MIEL++ : un entrepôt intégrant des données floues exprimées en graphes conceptuels, bases de données relationnelles et XML

Patrice Buche[1], Juliette Dibie-Barthélemy[2], David Doussot[2], Ollivier Haemmerlé[2,3], Gaëlle Hignette[2], Rallou Thomopoulos[4]

[1]INRA - Mét@risk, 16 rue Claude Bernard, F-75231 Paris Cedex 05, France
[2]UMR INA P-G/INRA MIA, 16 rue Claude Bernard, F-75231 Paris Cedex 05, France
[3]LRI (UMR CNRS 8623 - Université Paris-Sud) / INRIA (Futurs), Bâtiment 490, F-91405 Orsay Cedex, France
[4]INRA - UMR IATE - bat. 31, 2 Place Viala 34060 Montpellier Cedex 1

Mots-clés Intégration de données, graphes conceptuels, sous-ensembles flous, validation.

1 Introduction, domaine d'application

Nous présentons dans ce résumé les travaux menés par l'équipe de recherche en informatique de l'Institut National Agronomique Paris-Grignon au cours des 5 dernières années. Ces travaux ont trouvé leur champ d'application dans le domaine de la prévention du risque microbiologique dans les aliments. La sécurité des aliments est une préoccupation grandissante, les crises passées ayant la plupart du temps eu des répercussions catastrophiques d'un point de vue sanitaire mais également économique. L'évaluation du risque microbiologique ne peut se faire qu'à partir d'informations sur le comportement des germes pathogènes lors d'épisodes de contamination. Les équipes travaillant en microbiologie prévisionnelle ont été amenées à construire des bases de données permettant une capitalisation de ces connaissances.

Notre équipe travaille sur la construction d'un entrepôt de données, qui a vocation à contenir des données extraites de la littérature scientifique en microbiologie ou des données fournies par des partenaires industriels. Les données à stocker présentent un certain nombre de caractéristiques : (i) les données sont hétérogènes ; (ii) les données peuvent être imprécises ; (iii) l'entrepôt de données est incomplet par nature ; (iv) les sources de données ne sont pas de qualité égale.

Pour prendre en compte ces caractéristiques, nous travaillons sur les thèmes suivants :

- intégration de données contenues dans des bases fondées sur des formalismes différents ;
- représentation de données imprécises par le biais de sous-ensembles flous ;
- mécanismes d'interrogation élargie des bases palliant l'incomplétude ;
- enrichissement automatique de l'entrepôt de données par des données automatiquement extraites du Web ;
- validation de données.

2 Intégration de données, vue d'ensemble du système MIEL++

Nous avons choisi de traiter l'hétérogénéité de nos données en construisant un entrepôt de données intégrant trois sources de données distinctes par le biais d'une même interface d'interrogation (Cf. Fig. 1) : une base de données relationnelle, une base de graphes conceptuels, et une base de données XML.

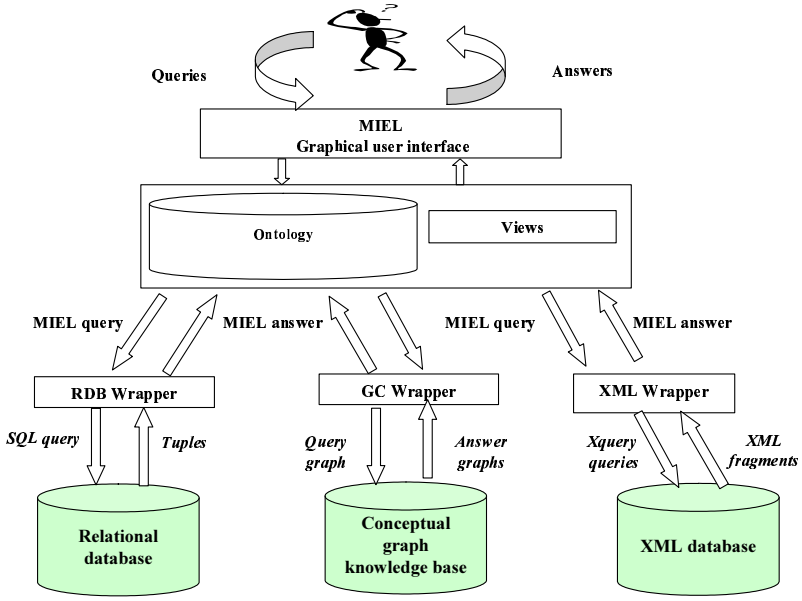

FIG. 1 – Le schéma global de l'entrepôt de données MIEL++

3 Données imprécises dans le modèle des graphes conceptuels

Les données expérimentales en microbiologie prévisionnelle présentent certaines formes d'imprécision : (i) variabilité liée à la complexité intrinsèque des processus biologiques ; (ii) imprécision des capteurs ; (iii) imprécision de l'expression des résultats dans les sources de données (publications scientifiques).

Nous avons choisi de représenter les données imprécises à l'aide de distributions de possibilités. Nous utilisons pour cela le modèle des sous-ensembles flous proposé par Zadeh. Nous avons étendu le modèle des graphes conceptuels autorisant une représentation de sous-ensembles flous dans les sommets concepts. Un sommet concept peut être porteur d'un type de concept flou ou d'un marqueur flou. L'opération de projection des graphes conceptuels reste définie comme un homomorphisme de graphes autorisant la spécialisation des étiquettes des sommets, en utilisant la nouvelle définition de la spécialisation des étiquettes des sommets concepts. L'extension du modèle à la représentation de sous-ensembles flous est utilisée pour représenter des données imprécises, mais également pour permettre à l'utilisateur d'exprimer des préférences dans les critères de sélection de ses requêtes à l'entrepôt de données.

4 Ouverture de l'entrepôt de données sur le Web

Le projet RNTL e.dot[1] a pour objectif la construction d'un entrepôt de données stockées au format XML alimenté automatiquement à partir du Web. Le domaine d'application choisi est également la prévention du risque microbiologique dans les aliments. Dans le cadre de ce projet, nous travaillons sur : (i) l'extraction d'informations à partir de publications scientifiques contenant des tableaux de données. Ces documents *pdf* sont transformés en documents XML étiquetés sémantiquement à partir de l'ontologie du système afin de permettre leur interrogation ; (ii) l'interrogation de ces documents étiquetés sémantiquement qui ne répondent pas nécessairement à un schéma prédéfini. Nous proposons une interrogation flexible de ces documents ; (iii) l'intégration des données aux bases de l'entrepôt.

Ces différentes parties de notre travail ont pour point commun d'être fondées sur l'utilisation d'une ontologie qui est proche de celle que l'on utilise dans le modèle des graphes conceptuels (un ensemble de types de concepts correspondant à une taxonomie de termes, un ensemble de relations, et des contraintes d'utilisation de ces concepts et relations).

5 Validation de graphes conceptuels

Ce travail a porté sur la validation syntaxique d'une base de connaissances exprimée en termes de graphes conceptuels. Nous avons proposé un ensemble de propriétés et avons montré que leur satisfaction globale par une base de connaissances garantissait la validité syntaxique de cette base. Nous avons également présenté des règles de réparation associées à chacune des propriétés.

Nous avons ensuite étudié la validation sémantique d'une base de connaissances relativement à des contraintes, que nous considérons comme des connaissances expertes extérieures à la base, uniquement fournies à des fins de validation et que nous supposons fiables. Ces contraintes sont classées en deux catégories : (i) des contraintes négatives qui permettent, dans une situation donnée, de s'assurer qu'une certaine conclusion ne peut pas être déduite. Cette forme de validation peut être comparée à l'étude de la cohérence de la base ; (ii) des contraintes positives qui permettent, dans une situation donnée, de s'assurer qu'une certaine conclusion peut être déduite. Cette forme de validation peut être comparée à l'étude de la complétude de la base.

[1] Entrepôt de Données Ouvert sur la Toile, mené avec l'équipe de Marie-Christine Rousset (LRI), l'équipe de Serge Abiteboul (INRIA) et la société Xyleme