Scansoft. Cette tâche est asynchrone. Les fichiers convertis en Word doivent être validés car il arrive parfois que les documents Word générés contiennent des erreurs, notamment dans les polices de caractères. La présentation des formats de tableaux peut être aussi simplifiée pour se ramener à un format de type XTAB (une seule ligne d'entête, sans groupement de cellules et plusieurs lignes de détail).

L'ensemble des tableaux de données se trouvant dans un fichier Word validé est converti automatiquement en un ensemble de documents XML au format XTAB. Ces documents XTab sont enfin enrichis sémantiquement par le module XTab2SML que nous avons développé.

5.2 Résultats de l'évaluation

Les résultats de l'évaluation de l'identification des relations sémantiques de l'ontologie dans les documents XTab ont montré que le rappel augmente significativement quand on conserve les relations partiellement identifiées. En effet, le rappel vaut 0,45 lorsque nous ne conservons que les relations parfaitement reconnues et vaut 0,65 lorsque nous gardons les relations parfaitement et partiellement reconnues. Ces résultats ont confirmé l'intérêt de cette première flexibilité dans la représentation des documents en SML.

Nous montrons dans ce qui suit les résultats de l'évaluation concernant la découverte de relations sémantiques candidates à l'enrichissement d'un entrepôt thématique. Ces relations candidates à l'enrichissement peuvent être représentées

[2] la majorité des publications scientifiques diffusées sur le Web sont dans le format PDF

par une relation comportant un ensemble d'attributs supplémentaires ou par une relation générique.

Relations comportant des attributs supplémentaires Ces relations comportent les attributs reconnus et intégrés dans des relations sémantiques de l'ontologie aussi que les attributs supplémentaires. Nous avons pu voir que 40% (soit 24/61) des tableaux testés contiennent des colonnes ne pouvant être associées à aucune relation sémantique identifiée ; ces colonnes sont donc représentées par des attributs supplémentaires. Ces derniers jouent essentiellement deux rôles : le premier consiste en la complétion d'une relation sémantique reconnue ce qui permet à l'utilisateur d'avoir une interprétation plus précise de la relation ; le second consiste en la représentation des données d'attributs non reconnus (oubliés) dans une relation déjà reconnue. Environ 67 % des attributs supplémentaires viennent compléter des relations sémantiques reconnues ou, par distribution des attributs, permettent de représenter de nouvelles relations. En effet, une distribution des attributs reconnus et des attributs supplémentaires peut représenter des relations sémantiques existantes ou nouvelles et candidates à l'enrichissement de l'entrepôt. Les résultats montrent que environ 33% (16/49) des attributs représentent des attributs oubliés dans des relations reconnues.

	Nombre d'attributs
Attributs supplémentaires	49
Capture des attributs oubliés	16
Complétion de relations	33

Fig. 10. *Résultat de l'évaluation de l'identification des attributs*

Une relation générique Elle permet de représenter les liens sémantiques existant entre les données d'un tableau, dans le cas où aucune des relations sémantiques de l'ontologie n'est reconnue. Une relation générique contient soit des attributs reconnus mais qui ne peuvent pas être combinés afin d'instancier une relation existante, soit des attributs représentant des colonnes non reconnues. La figure 11 présente les statistiques concernant l'identification de relations sémantiques de l'ontologie dans les documents XTab testés. Dans ces tableaux, 34 relations génériques ont été générées, 93 relations sémantiques ont été reconnues correctes et 48 relations sémantiques de l'ontologie ont été oubliées.

Parmi les relations génériques candidates à l'enrichissement, nous distinguons trois cas :

– des relations déjà existantes dans l'ontologie pour lesquelles nous n'avons pas pu reconnaître les attributs nécessaires à leur instantiation (au moins deux attributs). Il s'agit dans ce cas de la découverte d'instances de relations sémantiques existantes mais oubliées;

– des relations représentant des informations du domaine de l'ontologie (le risque microbiologique dans les aliments) et pour lesquelles la combinaison des attributs ne peut pas correspondre à une relation de l'ontologie. Il s'agit dans ce cas d'une relation du domaine, candidate à l'enrichissement non seulement de l'entrepôt mais également de l'ontologie du domaine ;
– des relations génériques représentant des données d'un domaine connexe. Dans ce cas, nous pouvons envisager de proposer aux experts du domaine ces relations découvertes, en vue d'étendre l'ontologie existante en une ontologie plus générale, par exemple, en transformant l'ontologie du risque microbiologique dans les aliments en une ontologie traitant également du domaine du risque chimique dans les aliments. En effet, dans notre ensemble de tests, nous avons traité des tableaux qui concernent le domaine du risque chimique et des tableaux qui concernent l'épidémiologie. Nous devons toutefois signaler que nous avons également traité 23% de tableaux qui n'étaient guère pertinents pour le domaine (comme par exemple les carences nutritionnelles du poisson chat).

	Nombre de relations
Relations trouvées	147
Relations trouvées correctes	93
Relations oubliées	48
Relations incorrectes	54
Relations génériques	34

Fig. 11. Résultat de l'évaluation de la découverte de relations sémantiques

Les résultats présentés dans figure 12 montrent que 20% (7/34) des relations génériques permettent de représenter des données de relations oubliées (parmi 48 relations oubliées). Ces résultats montrent également que 55% (19/34)des relations génériques représentent des relations pertinentes pour l'enrichissement de l'entrepôt et donc intéressantes, par la suite, pour l'interrogation par les utilisateurs. Le faible nombre de tableaux et donc de relations génériques qui ne présentent pas d'intérêt pour le domaine peut s'expliquer par la précision de la directive appliquée lors de la collecte des documents sur le Web. En effet, si les critères de sélection du document sont assez précis, les tableaux contenus dans le document ont de fortes chance d'être pertinents pour le domaine et donc de donner lieu à une relation générique intéressante.

6 Conclusion

Il existe une multitude de travaux dans le domaine de la découverte de connaissances dans les documents. Parmi ceux qui se sont intéressés à la découverte de

	Nombre de relations
Relations non pertinentes	8
Relations de l'ontologie oubliées	7
Relations pertinentes	19

Fig. 12. Intérêt des relations génériques

relations sémantiques, nous citons les systèmes MeatAnnot [5] et [8]. Nous citons également les approches proposées par [7] et [6].

MeatAnnot est un système d'extraction d'instances de concepts et de relations à partir de texte guidé par l'ontologie UMLS. MeatAnnot permet une génération semi-automatique d'annotations. Dans cette approche deux outils ont été utilisés : GATE pour l'étiquetage grammatical, Syntex pour l'extraction de syntagmes verbaux pouvant correspondre à des instances de relations. A partir de ces données, une grammaire d'extraction de relations est écrite manuellement. Le résultat de l'extraction est ensuite proposé à la validation de l'utilisateur et représenté en RDF.

Le système CAMELEON [8] permet d'extraire des instances de relations sémantiques entre concepts en utilisant des marqueurs sur un corpus d'apprentissage (exemple : la présence de "X être article-indéfini Y" pour la relation *est-un*).

Ces deux systèmes sont semi-automatiques. De plus, l'extraction de relations sémantiques est effectuée à partir de textes assez riches pour apprendre les règles d'extraction. Ce type d'approche ne peut être appliqué dans notre cas puisque le seul contexte dont nous disposons est celui du tableau (titres et contenu des lignes). Le travail de [7] propose une approche permettant de générer automatiquement un ensemble de méthodes (en utilisant une Frame-logique) permettant d'extraire une donnée d'un tableau. Chaque méthode est représentée par un ensemble de paramètres et par un type de retour et permet d'extraire des instances de relations sémantiques.

Dans le travail proposé par [6], l'ontologie comporte un ensemble de termes du domaine, leurs caractéristiques lexicales, leur contexte d'apparition à l'aide de mots clés et de relations sémantiques entre les termes. A partir de l'ontologie, l'outil peut concevoir et alimenter automatiquement une base de données relationnelle par des informations reconnues et extraites à partir des documents donnés en entrée. L'ontologie est construite manuellement par un expert du domaine visé. Cette approche suppose une importante homogénéité des documents traités.

Nous avons proposé dans ce travail une approche complètement automatique de découverte de relations sémantiques candidates à l'enrichissement d'un entrepôt thématique. Ces relations peuvent être utilisées, dans un premier temps, pour enrichir les réponses de l'utilisateur. Il peut s'agir soit d'informations qui viennent s'ajouter à une relation connue, soit de liens sémantiques potentiels entre valeurs présentes dans une même ligne de tableau. Ces relations peuvent

être, par la suite, proposées à un expert du domaine pour enrichir l'ontologie du domaine ou même pour l'étendre à une ontologie plus générale. En l'occurrence, il s'agirait ici d'étendre l'ontologie du risque microbiologique dans les aliments à une ontologie plus générale qui serait l'ontologie du risque alimentaire. Cette dernière représenterait les connaissances du domaine du risque microbiologique, du domaine du risque chimique et du domaine de l'épidémiologie. Une des améliorations possibles de cet outil est de représenter également en SML les parties moins structurées des tableaux (des listes de valeurs dans les cases du tableau) de façon à pouvoir les exploiter au mieux lors des requêtes. Cet outil pourra être intégré à la plate-forme logicielle qui sera développée dans le cadre du projet WebContent.

References

1. *http://www.symprevius.net.*
2. Patrice Buche, Juliette Dibie-Barthélemy, Ollivier Haemmerlé, and Mounir Houhou, *Towards flexible querying of xml imprecise data in a dataware house opened on the web*, Flexible Query Answering Systems (FQAS), Springer Verlag, june 2004.
3. e.dot, *Progress report of the e.dot project*, http://www-rocq.inria.fr/ amann/edot/, 2004.
4. Hélène Gagliardi, Ollivier Haemmerlé, Nathalie Pernelle, and Fatiha Saïs, *A semantic enrichment of data tables applied to food risk assessment*, DS '05, 8th International Conference on Discovery Science, Singapore, October 2005, Proceedings **3735** (2005), 374–376, LNCS-Lecture Notes in Computer Science.
5. Khaled Khelif and Rose Dieng-Kuntz, *Ontology-based semantic annotations for biochip domain*, EKAW (2004), 483–484 ().
6. Tok Wang Ling, Sudha Ram, and Mong-Li Lee (eds.), *Conceptual modeling - er '98, 17th international conference on conceptual modeling, singapore, november 16-19, 1998, proceedings*, Lecture Notes in Computer Science, vol. 1507, Springer, 1998.
7. A. Pivk, P. Cimiano, and Y. Sure, *From tables to frames*, Journal of Web Semantics **3** (2005).
8. Patrick Séguéla and Nathalie Aussenac-Gilles, *Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine*, IC, Ingéneirie des Connaissance, Palte-forme AFIA (1999).