Utilisation d'un outil de recherche d'information pour la mise en correspondance de thesaurus. Application aux sciences du vivant.

Laurie Planes[1], Stéphane Dervaux[1], Juliette Dibie-Barthélemy[1,2], Nicolas Guinet[3], Liliana Ibanescu[1,2]

[1] INRA - UNITÉ MET@RISK, 16 rue Claude Bernard, 75231 Paris Cedex 05
laurie.planes@agroparistech.fr,stephane.dervaux@paris.inra.fr

[2] AGROPARISTECH, 16 rue Claude Bernard, 75231 Paris Cedex 05
{juliette.dibie,liliana.ibanescu}@agroparistech.fr

[3] INRA - UNITÉ ALISS, 65 Boulevard de Brandebourg, 94200 Ivry-sur-Seine
nguinet@ivry.inra.fr

Résumé : Nous présentons dans cet article un outil d'étiquetage de plusieurs thesaurus afin d'aider des experts ayant des besoins différents à les mettre en correspondance. L'étiquetage des différents thesaurus, appelés thesaurus sources, est effectué à l'aide d'un thesaurus pivot. Il consiste à étiqueter chaque élément des thesaurus sources le plus largement possible avec des éléments du thesaurus pivot, ceci afin de prendre en compte les besoins des différents experts. Nous proposons de considérer la tâche d'étiquetage comme un problème de recherche d'information où chaque élément d'un thesaurus source est considéré comme une requête et chaque élément du thesaurus pivot comme un document sur lequel sera projeté la requête. Une expérimentation dans le cadre de la plateforme du Pôle Alimentation Parisien (PAP), permettant le traitement et l'exploitation de sources de données hétérogènes dans le domaine de l'Alimentation et de la Santé, est présentée.

Mots-clés : mise en correspondance de thesaurus, étiquetage, outil de recherche d'information, application aux sciences du vivant

1 Introduction

Nos travaux portent sur la mise en correspondance de thesaurus dans le cadre d'un projet interdisciplinaire visant à développer une plateforme [1] permettant le traitement et l'exploitation de sources de données hétérogènes, dans le domaine de l'Alimentation et de la Santé. Cette plateforme permet, d'une part, de croiser des sources de données d'origines et de natures diverses (données de consommation, données nutritionnelles, données de contamination, etc.) portant sur un même objet, l'aliment, afin de pouvoir

1. http ://www.inra.fr/pap

les exploiter et les analyser conjointement. Elle permet, d'autre part, de fournir aux experts de disciplines variées des outils adaptés à leurs méthodes scientifiques, de manière à les aider au plus proche de leurs besoins. Nous sommes ainsi confrontés à une double problématique : celle de la mise en correspondance de thesaurus d'aliments sur lesquel reposent différentes sources de données, et, celle de la prise en compte des différents besoins des experts. Jusqu'à présent, la mise en correspondance des thesaurus était effectuée manuellement par un expert. Cette mise en correspondance était dépendante de la discipline de rattachement de l'expert et de l'objectif de l'étude pour laquelle les données devaient être croisées. Outre le fait que ce travail était couteux et fastidieux, la mise en correspondance effectuée n'était pas réutilisable car trop dépendante de l'étude. Afin d'aider les experts à mettre en correspondance des thesaurus en fonction de leurs différents besoins, nous avons développé un outil d'étiquetage des éléments de ces thesaurus à l'aide d'un thesaurus pivot, qui s'appuie sur un moteur de recherche d'information. Nous présentons cet outil puis une expérimentation.

2 Etiquetage à l'aide d'un outil de recherche d'information

Les thesaurus sur lesquels portent notre étude sont composés d'aliments organisés en taxonomie, auxquels peuvent éventuellement être associés des caractéristiques telles que leur conditionnement, leur mode de production ou encore leur cuisson. Nous présentons dans cet article un outil permettant de faciliter la mise en correspondance de thesaurus sources par un étiquetage de leurs aliments à partir d'éléments (i.e. aliments et valeurs de leurs caractéristiques) d'un thesaurus pivot. L'étiquetage d'un aliment d'un thesaurus source consiste à lui attribuer un ensemble d'étiquettes permettant de le décrire avec le vocabulaire du thesaurus pivot. Cet étiquetage consiste donc à apparier cet aliment avec les éléments "les plus ressemblants" dans le thesaurus pivot. Nous proposons de ramener cette tâche d'étiquetage à un problème de Recherche d'Information (cf. Manning *et al.* (2008)). Dans cette optique, nous considérons chaque aliment d'un thesaurus source comme une requête et chaque élément du thesaurus pivot comme un document, sur lequel sera projeté la requête. Notre étiquetage revient alors à utiliser un moteur de recherche qui s'appuie sur un calcul de similarité permettant de quantifier la ressemblance entre un document et une requête sur la base du nombre de termes de la requête présents dans le document. Nous proposons de favoriser le rappel de l'étiquetage, c.à.d. de faire un étiquetage le plus large possible permettant d'établir des correspondances $n : m$ entre les éléments des deux thesaurus à croiser.

3 Expérimentation

L'expérimentation de notre outil d'étiquetage a été menée sur deux thesaurus d'aliments. Le premier thesaurus source est Codex [2] qui rassemble 500 aliments classés en

2. http ://www.codexalimentarius.net/gsfaonline/foods/index.html ?lang=en

4 familles, divisées en 7 groupes et 27 sous-groupes. Le second thesaurus source est INCA 2 [3] qui rassemble 1201 aliments. Nous avons choisi de prendre FoodEx (cf. EFSA (2011)) comme thesaurus pivot, qui dispose d'une bonne couverture avec les thesaurus sources testés. Le thesaurus FoodEx est composé d'une liste de 2673 aliments organisés en taxonomies et décrits par un ensemble de caractéristiques. Pour effectuer notre expérimentation, nous avons utilisé le moteur de recherche Terrier [4] qui produit en sortie des scores de similarité entre des aliments des requêtes et des éléments des documents, permettant de définir pour un ensemble d'aliments du thesaurus source, une liste d'étiquettes potentielles exprimées dans le vocabulaire du thesaurus pivot.

L'évaluation de notre outil a été effectuée manuellement par un expert. Pour mesurer la précision de notre méthode, nous avons utilisé le pourcentage de résultats jugés "'satisfaisants"' par l'expert, i.e. lorsque l'ensemble des éléments de la description de l'aliment a été reconnu dans les 20 premiers résultats ou lorsqu'au moins l'aliment principal a été reconnu (e.g. "'Shrimp or prawn"' reconuu dans "'Shrimp or prawn, boiled"'). Ce pourcentage est de 78% pour Codex et INCA 2. Ne disposant pas de "gold standard", nous avons utilisé, comme indicateur d'évaluation du rappel, la proportion d'aliments pour lesquels notre outil a ramené au moins une étiquette. Pour Codex, 93% des aliments ont pu être étiquetés (seuls 39 aliments n'ont pas d'étiquettes associées). Pour INCA 2, 99% des aliments ont pu être étiquetés (seuls 13 aliments n'ont pas d'étiquettes associées).

4 Conclusion

Ce premier travail nous a permis de développer rapidement un outil d'aide à la mise en correspondance de thesaurus en nous appuyant sur un moteur de recherche d'information existant. La prochaine étape de notre travail consistera à comparer les résultats de notre méthode avec ceux d'outils d'alignements d'ontologies existants, et en particulier ceux qui se sont bien classés dans la compétition OAIE 2012 [5] pour la tâche "Library" d'alignement de thesaurus en sciences sociales : GOMMA [6], LogMap [7], YAM++ [8].

Références

EFSA (2011). *The food classification and description system FoodEx 2*. Rapport interne, European Food Safety Authority.

MANNING C. D., RAGHAVAN P. & SCHTZE H. (2008). *Introduction to Information Retrieval*. New York, NY, USA : Cambridge University Press.

3. http ://www.anses.fr/index.htm
4. http ://terrier.org/
5. http ://oaei.ontologymatching.org/2012/
6. http ://dbs.uni-leipzig.de/GOMMA
7. http ://www.cs.ox.ac.uk/isg/tools/LogMap/
8. http ://www.lirmm.fr/∼dngo/