Une nouvelle méthode d'appariement entre deux vocabulaires d'annotation

V. Chhuo, Catherine Roussey, V. Soulignac, Stéphan Bernard, Jean-Pierre Chanet

Author-produced version of the paper presented at 4ème atelier Recherche d'Information SEmantique RISE
Conférence EGC 2012, January 31 2012, Bordeaux, France

Une nouvelle méthode d'appariement entre deux vocabulaires d'annotation

Vanna Chhuo *, Catherine Roussey*, Vincent Soulignac*,
Stephan Bernard*, Jean-Pierre Chanet*

*Irstea/Cemagref, 24 Av. des Landais, BP 50085, Aubiére, France

Résumé. KOFIS est un système de gestion des connaissances développé par Irstea/Cemagref pour améliorer la capitalisation des connaissances en agriculture biologique. Ce système se compose de deux applications web d'annotation de contenu. Chacune de ces applications disposent d'un vocabulaire d'annotation organisé hiérarchiquement. L'objectif de notre travail est de proposer une méthode d'appariement de vocabulaires hiérarchisés. Notre proposition combine une méthode d'appariement terminologique avec une méthode d'appariement structurel. Notre méthode d'appariement structurelle est une adaptation de l'approche "similarity flooding", qui prend en compte des alignements initiaux, et la transitivité de certaines relations hiérarchiques.

1 Introduction

L'agriculture est en pleine mutation : elle doit notamment modifier ses pratiques afin de limiter ses impacts négatifs sur l'environnement. L'agriculture intensive va évoluer vers une agriculture durable, voir biologique. Tant l'Europe que la France développe des programmes d'incitation au changement de pratiques agricoles. En France par exemple, le plan Ecophyto 2018 propose d'augmenter de 6% la surface agricole consacrée à l'agriculture biologique en 2012. Cependant, la conversion des pratiques agricoles vers une agriculture biologique est difficile car il y a peu de ressources, peu d'informations disponibles sur ce thème. Pour améliorer la capitalisation et la diffusion de ces savoirs, le Cemagref développe un système de gestion des connaissances en agriculture durable sur le web, intitulé KOFIS (Knowledge for Organic Farming and Innovative System), Soulignac et al. (2011). KOFIS se compose de deux applications web :

1. KOFIS_Innovation, un site web collaboratif construit à partir du système de gestion de contenu Drupal. Le contenu de cette application est annoté avec un ensemble de mots-clés organisés hiérarchiquement par une relation informelle.
2. KOFIS_Knowledge, un wiki sémantique construit à partir du moteur Semantic MediaWiki. Le contenu de cette application est annoté avec une ontologie du domaine composée de catégories organisées hiérarchiquement par une relation formelle.

Les deux applications KOFIS_Innovation et KOFIS_Knowledge sont indépendantes, elles ne partagent que leurs utilisateurs. Nous souhaitons mettre en place un système d'interrogation capable de retrouver à la fois des pages de KOFIS_Knowledge et de KOFIS_Innovation avec

la même requête. Dans un premier temps, le système d'interrogation doit permettre de construire une requête composée de mots clés du vocabulaire de KOFIS_Innovation pour retrouver des pages de KOFIS_Knowledge, c'est-à-dire des pages annotées par des catégories de KOFIS_Knowledge. Pour ce faire nous avons besoin de détecter automatiquement des appariements entre les mots clés de KOFIS_Innovation et les catégories de KOFIS_Knowledge. Cet article présente notre approche de détection d'appariements entre les deux vocabulaires d'annotation de KOFIS.

L'organisation de cet article est la suivante : la section 2 présente en détail le système KOFIS. Un état de l'art sur les méthodes d'appariement est proposé section 3. La section 4 présente notre approche d'appariement dédié au système KOFIS.

2 KOFIS : un système de gestion de connaissances en agriculture biologique

KOFIS est un système collaboratif de gestion des connaissances. KOFIS a pour objectif de partager et de diffuser les meilleures pratiques agricoles pour réduire les impacts négatifs de l'agriculture sur l'environnement. Ce système est accessible à différents types d'utilisateurs ayant des objectifs et des parcours professionnels variés. Parmi les profils d'utilisateurs intéressés par le système KOFIS, nous pouvons entre autres citer : les agriculteurs, les chercheurs en agronomie, les conseillers agricoles, les enseignants agricoles, etc. Chacun de ces profils a des droits d'accès différenciés dans le système KOFIS.

2.1 KOFIS_Innovation

KOFIS_Innovation est un espace ouvert où tous les utilisateurs de KOFIS peuvent créer des blogs ou poster des billets sur des thèmes relatifs à l'agriculture biologique. KOFIS_Innovation est construit à l'aide du système de gestion de contenu Drupal [1] .

Drupal permet de créer plusieurs types de contenu. Dans le cadre de KOFIS, uniquement deux types de contenu ont été retenus :

- les billets de blog : Un blog est une séquence de commentaires, appelés billets, postés par des auteurs différents. Cette suite de billets correspond à une discussion organisée au fil du temps. Le premier billet du blog pose le sujet de discussion auquel les billets suivants répondent. Dans le cadre de KOFIS_Innovation, un blog correspond à un problème lié à la production d'une culture biologique.
- Les pages de livre (book page) : Un livre permet d'organiser logiquement les pages de contenu en chapitre, section etc. Un outil de navigation affiche la structure logique de chaque livre. Dans KOFIS_Innovation un livre est associé à un type de culture et regroupe tous les blogs relatifs à ces cultures.

KOFIS_Innovation propose à tous ses auteurs, producteurs de contenu, d'annoter librement leur contenu avec des mots-clés. Cette annotation se fait en associant un champ à chaque type de contenu. Ce champ contiendra la liste des mots clés annotant le contenu. L'ensemble des mots clés forme un vocabulaire libre, organisé hiérarchiquement par une relation informelle comme la relation générique/spécifique des thésaurus. De plus, le vocabulaire d'annotation de

1. http ://drupal.org/

KOFIS_Innovation contient des mots clés issus du thésaurus Agrovoc géré par la FAO, Alonso et Sicilia (2007). Cette fonctionnalité est fournie par un module intitulé Agrovoc Field. Ce module interroge directement le service web d'Agrovoc pour aider l'utilisateur à annoter son texte avec des mots clés d'Agrovoc. Les relations hiérarchiques sont construites manuellement par un utilisateur ayant pour rôle de gérer le vocabulaire. Ces relations peuvent être modifiées à tout moment.

Ce vocabulaire est présenté aux utilisateurs dans un module de navigation, pour retrouver un contenu en fonction de son thème. En plus de cette fonctionnalité de recherche par navigation, KOFIS_Innovation dispose d'un module de recherche en "full text".

La figure 1 présente un extrait du vocabulaire de mots clés utilisé dans KOFIS_Innovation.

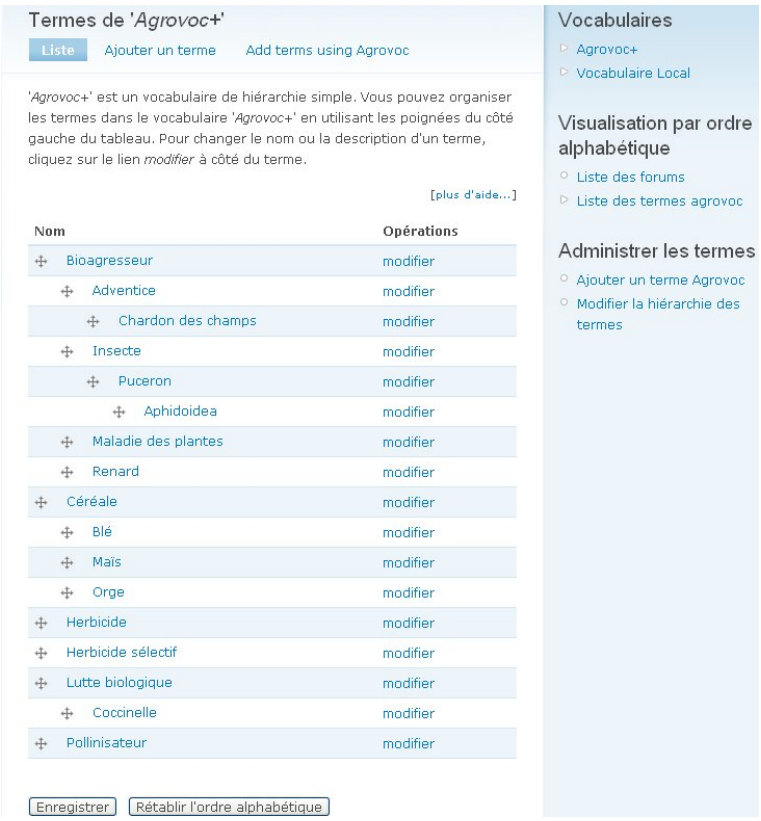

FIG. 1 – *le vocabulaire d'annotation de KOFIS_Innovation*

2.2 KOFIS_Knowledge

KOFIS_Knowledge est un espace fermé contenant uniquement des informations validées par des experts. KOFIS_Knowledge est construit à l'aide du moteur de wiki sémantique Se-

mantic MediaWiki (SMW), Völkel et al. (2006). Ce moteur est une extension du moteur de wiki MediaWiki utilisant des technologies Web Sémantique.

SMW est un wiki, c'est-à-dire un site web permettant la création et l'édition collaborative de pages de manière simple. SMW utilise des technologies Web Sémantique pour annoter les pages suivant un schéma de métadonnées prédéfini : une ontologie. Les annotations sont structurées c'est à dire composées de classes (appelées catégories) et de propriétés préalablement définies dans l'ontologie. SMW est un moteur de wiki de type "wiki for ontology", Meilender et al. (2010), l'annotation de contenu permet à la fois de mettre à jour l'ontologie et de la peupler avec des instances, mais la cohérence de la base de connaissance finale n'est pas garantie. Les auteurs peuvent définir deux éléments différents pour représenter la même information, et aucune inférence n'est utilisée pour valider l'ontologie.

Le vocabulaire d'annotation de KOFIS_Knowledge se compose donc de catégories et de propriétés. Les catégories sont organisées hiérarchiquement par une relation formelle "sous classe de" constituant la hiérarchie de classes de l'ontologie sous jacente. Ce vocabulaire est contrôlé, c'est-à-dire que seul l'utilisateur en charge de la gestion du vocabulaire à la possibilité d'ajouter ou de modifier des éléments de ce vocabulaire : par exemple, il peut ajouter de nouvelles catégories ou modifier la hiérarchie.

Grâce aux technologies Web Sémantique, KOFIS_Knowledge dispose, en plus d'une recher che "full text", d'un module d'interrogation structurée. Il est ainsi possible d'interroger l'ensemble des pages de KOFIS_Knowledge pour retrouver la liste des agresseurs biologiques du blé par exemple.

La figure 2 présente un exemple de page annotée par la catégorie "puceron". Cette page sur les Aphidoidea est une instance de la catégorie "Puceron". Deux autres propriétés ont été rajoutées à cette instance. En particulier, cette instance est liée par la propriété "lutte biologique" à la catégorie "syrphe" .

2.3 Un module d'interrogation commun

Les deux applications KOFIS_Innovation et KOFIS_Knowledge sont indépendantes, elles ne partagent que leurs utilisateurs. Chacune de ces applications dispose de ses propres modules de recherche d'information, mais à l'heure actuelle il n'existe pas de module de recherche capable de retrouver à la fois des pages de KOFIS_Knowledge et des billets de KOFIS_Innovation avec la même requête. Dans un premier temps, le module d'interrogation que nous envisageons doit permettre de construire une requête composée de mots-clés du vocabulaire de KOFIS_Innovation pour retrouver des pages de KOFIS_Knowledge, c'est-à-dire des pages annotées par des catégories de KOFIS_Knowledge. Pour ce faire nous avons besoin de construire automatiquement des correspondances entre les mots clés de KOFIS_Innovation et les catégories de KOFIS_Knowledge.

A l'installation, KOFIS_Knowledge contient déjà des pages annotées avec des catégories alors que KOFIS_Innovation est vide. Pour initialiser le vocabulaire de KOFIS_Innovation, les catégories de KOFIS_Knowledge sont dupliquées sous forme de mots clés. Ainsi au départ, les deux hiérarchies sont identiques et les correspondances entre les deux vocabulaires d'annotation sont connues. Au bout d'un certain temps d'utilisation, les deux vocabulaires évoluent indépendamment l'un de l'autre. Par conséquent, il devient nécessaire de détecter de nouvelles correspondances. Nous souhaitons mettre en place un système de détection semi automatique de correspondances utilisant les correspondances initiales et la structuration hiérarchique des

FIG. 2 – *Une page de KOFIS_Knowledge annotée*

vocabulaires. Le but de ce système est de proposer, à l'utilisateur en charge de la gestion des vocabulaires, une liste pondérée de correspondances à valider.

La figure 3 présente l'architecture générale de KOFIS. Les parties grisées représentent les nouveaux composants que nous souhaitons développer.

3 Etat de l'art sur l'appariement

L'appariement entre deux vocabulaires hiérarchisés est un processus de détection des correspondances entre des éléments appartenant à chacun des vocabulaires. Chaque correspondance doit être pondérée par un score. Le score d'appariement est un nombre réel qui évalue la similarité des deux éléments associés. Ce score prend une valeur réelle entre 0 et 1 : 1 signifiant que les éléments sont identiques, 0 que les éléments sont dissemblables. Une correspondance est

Une nouvelle méthode d'appariement entre vocabulaires

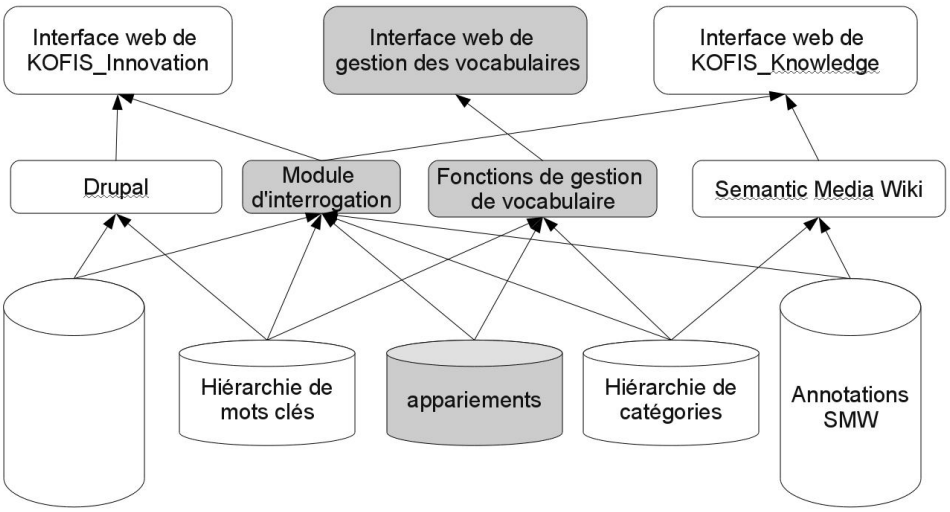

FIG. 3 – *architecture de KOFIS*

un couple (e_1, e_2) où e_1 est un élément du vocabulaire V_1 et e_2 est un élément du vocabulaire V_2.

Un appariement est composé d'une correspondance entre deux éléments (e_1, e_2) et de son score obtenu par une mesure de similarité, $\sigma(e_1, e_2)$.

Il existe plusieurs approches de détection d'appariements, ces approches sont utilisées dans différents domaines tels que l'appariement de schémas de bases de données, l'appariement d'ontologies du web sémantique, l'appariement de thésaurus. Dans un cadre plus général, chacun des objets à apparier (schéma, ontologie ou thésaurus) que nous nommerons vocabulaire hiérarchisé est un graphe dont les noeuds et les arcs sont étiquetés par des termes. Il existe plusieurs classifications des approches de détection d'appariements dans la littérature, Bellahsene et al. (2011), Kalfoglou et Schorlemmer (2003), Euzenat et Shvaiko (2007). On distingue plusieurs familles :

1. approche terminologique basée sur la comparaison de termes,
2. approche structurelle basée sur la structure des graphes,
3. approche sémantique utilisant une ressource externe pour déterminer l'interprétation des éléments à apparier,
4. approche hybride combinant plusieurs approches pour obtenir de meilleurs résultats.

Dans notre étude nous nous focaliserons uniquement sur les approches ne nécessitant pas de ressources externes.

3.1 Appariement terminologique

L'appariement terminologique détecte les correspondances entre des vocabulaires hiérarchisés à partir du contenu textuel associé aux éléments des vocabulaires. Ces approches se basent sur des techniques de comparaison de chaînes de caractères ou des techniques du Traitement Automatique du Langage Naturel (TALN).

Pour comparer des chaînes de caractères, il faut dans un premier temps normaliser et nettoyer ces chaînes. Les opérations sur les chaînes peuvent être par exemple :

- normalisation de la case, en remplacant chaque caractère par la minuscule correspondante : "Insecte" $\rightarrow$ "insecte"
- suppression des accents : "espèce" $\rightarrow$ "especе"
- suppression des caractères numériques ou des caractères de ponctuation "espèce1" $\rightarrow$ "espèce"
- remplacement de tout caractère de séparation de mot par un caractère espace : "espèce d'insecte" $\rightarrow$ "espèce d insecte"

Une fois les chaînes de caractères normalisées, une mesure de similarité entre chaînes est appliquée. Cette mesure peut être proportionnelle au nombre de caractères communs, au nombre de Ngrams communs, Kondrak (2005), à la longueur de la plus grande sous-chaîne commune, ou inversement proportionnelle au nombre de caractères dissemblables. Par exemple, la distance de Jaro Winkler, Winkler (1999), entre deux chaînes est proportionnelle au nombre de caractères communs. La distance de Hamming, Hamming (1950), évalue le nombre de positions dans les chaînes où les caractères différent. La distance de Levenshtein, Levenshtein (1965), détermine le nombre de transformations nécessaires pour obtenir une chaîne à partir d'une autre.

L'appariement terminologique, Safar et Reynaud (2009), utilisant des outils de TALN utilise en plus des opération de nettoyage sur les chaînes, des outils de normalisation linguistique pour éliminer les variations des termes propres à une langue donnée. Nous pouvons entre autres citer l'extraction des lemmes à partir des mots ("articles" $\rightarrow$ "article") , extraction des racines des lemmes ("travailler", "travailleur" $\rightarrow$ "travail"), etc. Il est aussi possible d'ajouter une ressource externe, comme un dictionnaire, pour détecter les termes synonymes.

3.2 Appariement structurel

L'appariement structurel détecte les correspondances en fonction de la structure des graphes.

Anchor-PROMPT est une des premières méthodes d'appariement structurel utilisée pour aligner des ontologies du web sémantique, Noy et Musen (2001). Cette méthode prend en entrée un ensemble d'ancres (des correspondances exactes entre deux classes) et retourne un nouvel ensemble de correspondances entre classes. Cette méthode considère l'ontologie comme un graphe dans lequel les classes sont des noeuds du graphe et les propriétés des classes sont des arcs. Cette méthode analyse les chemins de même longueur entre deux ancres (voir fig. 4). Deux noeuds de deux chemins qui apparaissent dans la même position obtiennent un score non nul. Le score entre ces deux noeuds augmentera s'ils apparaissent à la même position dans deux autres chemins. Enfin, les correspondances obtenues en sortie sont les couples de classes

ayant un score élevé. Cette méthode ne prend pas en compte l'étiquette des arcs (le nom des propriétés) entre les noeuds. Cette méthode obtient de bons résultats sur l'appariement d'ontologies : 75% de réponses correctes.

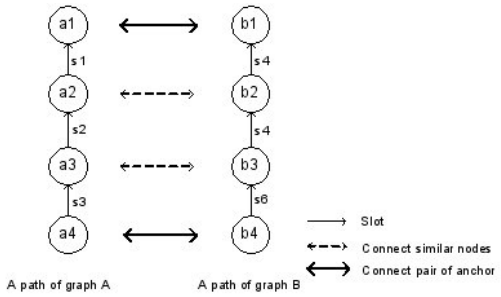

FIG. 4 – *Exemple d'analyse de chemins par Anchor-PROMPT : (a1,b1) et (a4,b4) sont des ancres ; (a2,b2) et (a3,b3) ont un score non nul.*

3.3 Appariement hybride par propagation de similarité

La méthode d'appariement intitulée "similarity flooding", Melnik et al. (2002), est une méthode d'appariement de graphes propageant des similarités terminologiques en fonction des arcs des graphes. Cette méthode part de l'hypothèse que la similarité de deux noeuds a et b appartenant à deux graphes G_1, G_2 augmente si il existe une similarité entre les noeuds adjacents de a et les noeuds adjacents de b .

Cette méthode nécessite la construction d'un graphe de connectivité par paire, intitulé PCG pour Pairwise Connectivity Graph . Un PCG est un graphe composé d'un ensemble de noeuds N et d'un ensemble d'arcs E : $PCG = \{N, E\}$.

- Un noeud du PCG représente une correspondance possible entre un noeud a de A et un noeud b de B : une paire. $N = \{(a, b), (a_1, b_1), ...\}$ avec $a, a_1 \in A$ et $b, b_1 \in B$.
- Un arc est un lien entre deux noeuds de N. Les arcs sont orientés et typés par un nom de relation. Nous représenterons les arcs sous forme de triplets < noeud source, relation, noeud cible>. L'arc du PCG de la figure 5 partant du noeud (a, b) vers le noeud (a_1, b_1) et typé par la relation l_1 se formalise de la manière suivante : $< (a, b), l_1, (a_1, b_1) >$.

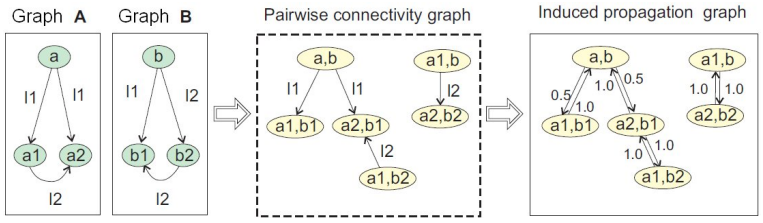

FIG. 5 – *Exemple de PCG*

Le PCG est construit suivant la formule :

$$\begin{aligned} < a, l_1, a_1 > \in A, < b, l_1, b_1 > \in B \quad &\Rightarrow \quad (a,b) \in N, (a_1,b_1) \in N \\ &\Rightarrow \quad < (a,b), l_1, (a_1,b_1) > \in E. \end{aligned} \tag{1}$$

Pour chaque arc du PCG, on construit un arc allant dans le sens opposé.

Les poids expriment le degré de propagation pc de la similarité d'une paire à ses voisins. On suppose que tous les arcs sortant d'un noeud et typés par la même relation ont une égale contribution.

Soit nb le nombre d'arcs sortant du noeud (a, b) typé par la relation R.

$$pc(< (a,b), R, (a_1,b_1) >) \quad = \quad \frac{1}{nb} \tag{2}$$

Itération apres itération la similarité initiale se propage sur tout le graphe PCG.

$$\begin{aligned} \sigma^i(a,b) \quad = \quad \sigma^{i-1}(a,b) + &\sum_{<x,R,a> \in A, <y,R,b> \in B} \sigma^{i-1}(x,y) * pc(< (x,y), R, (a,b) >) + \\ &\sum_{<a,R,x'> \in A, <b,R,y'> \in B} \sigma^{i-1}(x',y') * pc(< (a,b), R, (x',y') >) \end{aligned} \tag{3}$$

Cette similarité initiale, σ^0, est donnée par une mesure de similarité de chaînes. Pour chaque itération, on normalise les similarités en les divisant par la plus grande valeur de similarité obtenue.

L'algorithme se termine à l'obtention d'un point fixe : les similarités de toutes les paires se stabilisent. Si la convergence n'est pas possible, le nombre d'itérations est limité par une borne max.

$$\Delta(\sigma^n, \sigma^{n+1}) < \varepsilon \tag{4}$$

4 Proposition : un système de détection des appariements

Nous souhaitons intégrer à KOFIS un système de détection d'appariements entre le vocabulaire d'annotation de KOFIS_Innovation et le vocabulaire d'annotation de KOFIS_Knowledge. Dans un premier temps, nous ne souhaitons travailler que sur un sous-ensemble des vocabulaires de KOFIS :

– Le vocabulaire de KOFIS_Innovation est composé d'un ensemble de mots clés T organisés par une relation hiérarchique informelle, intitulé $skos : broader$, identique à la relation générique/spécifique des thésaurus. Ce vocabulaire constitue donc une hiérarchie de mots clés.
$\exists t, t' \in T,\ tq\ skos : broader(t, t') \in H_t$

– Le vocabulaire de KOFIS_Knowledge est composé en partie d'une hiérarchie de catégories organisées par la relation formelle $subClassOf$.
$\exists c, c' \in C,\ tq\ subClassOf(c, c') \in H_c$

Le but du système d'appariement est de découvrir des correspondances entre les mots clés t de KOFIS_Innovation et les catégories c de KOFIS_Knowledge, et d'associer à chaque correspondance (t, c) une valeur réelle représentant un degré de similarité fourni par la mesure de similarité $\sigma(t, c)$. Un mot clé peut être aligné avec plusieurs catégories et inversement.

Nous proposons de définir une nouvelle mesure de similarité entre des mots clés et des catégories en utilisant plusieurs informations propres au contexte de KOFIS :

– l'existence de correspondances intiales au démarrage de KOFIS,
– les relations hiérarchiques des vocabulaires.

Notre méthode de détection d'appariement se compose de plusieurs étapes, comme l'indique la figure 6 :

– calcul de la similarité initiale par la mesure σ_{init} représentant les alignements initiaux.
– calcul de la similarité terminologique, par la mesure σ_{termi}, entre les chaînes de caractères correspondant aux mots clés et aux noms des catégories. Cette mesure de similarité est basée sur le nombre de bigrammes communs que partagent chacune des chaînes de caractères.
– calcul de la similarité structurelle, par la mesure σ_{struct}, qui est une adaptation de l'approche de "similarity flooding" utilisant les valeurs de σ_{init} et σ_{termi} et la transitivité de la relation $subClassOf$.
– calcul de la similarité finale, par la mesure σ_{final}, qui combine les similarités terminologiques et structurelles.

4.1 Mesure de similarité initiale

Puisqu'à l'initialisation de KOFIS les noms de catégories de KOFIS_Knowledge sont dupliquées comme mots clés dans le vocabulaire d'annotation de KOFIS_Innovation, il existe dès le départ un ensemble de correspondances exactes que nous nommerons AI pour Alignement Initial. Cet ensemble de correspondances peut évoluer et correspondre à l'ensemble des correspondances exactes validées manuellement par l'utilisateur en charge de la gestion des vocabulaires.

$$\begin{aligned} \forall t \in T, \forall c \in C,\ tq\ (t, c) \in AI \quad &\Rightarrow \quad \sigma_{init}(t, c) = 1 \\ (t, c) \notin AI \quad &\Rightarrow \quad \sigma_{init}(t, c) = 0 \end{aligned} \tag{5}$$

4.2 Mesure de similarité terminologique

Les méthodes d'appariement utilisent souvent des mesures de comparaison de chaînes de caractères. Plusieurs mesures sont d'ailleurs proposées dans la littérature. Dans notre système, une mesure de comparaison de chaînes est utilisée pour comparer les mots clés de KOFIS_Innovation avec les noms des catégories de KOFIS_Knowledge.

Pour évaluer la similarité il est nécessaire, au préalable, de normaliser les chaînes de caractères. Cette normalisation consiste en plusieurs étapes :

– mise en minuscule des caractères,

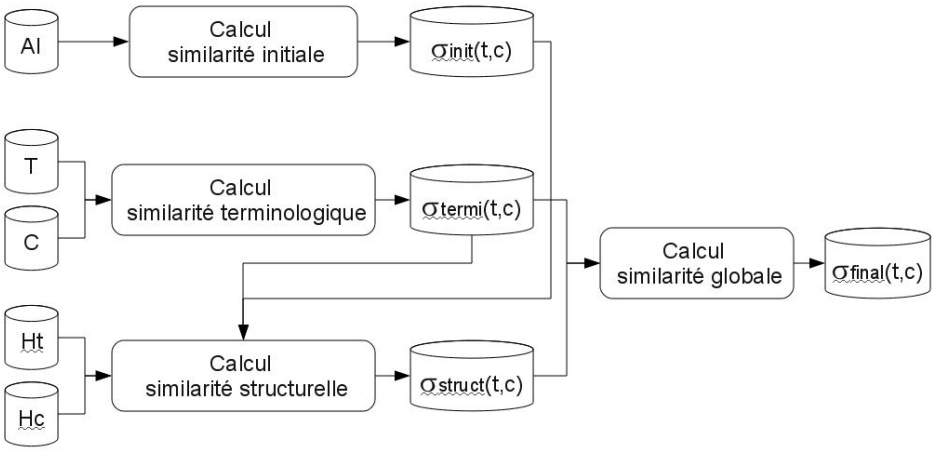

FIG. 6 – *présentation générale de notre méthode d'appariement*

- remplacement des caractères de séparation de chaîne par des espaces et réduction des suites d'espaces,
- suppression des caractères de ponctuation,
- remplacement des espaces par le caractère souligné,
- ajout en début et fin de chaîne d'un caractère souligné.

Ainsi la chaîne "article scientifique" est transformée en "_article_scientifique_".

La mesure de similarité entre deux chaînes de caractères normalisées évalue le nombre de bigrammes communs à l'aide du coefficient de Dice, comme proposé dans Kondrak (2005). Le fait d'ajouter un caractère en début et fin de chaîne permet de ne pas défavoriser les premiers et derniers caractères de la chaîne. Chaque caractère apparait deux fois dans l'ensemble des bigrammes. Par exemple l'ensemble des bigrammes issus de la chaîne "_article_scientifique_" est donné par la fonction $bigram(\text{"}\_article\_scientifique\_\text{"}) =$
$\{\_a, ar, rt, ti, ic, cl, le, e\_, \_s, sc, ci, ie, en, nt, if, fi, iq, qu, ue\}$

Nous utilisons la fonction $chane$ qui retourne pour un élément de vocabulaire la chaîne de caractère associée. Ainsi la similarité terminologique se calcule suivant la formule suivante :

$$\forall t \in T, \forall c \in C,$$

Une nouvelle méthode d'appariement entre vocabulaires

$$\sigma_{termi}(t,c) \quad = \quad \frac{2 * |bigram(chane(t)) \cap bigram(chane(c))|}{|bigram(chane(t))| + |bigram(chane(c))|} \tag{6}$$

4.3 Mesure de similarité structurelle

Notre mesure de similarité structurelle est basée sur la méthode d'appariement "similarity flooding". Nous allons adapter la construction du PCG en fonction du contexte de KOFIS. Un PCG est un graphe composé d'un ensemble de noeuds N et d'un ensemble d'arcs E : $PCG = \{N, E\}$.

Choix des graphes à aligner Notre méthode a pour objectif d'aligner la hiérarchie des mots clés de KOFIS_Innovation H_t avec la hiérarchie des catégories de KOFIS_Knowledge H_c.

Dans un premier temps, nous allons enrichir la hiérachie des catégories H_c en ajoutant de nouvelles relations $subClassOf_{trans}$ par transitivité de la relation $subClassOf$. L'objectif de cet ajout est de détecter des correspondances même si la hiérarchie de KOFIS_Knowledge est plus détaillée que celle de KOFIS_Innovation. Nous voulons pouvoir apparier des chemins de longueur différente. Nous souhaitons obtenir des correspondances qui ne suivent pas strictement la structure des graphes. D'après l'exemple de la figure 7, les correspondances trouvées à partir de la méthode "similarity flooding" traditionnelle seront ["céréale", "céréale"] et ["blé tendre", "blé"]. Nous souhaitons trouver un autre ensemble de correspondances ["céréale", "céréale"] et ["blé tendre", "froment"].

$$\begin{aligned} &\forall c, c', c'' \in C \; tq \; subClassOf(c, c') \in H_c, subClassOf(c', c'') \in H_c \\ &\qquad subClassOf(c, c'') \notin H_c \;\Rightarrow\; subClassOf_{trans}(c, c'') \in H_c \end{aligned} \tag{7}$$

Construction des noeuds du PCG : N Contrairement à l'approche initiale de "similarity flooding", les noeuds du PCG ne vont pas représenter l'ensemble des correspondances possibles. Nous allons limiter les correspondances aux mots clés de KOFIS_Innovation qui n'ont pas encore été associés dans les alignements initiaux AI, auxquels on rajoute les correspondances validées dans les alignements initiaux.

Nous partons de l'hypothèse qu'un mot clé de KOFIS_Innovation qui a deja été associé dans les AI n'a pas besoin d'etre associé à une autre catégorie de KOFIS_Knowledge.

$$\begin{aligned} \forall t \in T, \forall c \in C \; tq \; (t,c) \in AI \quad &\Rightarrow \quad (t,c) \in N \\ \left.\begin{array}{l} \forall t, t' \in T, tq \nexists (t,c'') \in AI, \nexists (t',c''') \in AI, \\ skos : broader(t,t') \in H_t, subClassOf(c,c') \in H_c \end{array}\right\} \quad &\Rightarrow \quad \left\{\begin{array}{l} (t,c) \in N, \\ (t',c') \in N \end{array}\right. \\ \left.\begin{array}{l} \forall t, t' \in T, tq \nexists (t,c'') \in AI, \nexists (t',c''') \in AI, \\ skos : broader(t,t') \in H_t, subClassOf_{trans}(c,c') \in H_c \end{array}\right\} \quad &\Rightarrow \quad \left\{\begin{array}{l} (t,c) \in N, \\ (t',c') \in N \end{array}\right. \end{aligned} \tag{8}$$

Construction des arcs du PCG : E Normalement le PCG est calculé à partir de graphes ayant des types de relations identiques. Nous allons calculer les arcs du PCG en considérant que la relation $skos : broader$ est équivalente aux relations $subClassOf$, $subClassOf_{trans}$. Pour se faire, nous allons définir deux relations pour typer les arcs du PCG : R et R_{trans}.

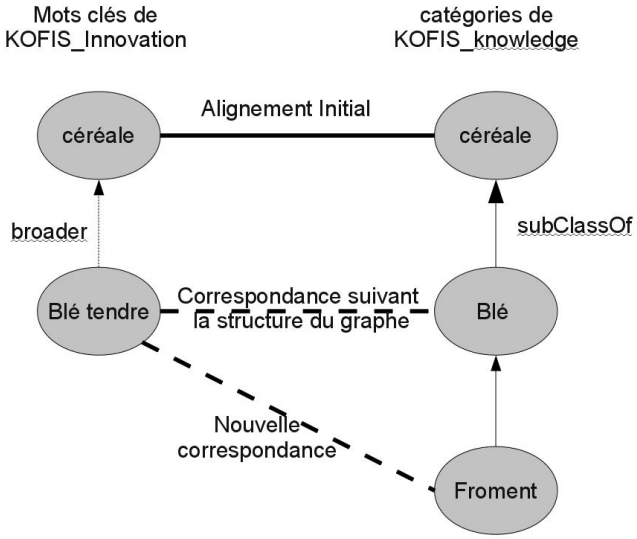

FIG. 7 – *Un exemple de correspondances possibles*

$$\forall skos : broader(t,t') \in H_t, \forall subClassOf(c,c') \in H_c \quad \Rightarrow \quad <(t,c),R,(t',c')> \in E \quad (9)$$
$$\forall skos : broader(t,t') \in H_t, \forall subClassOf_{trans}(c,c') \in H_c \quad \Rightarrow \quad <(t,c),R_{trans},(t',c')> \in E$$

Pour chaque arc du PCG on construit un arc allant dans le sens opposé et typé par le même nom de relation.

Calcul des degrés de propagation des arcs du PCG : pc Les poids des arcs expriment le degré de propagation de la similarité d'un noeud à ses voisins. L'ajout des relations transitives dans H_c va favoriser les noeuds du PCG contenant les catégories les plus génériques. Pour limiter cet impact, le degré de propagation d'un arc issu d'une relation directe devra être supérieur au degré de propagation d'un arc issu d'une relation transitive. Nous posons arbitrairement qu'un arc typé par la relation R aura deux fois plus de poids qu'un arc typé avec la relation R_{trans}.

Pour un noeud du PCG (t,c) donné, soit nb_d le nombre d'arcs typés par la relation R sortant de ce noeud $<(t,c),R,(t',c')>$ et soit nb_{ind} le nombre d'arcs typés par la relation R_{trans} sortant de ce noeud $<(t,c),R_{trans},(t',c')>$.

Le degré de propagation pc des arcs sortant de ce noeud est fixé par la formule suivante :

$$\begin{aligned} nb &= nb_d + \frac{nb_{ind}}{2} \\ pc(<(t,c),R,(t',c')>) &= \frac{1}{nb} \\ pc(<(t,c),R_{trans},(t',c')>) &= \frac{1}{2*nb} \end{aligned} \tag{10}$$

La somme des degrés de propagation des arcs sortant d'un noeud du PCD donné sera égale à 1.

Algorithme "similarity flooding" Une fois le PCG construit, nous allons appliquer un algorithme itératif. Itération après itération, la similarité initiale σ^0_{struct} des noeuds (t,c) se propage sur tout le graphe PCG. L'algorithme se termine à l'obtention d'un point fixe : les similarités de tous les noeuds se stabilisent.

La formule pour calculer $\sigma^i_{struct}(t,c)$, la similarité structurelle à l'itération i entre le mot clé t et la catégorie c, est donnée par la formule suivante :

$$\begin{aligned} \sigma^i_{struct}(t,c) = \quad & \sigma^{i-1}_{struct}(t,c) \quad + \\ & \sum_{<(t,c),R,(t',c')>\in E} \sigma^{i-1}_{struct}(t',c') \quad *pc(<(t,c),R,(t',c')>) + \\ & \sum_{<(t,c),R_{trans},(t',c')>\in E} \sigma^{i-1}_{struct}(t',c') \quad *pc(<(t,c),R_{trans},(t',c')>) + \\ & \sum_{<(t'',c''),R,(t,c)>\in E} \sigma^{i-1}_{struct}(t'',c'') \quad *pc(<(t'',c''),R,(t,c)>) + \\ & \sum_{<(t'',c''),R_{trans},(t,c)>\in E} \sigma^{i-1}_{struct}(t'',c'') \quad *pc(<(t'',c''),R_{trans},(t,c)>) \end{aligned} \tag{11}$$

Pour que la similarité structurelle soit une valeur entre 0 et 1, il faut, à chaque itération, normaliser les similarités. Pour ce faire, nous divisons chaque mesure par la plus grande valeur de σ^i_{struct} trouvée dans le PCG à l'itération i.

L'algorithme s'arrête quand le point fixe est atteint ou quand le nombre d'itérations atteint la borne max :

$$\Delta(\sigma^n_{struct}, \sigma^{n+1}_{struct}) < \varepsilon \tag{12}$$

Initialisation de la similarité structurelle Pour débuter l'algorithme, il faut pondérer chaque noeud (t,c) du PCG avec une mesure de similarité initiale. Si la correspondance (t,c) appartient aux alignements initiaux cette valeur est égale à 1 ; sinon elle correspond à la mesure de similarité terminologique.

$$\begin{aligned} \forall t \in T, \forall c \in C,\ tq(t,c) \in N(t,c) \in AI \quad &\Rightarrow \quad \sigma^0_{struct}(t,c) = \sigma_{init}(t,c) \\ (t,c) \notin AI \quad &\Rightarrow \quad \sigma^0_{struct}(t,c) = \sigma_{termi}(t,c) \end{aligned} \tag{13}$$

4.4 Mesure de similarité finale

Pour donner plus ou moins d'impact à l'une des similarités que nous avons défini, nous proposons de calculer une similarité finale qui est la somme pondérée des similarités terminologiques et structurelles. Soit β le poids fixé arbitrairement à la similarité terminologique, nous obtenons la formule suivante :

$$\begin{array}{l} \forall t \in T, \forall c \in C \\ \quad \sigma_{final}(t,c) \quad = \quad \beta * \sigma_{termi}(t,c) + (1-\beta) * \sigma_{struct}(t,c) \end{array} \tag{14}$$

En faisant varier β, nous allons pouvoir détecter avec le même système plusieurs types d'appariement :

- la polysémie ($\beta = 1$) : Dans ce cas nous n'utilisons que la similarité terminologique. Si un mot clé est polysémique, il est associé à deux catégories portant des noms trop proches.
- les correspondances exactes ($\beta = 0,5$) : Un mot clé correspond bien à une catégorie.

5 Conclusion

Dans cet article, nous avons présenté KOFIS un système de gestion des connaissances développé par Irstea/Cemagref pour améliorer la capitalisation des connaissances en agriculture biologique. Ce système se compose de deux applications web d'annotation de contenu. Chacune de ces applications disposent d'un vocabulaire d'annotation organisé hiérarchiquement. Dans le but d'intégrer ces deux applications, nous avons proposé une méthode d'appariement de vocabulaires hiérarchisés. Notre proposition combine une mesure de similarité terminologique avec une mesure de similarité structurelle. Notre mesure de similarité structurelle est une adaptation de l'approche "similarity flooding", qui prend en compte des alignements initiaux, et la transitivité de certaines relations hiérarchiques.

Références

Alonso, S. S. et M.-Á. Sicilia (2007). Using an agrovoc-based ontology for the description of learning resources on organic agriculture. In M.-Á. Sicilia et M. D. Lytras (Eds.), *MTSR*, pp. 481–492. Springer.

Bellahsene, Z., A. Bonifati, et E. Rahm (2011). *Schema matching and mapping*. Data-Centric Systems and Applications. Springer.

Euzenat, J. et P. Shvaiko (2007). *Ontology matching*. Springer.

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal 29*(2), 147–160.

Kalfoglou, Y. et M. Schorlemmer (2003). Ontology mapping : the state of the art. *The knowledge engineering review 18*(01), 1–31.

Kondrak, G. (2005). *N*-gram similarity and distance. In *SPIRE*, Volume 3772 of *LNCS*, pp. 115–126. Springer.

Une nouvelle méthode d'appariement entre vocabulaires

Levenshtein, V. (1965). Binary codes capable of correctng deletions, insertions and reversals. pp. 707–710.

Meilender, T., N. Jay, J. Lieber, et T. Palomares (2010). Les moteurs de wikis sémantiques : un état de l'art. rapport technique hal-00542813, INRIA, CNRS : UMR7503, Université Henri Poincaré, Nancy I, Université Nancy II, Institut National Polytechnique de Lorraine, Nancy, France.

Melnik, S., H. Garcia-Molina, et E. Rahm (2002). Similarity flooding : A versatile graph matching algorithm and its application to schema matching. In *ICDE*, pp. 117–128. IEEE Computer Society.

Noy, N. et M. Musen (2001). Anchor-PROMPT : using non-local context for semantic matching. In *Proceedings of the workshop on ontologies and information sharing at the international joint conference on artificial intelligence (IJCAI)*, Washington, USA, pp. 63–70.

Safar, B. et C. Reynaud (2009). Alignement d'ontologies basé sur des ressources complémentaires illustration sur le système taxomap. *Technique et Science Informatiques 28*(10), 1211–1232.

Soulignac, V., J. Ermine, J. Paris, O. Devise, et J. Chanet (2011). A knowledge server for sustainable agriculture. Bangkok, pp. 14.

Völkel, M., M. Krötzsch, D. Vrandecic, H. Haller, et R. Studer (2006). Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, Scotland, pp. 585–594. ACM.

Winkler, W. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Census Bureau.

Summary

The project associated to this work is the knowledge management system in sustainable agriculture called KOFIS. KOFIS consists of two tools for web pages annotation using two different vocabularies. The elements of the annotation vocabularies are organized hierarchically. The objective of the work is to propose a matching system for matching two annotation vocabularies. A study on the context of project is presented to identify research problems. Then a state of the art on matching method is established. Finally, a matching system is proposed consisting of several methods from existing works. Our main contribution is a new method of structure based matching approach adapted from "Similarity Flooding" [1]. This method takes into account the initial alignment and the transitivity of some hierarchical relations.