Actes du quatrième atelier Recherche d'Information SEmantique associé à la conférence EGC 2012

Catherine Roussey, J-P. Chevallet

Actes des Ateliers d'EGC 2012

Quatrième Atelier Recherche d'Information SEmantique RISE 2012

Contexte

Alors que la production actuelle de documents se fait essentiellement de manière numérique, les documents plus anciens sont en passe à leur tour d'être accessibles numériquement. La numérisation de l'information fait évoluer la notion même de document, et induit de nouvelles situations de communication (blogs, SMS, réseaux sociaux). Toutefois, même si cette masse informations est disponible, la difficulté majeure réside dans l'extraction et l'accès à de l'information ciblée, c'est à dire réellement en adéquation avec un besoin personnel et ponctuel. Ainsi, pour un accès plus "sémantique" à l'information, il est nécessaire d'en extraire des connaissances utilisées dans des moteurs de recherche d'un nouveau genre : les outils de Recherche d'Information Sémantique. Cependant, la nature peu structurée des documents, et leur expression en langue naturelle présente un obstacle à une mise en forme efficace des connaissances qu'ils contiennent. Pour résoudre ces problèmes, les travaux en Recherche d'Information (RI) s'orientent vers les technologies issues du Web Sémantique comme l'usage des ressources sémantiques (ontologies, thésaurii ou les bases de données lexicales), et celles issues du Traitement Automatique des Langues. Cet atelier est alors à la confluence de l'Extraction et de la Gestion de Connaissances et la Recherche d'Information.

Objectifs

Les travaux sur les ontologies ou les ressources sémantiques sont actifs dans les différentes communautés informatique comme : le Web, la bioinformatique, de domaine médical ou les systèmes d'information géographiques. Ainsi, les ressources sémantiques comme les ontologies, les bases de données lexicales, les thésaurii, se développent et sont facilement disponibles. Les techniques de fouilles et d'extraction d'information permettent de construire, nettoyer et enrichir ces ressources sémantiques. L'atelier RISE est donc spécialement dédié à l'usage des techniques de fouilles pour développer des ressources sémantiques utilisées dans des systèmes de Recherche d'Information. Ainsi, cet atelier a pour but de proposer un lieu de rencontre entre des chercheurs issus de différentes communautés comme l'Extraction de Connaissances, la Recherche d'Information, mais aussi le Web Sémantique et le Traitement Automatique des langues.

Thèmes

Les principaux thèmes abordés sont alors (liste non exhaustive, d'autres thèmes connexes peuvent être traités par les auteurs) :

- Extraction de Connaissance pour la Recherche d'Information ;
- Techniques de Fouilles pour la construction et l'enrichissement

d'Ontologies ou de Ressources Sémantiques ;
- Recherche d'Information Sémantique, Annotation Sémantique ;
- Alignement d'Ontologie et Correspondance pour la Recherche d'Information ;
- Langages de Représentation des connaissances pour la Recherche d'Information ;
- Usage de larges Bases de Connaissances pour la Recherche d'Information.

Contenu

Le programme de l'atelier RISE 2012 est disponible sur le site RISE.

L'atelier débute par une conférence invitée d'Olivier CORBY, Chargé de Recherche à l'INRIA intitulée "Un peu de (Web) sémantique pour la Recherche d'Information"
Dans cet exposé nous passerons en revue la nouvelle version du langage W3C SPARQL 1.1 dans une perspective de RI sémantique. Nous nous focaliserons en particulier sur les nouveaux énoncés : select expression, aggrégation, negation, chemin, sous-requête. Nous verrons quelques extensions de SPARQL 1.1 conçues dans l'équipe Edelweiss: recherche approchée, chemins étendus, XPath et XML. Enfin nous verrons comment les graphes nommés RDF et l'énoncé graph de SPARQL permettent de contextualiser la recherche d'information.

Quatre contributions sont ensuite présentées:

- Vanna Chhuo, Catherine Roussey, Vincent Soulignac, Stephan Bernard, Jean-Pierre Chanet "Une nouvelle méthode d'appariement entre deux vocabulaires d'annotation".
- Iana Atanassova, Marc Bertin, Jean-Pierre Descles "Ordonnancement des réponses pour une recherche d'information sémantique à partir d'une ontologie discursive".
- Karam Abdulahhad, Jean-Pierre Chevallet, Catherine Berrut "Matching Fusion with Conceptual Indexing".
- Nada Mimouni, Adeline Nazarenko, Sylvie Salotti "Analyse formelle et relationnelle de concepts pour la modélisation et l'interrogation d'une collection documentaire"

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.

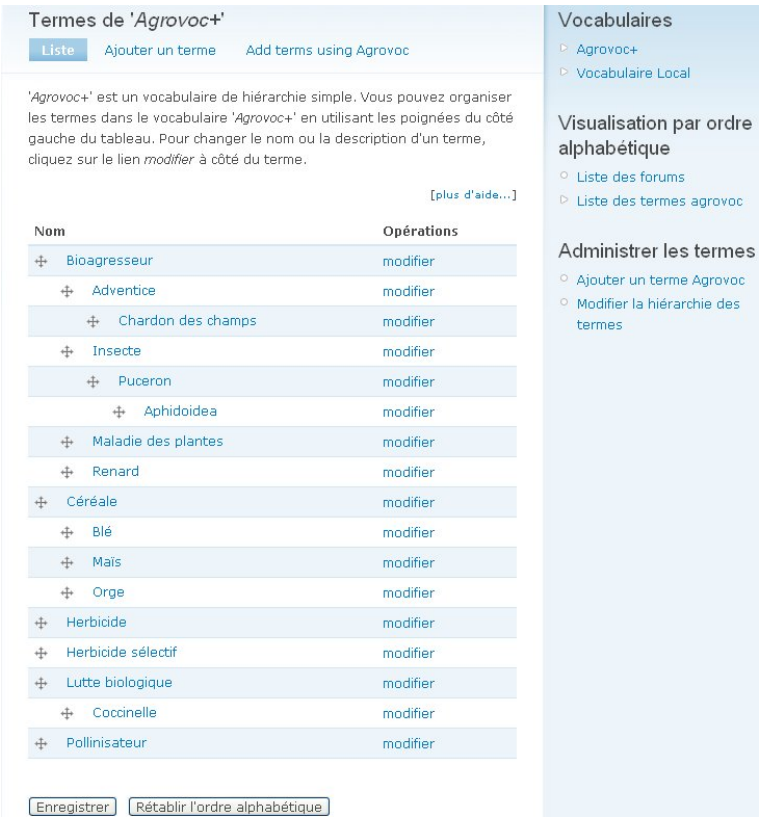

FIG. 1 – *le vocabulaire d'annotation de KOFIS_Innovation*

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

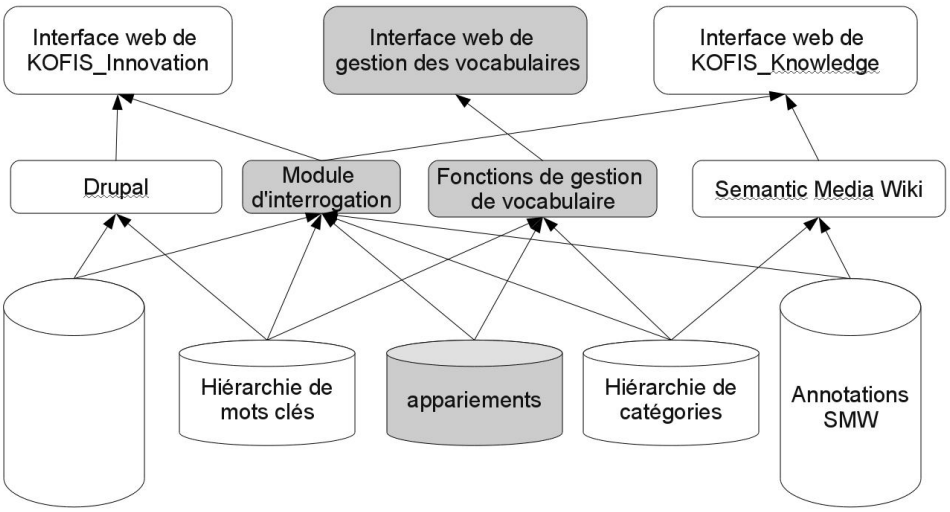

FIG. 3 – *architecture de KOFIS*

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.

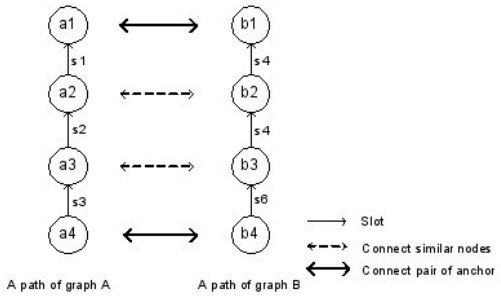

FIG. 4 – *Exemple d'analyse de chemins par Anchor-PROMPT : (a1,b1) et (a4,b4) sont des ancres ; (a2,b2) et (a3,b3) ont un score non nul.*

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 : $< (a, b), l_1, (a_1, b_1) >$.

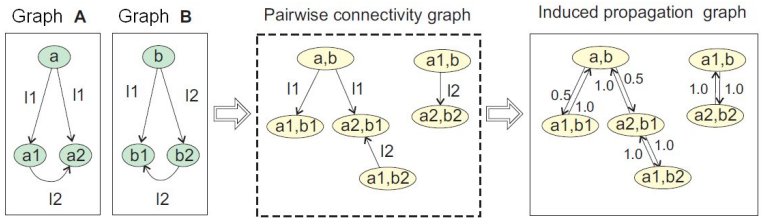

FIG. 5 – *Exemple de PCG*

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,

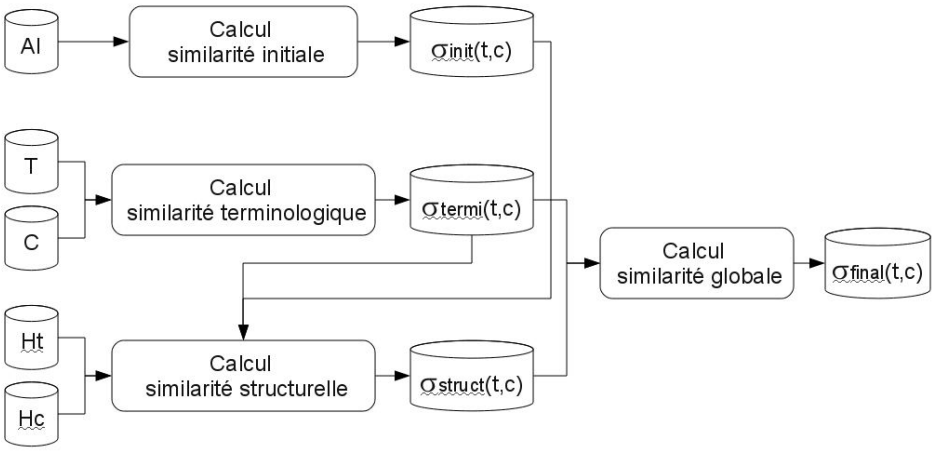

FIG. 6 – *présentation générale de notre méthode d'appariement*

- 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 :

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

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

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}.

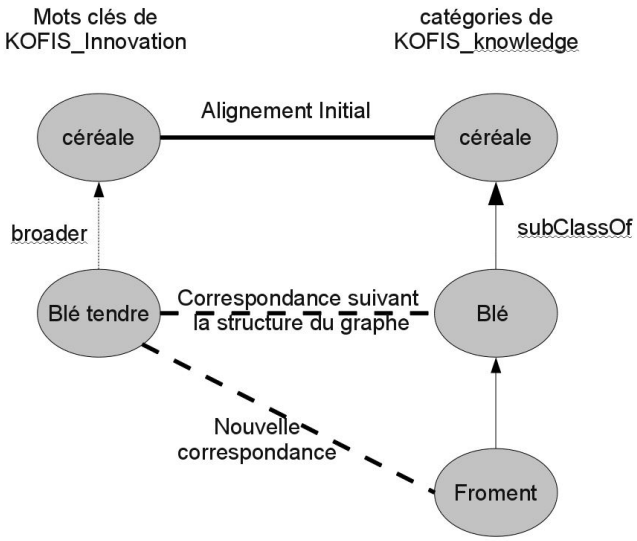

FIG. 7 – *Un exemple de correspondances possibles*

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

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.

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.

Ordonnancement des réponses pour une recherche d'information sémantique à partir d'une ontologie discursive

Iana Atanassova, Marc Bertin, Jean-Pierre Desclés

LaLIC (Langues, Logiques, Informatique, Cognition)
Université Paris-Sorbonne
Maison de la recherche
28 rue Serpente 75006 Paris
{iana.atanassova | marc.bertin | jean-pierre.descles}@paris-sorbonne.fr
http://lalic.paris-sorbonne.fr

Résumé. Nous avons élaboré un système de recherche d'information sémantique en utilisant les annotations automatiques issues d'une ontologie discursive. Le système permet d'effectuer des recherches selon différentes catégories sémantiques liées aux citations bibliographiques dans des publications scientifiques. Nous avons proposé de nouveaux critères d'ordonnancement qui exploitent les annotations et la structure de lontologie. Ces critères prennent en compte la spécificité et la diversité des annotations, ainsi que les types de marqueurs linguistiques. Nous avons obtenu un score de pertinence que nous avons évalué en comparant les résultats du système avec des jugements de pertinence humains. L'évaluation montre que les algorithmes proposés produisent un ordonnancement de qualité pour les résultats se trouvant en début de liste.

1 Introduction

L'annotation sémantique permet d'accéder au contenu textuel de façon plus pertinente que par une recherche uniquement par mots clés. Il s'agit d'étendre le modèle de recherche classique afin de prendre en compte des annotations sémantiques automatiques des relations dans des textes (*hypothèse, rencontre, définition, citation,* ...), qui sont identifiées automatiquement par un moteur d'annotation. Notre approche de la recherche et extraction d'informations revient à croiser deux types de requêtes, l'une portant sur la sélection d'un point de vue et l'autre sur un ensemble de termes. Les points de vue correspondent aux catégories annotées dans les textes et qui permettent de filtrer l'ensemble des occurrences des termes de la requête. Si les annotations font partie de l'index du système, elles sont également utilisées par le langage des requêtes, donnant ainsi la possibilité d'exprimer le besoin informationnel à partir de relations sémantiques.

2 Ontologie linguistique et annotation sémantique automatique

Nous traitons des corpus annotés automatiquement par des catégories discursives, organisées dans une ontologie linguistique, appelée également *carte sémantique*. Une carte sémantique est le produit d'une conceptualisation des relations sémantiques dans les textes. Elle est un treillis dont les nuds sont des classes de concepts, et dont les arcs orientés représentent des liaisons de spécifications et de généralisations entre ces classes. Les concepts (ou points de vue) dans la carte sémantique peuvent être de nature grammaticale (par exemple *accompli/inaccompli*) ou discursive (par exemple *annonce thématique, rencontre, définition*). Les catégories sont instanciées par des classes de marqueurs linguistiques, appelés *indicateurs*, qui sont des expressions (continues ou discontinues) porteuses de la sémantique de chaque catégorie. Ces expressions sont identifiables à la surface des textes. Étant donné que les formes linguistiques sont souvent polysémiques, l'occurrence d'un indicateur dans un segment n'est pas suffisante pour attribuer la catégorie d'annotation. La désambiguïsation se fait par l'examen du contexte en vérifiant la présence ou absence d'un certain nombre d'indices contextuels. L'annotation s'inscrit ainsi dans la méthode d'Exploration Contextuelle (EC) Desclés et al. (1997) qui est une technique opératoire permettant de tenir compte du contexte pour lever l'indétermination sémantique des formes linguistiques.

Au-delà d'être un réseau de concepts, la carte sémantique organise les marqueurs et les règles d'EC sous-jacentes dans un réseau avec des relations de spécification, entre les concepts, ainsi qu'entre les classes de marqueurs linguistiques, et d'instanciation, entre chaque concept et la classe de marqueurs qui lui sont associés. Elle est issue d'une étude linguistique qui a pour but de systématiser les marqueurs linguistiques par lesquels points de vue se réalisent à la surface des textes.

Les règles d'EC sont écrites sous forme déclarative, où la vérifications de la présence ou de l'absence de formes de surface spécifiques (indices contextuels) déclenche l'annotation du segment (voir la figure 1). Comme le montre Atanassova (2012), la métode d'EC permet de reconnaître une classe de langages plus grande que celle des langages régulières et elle a une complexité linéaire par rapport à la longueur du segment annoté pour un ensemble de règles non-récursif.

Nous avons appliqué notre méthode en utilisant l'ontologie de la Bibliosémantique (voir la figure 2), qui propose une catégorisation des citations entre auteurs dans des publications scientifiques Bertin (2008, 2011). Elle est issue de l'analyse des contextes de références bibliographiques afin de dégager les différentes motivations d'un auteur pour citer une publication, par exemple pour introduire une méthode, comparer ses travaux aux autres, reprendre une définition, etc. L'annotation permet, à travers une étude de la bibliographie, d'identifier les relations entre auteurs. Les éléments de base pour l'annotation sont les phrases. Les figures 3 et 4 présentent quelques exemples de phrases annotées obtenues par le système. Les annotations multiples sont possibles lorsque la même phrase est porteuse de plusieurs relations sémantiques de l'ontologie.

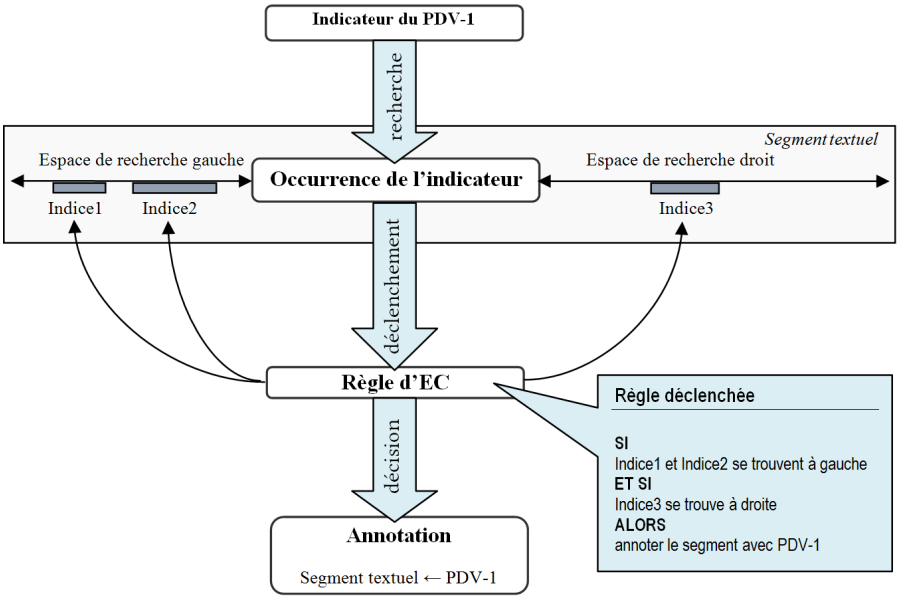

FIG. 1 – *Schéma de fonctionnement de la méthode d'EC*

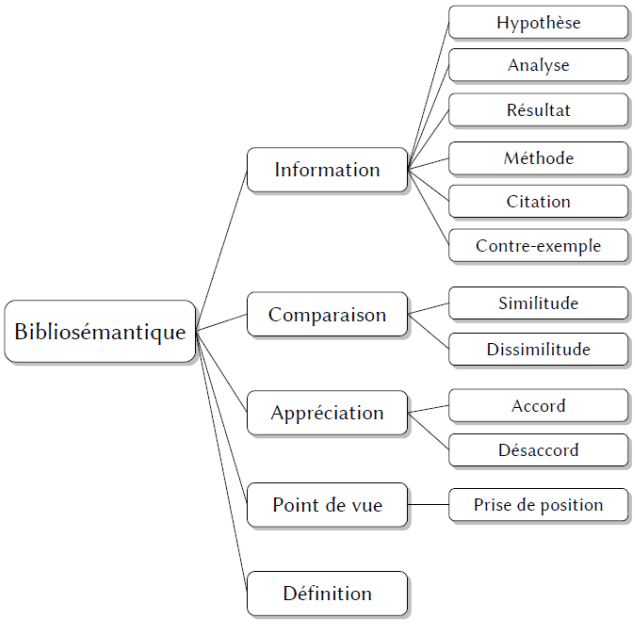

FIG. 2 – *Ontologie de la Bibliosémantique*

19. fr.utf8.lalic.these.Valliez.Chap3.txt.xml (These-Valliez)
A partir de ce principe, défini comme un « processus informatique de changements de représentations qui crée des représentations intermédiaires à différents niveaux » [DES 96, p. 105], fut proposée dans [Abraham, Desclés 92] une architecture cognitive d'interprétation sémantique des textes par des représentations iconiques (Fig. 111.2). **(definition, résultat)**

20. fr.utf8.lalic.these.pascu.Chap2.txt.xml (These-Pascu)
L'idée de définir le quantificateur comme un opérateur est reprise plus tard et indépendamment par Curry [Cur58] . **(definition, similitude)**

21. fr.utf8.alsic.2006.v9.ADEN.txt.xml (ALSIC2006)
Deyrich ([Deyrich05] : 127) rappelle la notion " d'illusion d'optique sociale" introduite par Martinon [Martinon96] par laquelle les sujets perçoivent les discours multimédiatiques à partir de leurs référents culturels, en effet, la mise en réseau des informations collectées reste, le plus souvent, ethnocentrée. **(definition, citation)**

FIG. 3 – *Phrases annotées : définition*

18. TheseIanaFinale.txttest.xml (Erudit)
Ces techniques introduisent un certain biais par rapport aux résultats de recherche et la visibilité des pages [Samier et al., 2007] : ce n'est plus le contenu d'une page qui est le principal facteur de son positionnement, mais l'expertise technique du créateur du site et l'argent investi dans la maintenance [Weideman, 2004]. **(information, prise de position)**
Fiche de synthèse

19. fr.utf8.intellectica.2002.Vol1.Num34_11_Lecuyer.txt.xml (Intellectica1991-2002)
Parfois même, c'est un paradigme expérimental précis qui est contesté (Haith, 1998; Cashon & Cohen (2000) . **(prise de position)**
Fiche de synthèse

20. fr.utf8.lalic.these.Bertin.Complete5.txt.xml (These-Marc)
Les principales critiques ont été exprimées par [Seglen, 1997] ,[Amin et Mabe, 2003], [Monastersky, 2005], [Ewing, 2006], [Adler, 2007], [Hall, 2007]. **(prise de position)**
Fiche de synthèse

FIG. 4 – *Phrases annotées : prise de position*

3 Ordonnancement des réponses

3.1 Méthodes existantes

La pertinence des résultats dans une grande partie des moteurs de recherche est modélisée par une valeur numérique qui se calcule dans un premier temps à partir des mesures de similarité entre le document et la requête. Les mesures et les méthodes de calcul dépendent du modèle de recherche d'information employé. L'estimation de la pertinence obtenue, constituant le *score de contenu* Baeza-Yates et Ribeiro-Neto (1999), est calculée en temps réel puisqu'elle est dépendante de la requête. Parmi les fonctions d'ordonnancement simples, nous pouvons citer la fréquence des termes, représentant la somme des nombres d'occurrences des termes de la requête dans le document, ainsi que la mesure *tf-idf*. En se basant sur la mesure *tf-idf*, Robertson et al. (1995); Cummins et ORiordan (2006) définissent la fonction *Okapi-BM25* dans le cadre du modèle probabiliste.

Les scores indépendants de la requête sont liés à la nécessité de développer des moteurs de recherche efficaces sur le Web. Ces nouveaux scores peuvent rendre compte de l'importance et de la fiabilité d'un document dans un corpus hyper-textuel. Plusieurs alternatives aux scores de contenu ont été proposées comme celles s'intéressant à la structure du graphe hypertexte et les relations entre les documents au sein de la collectionKleinberg (1999); Page et al. (1998);

Baeza-Yates et al. (2006); Krishnan et Raj (2006). Ces nouveaux scores, appelés également *scores d'importance*, sont calculés hors ligne et expriment une pertinence relative du document pour une requête quelconque. Deux types d'information sont alors utilisées : la structure du corpus, exprimée par le graphe hyper-texte, et certains aspects du comportement des internautes. Les scores *PageRank* Page et al. (1998) et *HITS* Kleinberg (1999) mesure la popularité des pages web à partir des hyperliens. Des essais ont été faits pour améliorer l'efficacité du score PageRank en prenant en compte de nouvelles informations. Par exemple, Haveliwala (2003) propose *Topic-sensitive PageRank* et Richardson et Domingos (2002) introduit une variante de PageRank dépendante de la requête dans un modèle probabiliste. Le score *TrustRank* Gyöngyi et al. (2004) est une estimation de la fiabilité des pages, afin de détecter le spam, calculé de façon semi-automatique à partir d'un petit ensemble de pages contrôlées manuellement. Le score de ces pages est propagé à travers le graphe par les hyper-liens. Afin de pallier certains biais introduits par la structure hyper-textuelle, d'autres scores comme BrowseRank Liu et al. (2008) et TrafficRank Tomlin (2003) prennent en compte l'historique de la navigation web et considèrent le temps et la fréquence des visites effectuées par les internautes.

Récemment la problématique de l'ordonnancement a été traitée par des techniques d'apprentissage automatique. La pertinence étant une notion complexe et multi-dimensionnelle Saracevic (1970, 1996); Denos (1997); Simonnot (2002), l'utilisation de ces méthodes est motivée par le fait que l'apprentissage pourrait théoriquement rendre compte de multiples critères d'ordonnancement, sans pour autant devoir analyser et modéliser tous les phénomènes autour de la pertinence.

Étant donnée une collection de documents, l'apprentissage automatique des ordonnancements suppose l'existence d'un corpus d'entraînement, qui consiste en un ensemble de requêtes et un ensemble de documents jugés pertinents pour chacune des requêtes ainsi que les valeurs du score de pertinence. La première phase d'entraînement consiste à créer une fonction d'ordonnancement qui a la propriété de produire exactement toutes les listes ordonnées du corpus d'entraînement. Dans une deuxième phase de test, cette fonction est utilisée afin d'ordonner des réponses de nouvelles requêtes. La mise en uvre des algorithmes d'apprentissage des ordonnancements est confrontée à une difficulté importante, qui est la définition des critères d'apprentissage pertinents pour l'optimisation des mesures d'évaluation. En effet, les mesures telles que la précision moyenne MAP et $nDCG$, sont difficiles à optimiser directement. Une première approche a été proposée par Cohen et al. (1999), qui considèrent l'ordre relatif des documents en utilisant une fonction binaire de préférence entre chaque couple de documents. Cette idée a été exploitée par d'autres travaux, qui optimisent le rang moyen des documents pertinents en minimisant les erreurs dans la fonction de préférence Freund et al. (2003); Cao et al. (2006); Usunier (2006). Les approches plus récentes se focalisent sur l'optimisation des erreurs en début de la liste, par exemple Volkovs et Zemel (2009). Nous pouvons également citer AquaLog Lopez et al. (2005, 2007) qui est un système de question/réponse reposant sur un moteur d'inférence et utilisant une ontologie.

3.2 Une nouvelle méthode s'appuyant sur une ontologie discursive

Les annotations sémantiques dans le corpus reflètent en partie le contenu sémantique des segments annotés. Ainsi, nous considérons que ces annotations portent des informations sur la pertinence des segments, surtout lorsque la recherche s'effectue selon des requêtes utilisant les catégories d'annotation Desclés et Djioua (2006); Atanassova et al. (2008).

L'unité textuelle de base dans cette approche est la phrase : tous les documents sont segmentés en phrases qui constituent un contexte minimal pour le traitement. Une phrase est constituée d'un contenu textuel (une suite de termes) et des caractéristiques sémantiques qui sont attribuées à une partie ou à la totalité de la phrase. De même, les requêtes sont constituées de deux éléments non obligatoires : une suite de termes, reliés éventuellement avec des opérateurs logiques, et un ensemble de catégories sémantiques sélectionnés à partir de l'ontologie.

Soit une ontologie qui s'exprime par un graphe $C = (V, A)$, où V est l'ensemble des nuds et A est l'ensemble des arcs. Nous pouvons alors formuler les deux hypothèses suivantes :

Hypothèse 1 : Une phrase $p = (T_p, R_p)$ est pertinente pour la requête $q = (T_q, R_q)$, si $d(T_q, T_p) < \alpha$ et $R_q \subset R_p$, dans le cas où l'ensemble R_q est non-vide,
où T_p est le texte du segment p, T_q est le texte du segment q, $R_p \subset V$ est l'ensemble des catégories d'annotation de p, $R_q \subset V$ est l'ensemble des catégories d'annotation de q, et d est une distance définie entre segments textuels.

Hypothèse 2 : Les pertinences relatives des phrases annotées sont déterminées en partie par les relations annotées.

La pertinence peut être aussi évaluée en termes de proximité entre les relations annotées et le besoin exprimé par la requête. L'ontologie sous-jacente définit une organisation hiérarchique entre les catégories, en utilisant des relations de spécification et généralisation. Nous proposons des critères d'ordonnancement des réponses utilisant cette structure, permettant ainsi de définir un score d'annotation tenant compte des types de relations exprimées dans les phrases.

3.2.1 Position des catégories dans la hiérarchie

L'organisation des catégories dans l'ontologie est liée à une organisation des marqueurs linguistiques sous-jacents. Considérons un couple de catégories $(pdv_1, pdv_2) \in A$, où pdv_2 est plus spécifique que pdv_1 dans l'ontologie. Nous pouvons formuler les observations suivantes :

- Si une phrase a été annotée par pdv_1 et non par pdv_2, cela signifie qu'elle contient les marqueurs linguistiques permettant d'attribuer l'annotation pdv_1. Comme l'annotation par pdv_2 n'a pas été déclenchée, la phrase ne contient pas assez de marqueurs linguistiques permettant d'affiner l'annotation.
- L'annotation par pdv_2 implique l'annotation par pdv_1, puisque $(pdv_1, pdv_2) \in A$, c'est-à-dire pdv_2 est une spécification de pdv_1. Cela signifie qu'une phrase annotée par pdv_2 contient suffisamment de marqueurs linguistiques, indicateurs désambiguïsés par des indices contextuels, permettant de lui attribuer cette annotation ainsi que l'annotation par pdv_1.

Nous pouvons énoncer l'hypothèse suivante :

Hypothèse 3 : Si $(pdv_1, pdv_2) \in A$, une phrase annotée par pdv_2 est plus pertinente qu'une phrase annotée par pdv_1.

Pour un segment annoté $p = (T_p, R_p)$ nous pouvons alors définir le score suivant :

$$SPos(p) = \frac{1}{M}. \max_{pdv \in R_p} \{N(pdv)\}, \quad (1)$$

où M est le niveau maximal de l'ontologie, c'est-à-dire le niveau contenant les catégories les plus spécifiques $M = \max_{pdv \in V}\{N(pdv)\}$, et la fonction $N : R_p \rightarrow \mathbb{N}$ associe à chaque catégorie son niveau dans la structure de l'ontologie, la racine ayant le niveau 1. Le score $SPos$ représente le niveau de spécificité de l'annotation sémantique d'une phrase.

$$N(pdv) = \begin{cases} 1, & \text{si} \quad \nexists pdv_1 \in V \text{ t.q. } (pdv_1, pdv) \in A \\ k+1, & \text{si} \quad \exists (pdv_1, pdv_2, \ldots, pdv_k) \in V^k \text{ t.q. } (pdv_k, pdv) \in A \\ & \quad \text{et } (pdv_{i-1}, pdv_i) \in A, i = 2, \ldots, k. \end{cases} \tag{2}$$

Le score $SPos$ prend des valeurs entre 0 et 1, où 1 correspond au niveau le plus spécifique de la carte sémantique.

3.2.2 Annotations multiples

L'annotation multiple d'un segment textuel indique la présence de plusieurs relations sémantiques, identifiées pendant la phase d'annotation. Nous parlerons d'annotations multiples uniquement dans le cas où il n'existe pas de relation de spécification/généralisation entre les catégories d'annotation. Leur annotation s'effectue de façon indépendante. Un segment qui est annoté plusieurs fois cumule l'ensemble des relations sémantiques annotées grâce à la présence des marqueurs linguistiques permettant d'identifier chacune de ces relations. Aussi, nous supposons que la pertinence d'une phrase annotée augmente avec le nombre d'annotations qu'elle porte.

Pour un segment $p = (T_p, R_p)$, nous pouvons définir l'ensemble de ses annotations indépendantes dans $C = (V, A)$ de la façon suivante :

$$S_p = \{pdv \mid (pdv \in R_p) \ \& \ (\nexists pdv'(pdv' \in R_p) \& ((pdv, pdv') \in A))\} \tag{3}$$

Nous pouvons formuler l'hypothèse suivante :

Hypothèse 4 : La pertinence d'une phrase p est proportionnelle à la taille de S_p.

Nous définissons le score lié aux annotations multiples de la manière suivante :

$$SNum(p) = \sum_{pdv \in S_p} \frac{1}{M} \{N(pdv)\} \tag{4}$$

Le score $SNum$ représente la diversité de l'annotation sémantique au sein de l'ontologie.

3.2.3 Types de marqueurs linguistiques

Par définition, l'annotation d'un segment textuel s'appuie sur plusieurs marqueurs linguistiques, puisque l'occurrence d'un simple marqueur peut être polysémique. La nécessité d'identifier un grand nombre de marqueurs dans un segment traduit le fait que chacun de ces marqueurs est polysémique et ne suffit par pour attribuer l'annotation. En effet, si une règle d'annotation nécessite un seul marqueur pour effectuer l'annotation, nous pouvons considérer qu'il s'agit d'un marqueur qui, en plus d'être porteur de la sémantique de la relation de l'annotation, est non polysémique. Nous appelons ce type de marqueurs *forts*. Au contraire, si la phrase ne contient pas de marqueur fort, son annotation s'appuie sur plusieurs marqueurs, chacun entre eux étant polysémique. Dans ce deuxième cas, nous pouvons considérer que la phrase est moins pertinente qu'une phrase contenant un marqueur fort, ce que nous formulons dans l'hypothèse suivante :

Hypothèse 5 : Le nombre de marqueurs linguistiques ayant déclenché l'annotation d'une phrase est inversement proportionnel à la pertinence de la phrase.

Nous définissons le score suivant :

$$SType(p) = \sum_{pdv \in R_p} \frac{1}{K(pdv) + 1}, \tag{5}$$

où la fonction $K : R_p \rightarrow \mathbb{N}$ donne le nombre d'indices contextuels qui ont été utilisés. Le score $SType$ prend des valeurs réelles entre 0 et 1 pour une annotation simple.

3.2.4 Score d'annotation

Nous constatons que les trois scores $SPos$, $SNum$ et $SType$ que nous avons définis utilisent en partie des propriétés communes des segments annotés, comme le nombre d'annotations différentes. Par conséquent, les trois valeurs attribuées à un segment donné ne sont pas entièrement indépendantes. Afin d'exprimer les relations entre ces scores, considérons un segment p annoté par N_p points de vue. Nous avons les inégalités suivantes, nous avons les inégalités suivantes :

$$\left| \begin{array}{l} 0 < SPos(p) \leq 1 \\ 0 < SPos(p) \leq SNum(p) \leq N_p \\ 0 < SType(p) \leq N_p \end{array} \right. \tag{6}$$

Nous définissons une fonction d'ordonnancement qui combine les trois scores définis ci-dessus. Cette fonction est définie de façon à ce que le résultat de l'ordonnancement soit identique à l'application consécutive des scores $SPos$, $SNum$, $SType$. Nous l'appellerons *score d'annotation* puisqu'il s'agit d'un score issu principalement de l'annotation sémantique et qui traduit donc le degré de pertinence des phrases selon les annotations par l'ontologie linguistique. L'adéquation du score d'annotation ainsi défini se confirme par l'évaluation (voir la section 5).

La fonction d'ordonnancement peut être donnée sous la forme :

$$SAnnot(p) = \alpha_1 SPos(p) + \alpha_2 SNum(p) + \alpha_3 SType(p) \tag{7}$$

Le vecteur $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ définit une pondération entre les trois scores $SPos$, $SNum$ et $SType$, et dans un cas particulier, permet de définir un ordre d'application de ces critères. Les coefficients $\alpha_i \in]0, 1]$ sont fixés de façon à ce que :

$$\left| \begin{array}{ll} (SPos(p_1) > SPos(p_2)) & \Rightarrow (SAnnot(p_1) > SAnnot(p_2)) \\ (SPos(p_1) = SPos(p_2)) \& (SNum(p_1) > SNum(p_2)) & \Rightarrow (SAnnot(p_1) > SAnnot(p_2)) \end{array} \right. \tag{8}$$

Des valeurs possibles pour les coefficients α_i sont : $\alpha = (1; 0,25; 0,04)$. Ce vecteur satisfait les conditions (8) pour l'ontologie que nous étudions.

Le score d'annotation est calculé hors ligne et reste indépendant de la requête. Ce dernier peut alors être considéré comme un score d'importance. Cependant, plusieurs propriétés le différencient des autres scores d'importance existants. Premièrement, il ne dépend ni de la structure globale du corpus, ni du comportement de l'ensemble des utilisateurs. Deuxièmement, le score est issu indirectement du contenu textuel en se basant sur une annotation qui explicite des relations sémantiques présentes dans des textes. Dans ce sens, il se rapproche plus aux

scores de contenu, qui expriment une proximité entre la requête et les documents selon leurs contenus. Pour le calcul de notre score, l'évaluation d'un document dépend uniquement de son contenu et non pas du reste du corpus. Troisièmement, le score d'annotation s'appuie sur des ressources externes : il exploite, d'une part, les annotations automatiques, et d'autre part, la structure de l'ontologie linguistique qui encode certaines connaissances linguistiques.

4 Système

Nous nous intéressons ainsi au système informatique, permettant l'indexation de documents textuels structurés au format DocBook et annotés par des catégories discursives, proposant des traitements plein texte et des fonctionnalités de gestion des contenus, d'extraction et de génération de documents secondaires dans une perspective de recherche d'information. Le schéma de la figure 5 présente les principales étapes du traitement automatique. Le moteur d'annotation sémantique Excom Djioua et al. (2006); Alrahabi et Desclés (2008) produit des fichiers annotés en format XML qui sont ensuite indexés à l'aide d'une base de données dédiée. Les segments textuels annotés sont indexés en plein texte.

L'annotation par la méthode d'Exploration Contextuelle s'appuie sur des ressources linguistiques (les marqueurs et l'ontologie linguistique) qui sont également utilisées pour structurer les interfaces d'accès à l'information, offrant à l'utilisateur la possibilité de formuler des requêtes selon catégories discursives. De plus, l'organisation de l'ontologie et une partie des marqueurs sont utilisés par les algorithmes d'ordonnancement des réponses que nous avons mis en place spécifiquement pour ce type de systèmes.

L'interface, élaborée en PHP/AJAX, offre des fonctionnalités de recherche d'information et de navigation textuelle en se basant sur les annotations sémantiques. L'utilisateur peut visualiser les résultats sous forme de phrases extraites, à partir desquelles il peut accéder aux contextes des phrases, aux documents sources, ou à des nouvelles représentations des documents telles que des extractions utilisant les annotations et des bibliographies enrichies permettant de retrouver les segments textuels liés à chaque référence.

5 Évaluation

Notre objectif est d'évaluer l'algorithme d'ordonnancement parmi les dix premiers résultats de chaque requête. Notons que cette évaluation n'a pas pour but d'évaluer les performances globales du système. Une évaluation complète serait plus coûteuse du fait qu'un très grand nombre de résultats doit être examiné manuellement. Seules les phrases qui figurent parmi les résultats des requêtes ont été évaluées, ce qui rend l'évaluation inadéquate pour estimer le rappel.

L'évaluation consiste en deux étapes :

1. Établir les pertinences de référence, correspondantes aux attentes des utilisateurs pour les résultats d'un ensemble de requêtes. Pour cela, nous avons utilisé des jugements de pertinence attribués manuellement aux résultats du système par trois juges humains.
2. Comparer les résultats du système avec les valeurs de pertinence de référence afin de calculer les valeurs $P@i$, $nDCG$, ainsi que la précision moyenne MAP.

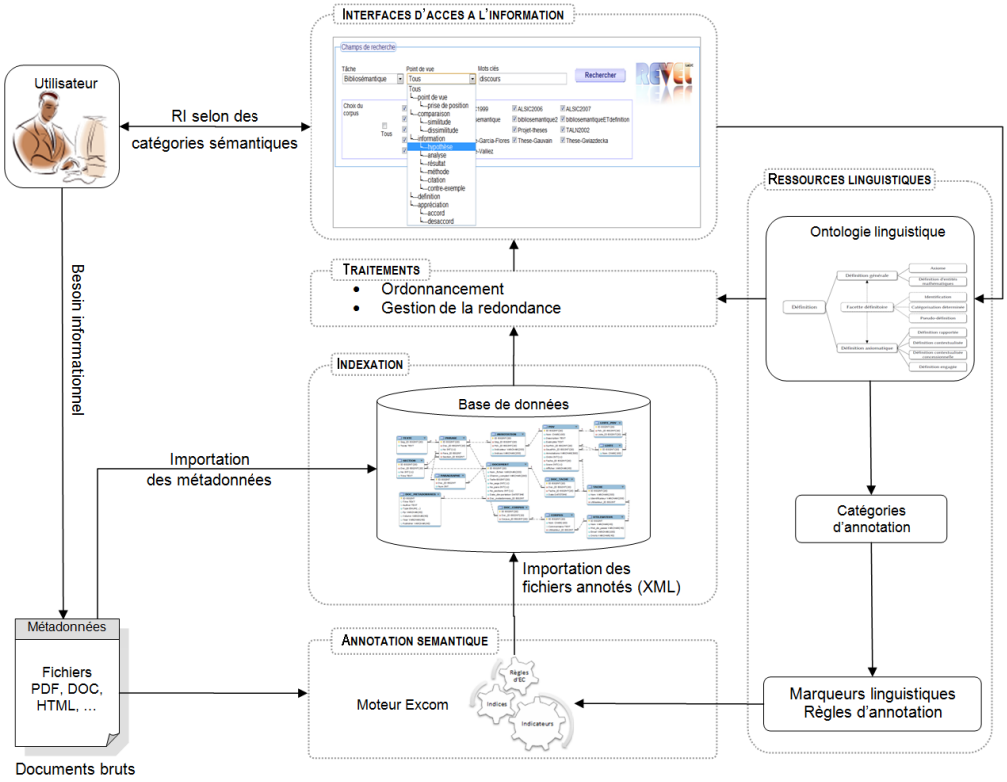

FIG. 5 – *Schéma général du traitement automatique*

5.1 Corpus

Pour tester le système nous avons utilisé un corpus de 330 documents de différentes tailles. Ce corpus comprend 8 thèses de l'Université Paris-Sorbonne, ainsi que des articles scientifiques issus de TALN 2002, Intellectica 1991-2002, ALSIC 1998, 1999, 2006 et 2007. La table 1 présente quelques statistiques du corpus traité. Tous les documents ont été annotés de façon automatique.

Phrases	Nb de mots	Mots par phrase	Phrases annotées	Pourcentage
359 354	7 890 558	21,96	8 672	2,41%

TAB. 1 – *Corpus*

Les annotations sémantiques représentent autour de 2,41% de tous les segments textuels. Cette annotation constitue un premier filtrage permettant d'identifier les phrases pertinentes pour l'analyse des citations bibliographiques. Pour la recherche d'information, nous considérerons que les réponses pertinentes se trouvent parmi les phrases annotées car elles sont porteuses des relations sémantiques recherchées, ce qui nous permet d'éliminer une grande partie du bruit parmi les résultats de la recherche d'information.

5.2 Protocole

Le corpus de requêtes pour l'évaluation consiste en 7 requêtes, où chaque requête est constituée d'une catégorie sémantique de recherche et un ou plusieurs mots clés. Pour ces requêtes, parmi toutes les phrases contenant les mots clés dans le corpus, le pourcentage des phrases annotées par les catégories recherchées varie entre 0,27% et 2,23%, ce qui montre que l'annotation permet dans un premier temps d'élire un ensemble très restreint de phrases pertinentes.

Les 10 premiers résultats de chaque requêtes ont été évalués par trois juges indépendants. Les jugements prennent trois valeurs possibles : 1 (*pertinent*), 0,5 (*peu pertinent*), ou 0 (*non pertinent*). Les résultats du système qui sont présentés aux juges sont ordonnés de façon aléatoire. Cet ordre est différent pour chacun des juges afin de s'assurer qu'il n'introduit pas de biais dans l'estimation de la pertinence. En dehors des étiquettes des annotations sémantiques et les contenus textuels des phrases, les juges n'ont aucune indication de l'importance ou de la qualité d'une phrase : ils n'ont pas accès aux valeurs des scores attribuées par le système, ni à l'ordonnancement qui en résulte.

Dans un premier temps, nous cherchons à estimer les valeurs de pertinence de référence, à partir des trois jugements pour chaque résultat. Pour un résultat r ayant obtenu les jugements $j_1(r)$, $j_2(r)$ et $j_3(r)$, nous considérons la moyenne :

$$Pert(r) = \frac{1}{3}\sum_{i=1}^{3} j_i(r) \tag{9}$$

Les valeurs $Pert(r)$ indiquent l'appréciation de la pertinence d'un résultat par l'ensemble des juges. Cette approche nous permet de modéliser la pertinence par une échelle à plusieurs valeurs entre 0 et 1, afin de mieux rendre compte du phénomène de pertinence. En effet, Kekäläinen et Järvelin (2002) soulignent que les jugements binaires sont à éviter pour les évaluations de la pertinence puisqu'ils ne peuvent refléter la variabilité et la complexité de celle-ci. Ils suggèrent que les jugements de pertinence soient plutôt modélisés par un ensemble de valeurs continu et proposent une généralisation des mesures de précision et de rappel pour de tels jugements.

Par la suite, étant donné un résultat r, nous considérerons que sa pertinence est représentée par la valeur $Pert(r) \in [0, 1]$. Nous allons considérer que r est non-pertinent si $Pert(r) = 0$.

5.3 Résultats

Pour chaque requête nous avons considéré l'ordonnancement des réponses par le système et nous avons calculé les valeurs $P@i$ ainsi que les valeurs $nDCG(i)$, où $i = 1, ..., 10$ est la position dans la liste produite par le système.

$$\begin{aligned} P@j &= \frac{1}{j}\sum_{i=1}^{j} Pert(r_i) \\ AP &= \frac{\sum_{j\in J} P@j}{|J|}, \text{ où } J = \{k \in [1;10] | Pert(r_k) > 0\} \end{aligned} \tag{10}$$

Les précisions $P@j$ indiquent la précision en considérant la tranche des j premiers résultats, et AP donne une précision moyenne sur toute la liste ordonnée, en pénalisant d'avantage le bruit en début de la liste.

Nous avons également considéré la mesure $nDCG$ (*Normalized Discounted Cumulative Gain*) Järvelin et Kekäläinen (2002) qui utilise l'hypothèse que la pertinence diminue de façon logarithmique par rapport à la position dans la liste. La figure 6 donne les valeurs $P@i$ et $nDCG(i)$ obtenues à partir de l'ensemble des requêtes évaluées. L'abscisse représente le numéro de la réponse et l'ordonnée représente les valeurs pour les deux types de mesures.

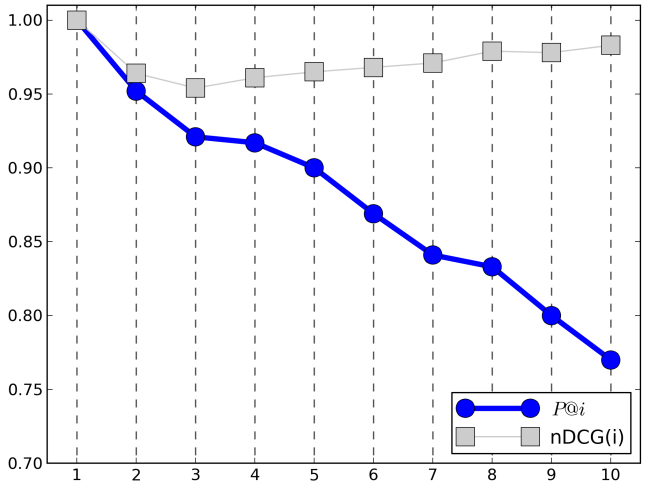

FIG. 6 – *Valeurs $P@i$ et $nDCG(i)$ pour les 10 permiers résultats*

Pour la précision moyenne MAP, nous avons obtenu :

$$MAP = \frac{1}{7}\sum_{i=1}^{7} AP(q_i) = 0,893. \tag{11}$$

6 Conclusion

Nous avons proposé un nouveau algorithme d'ordonnancement en nous appuyant sur la structure d'une ontologie discursive et sur l'annotation sémantique. L'évaluation montre que cet algorithme est très performant en ce qui concerne les premiers résultats fournis par le système. En effet, la figure 6 montre bien que le système a une tendance prononcée à classer les résultats les plus pertinents en début de la liste, et donc d'ordonnancer les résultats dans un ordre proche de celui des juges humains.

Cette première évaluation est encourageante, même si nous sommes conscients qu'il sera nécessaire de confirmer ce résultat à une plus grande échelle. Les valeurs obtenues pour la précision moyenne, notamment $MAP = 0,893$, confirment que l'annotation sémantique apporte une meilleure qualité à la recherche d'information. Au-delà de l'algorithme d'ordonnance-

ment, les valeurs des précisions très élevées reflètent également la précision des annotations sémantiques.

Références

Alrahabi, M. et J.-P. Desclés (2008). Automatic annotation of direct reported speech in Arabic and French, according to a semantic map of ennunciative modalities. In *6th Internatilnal Conference of NLP, GOTAL*, Gothenburg, Sweden.

Atanassova, I. (2012). *Exploitation informatique des annotations sémantiques automatiques d'Excom pour la recherche d'informations et la navigation*. Ph. D. thesis, Université Paris-Sorbonne.

Atanassova, I., J.-P. Desclés, A. Franchi, et F. Le Priol (2008). La plate-forme excom comme outil automatique d'annotations sémantiques des textes pour la catégorisation des informations sur le web. In *Colloque ń Internet : besoin de communiquer autrement ż, Université St. Clément d'Ohride*, Sofia, Bulgarie.

Baeza-Yates, R., P. Boldi, et C. Castillo (2006). Generalizing PageRank : damping functions for link-based ranking algorithms. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval 1*, 308–315.

Baeza-Yates, R. et B. Ribeiro-Neto (1999). *Modern Information Retrieval* (1^e ed.). Addison Wesley.

Bertin, M. (2008). Categorizations and annotations of citation in research evaluation. In *FLAIRS*, Coconut Grove, Floride. AAAI Press.

Bertin, M. (2011). *Bibliosématique : une technique linguistique et informatique par exploration contextuelle*. Ph. D. thesis, Université Paris-Sorbonne.

Cao, Y., J. Xu, T. Liu, H. Li, Y. Huang, et H. Hon (2006). Adapting ranking SVM to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in IR*, pp. 186–193.

Cohen, W., R. Schapire, et Y. Singer (1999). Learning to order things. *Journal of Artificial Intelligence Research 10*, 243–270.

Cummins, R. et C. ORiordan (2006). Evolving local and global weighting schemes in information retrieval. *Information Retrieval 9*(3), 311–330.

Denos, N. (1997). *Modélisation de la pertinence en recherche d'information : modèle conceptuel, formalisation et application*. Ph. D. thesis, Université de Grenoble 1.

Desclés, J.-P., E. Cartier, A. Jackiewicz, et J.-L. Minel (1997). Textual Processing and Contextual Exploration Method. *CONTEXT'97, Rio de Janeiro 1*, 189–197.

Desclés, J.-P. et B. Djioua (2006). Machines d'annotation et d'indexation discursives de textes : EXCOM/MOCXE. In *Annotation automatique de relations sémantiques et recherche d'informations : vers de nouveaux accès aux savoirs*, Paris.

Djioua, B., J. G. Flores, A. Blais, J.-P. Desclés, G. Guibert, A. Jackiewicz, F. Le Priol, L. Nait-Baha, et B. Sauzay (2006). Excom : an automatic annotation engine for semantic information. *The 19th international FLAIRS Conference, Melbourne, Floride 1*, 285–290.

Freund, Y., R. Iyer, R. Schapire, et Y. Singer (2003). An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research 4*, 933–969.

Gyöngyi, Z., H. Garcia-Molina, et J. Pedersen (2004). Combating web spam with trustrank. In *Proceedings of the 30th international conference on Very large data bases*, Volume 30, pp. 576–587.

Haveliwala, T. (2003). Topic-sensitive pagerank : A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 784–796.

Järvelin, K. et J. Kekäläinen (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS) 20*(4), 422–446.

Kekäläinen, J. et K. Järvelin (2002). Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology 53*(13), 1120–1129.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM 46*(5), 604–632.

Krishnan, V. et R. Raj (2006). Web spam detection with anti-trust rank. *2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

Liu, Y., B. Gao, T. Liu, Y. Zhang, Z. Ma, S. He, et H. Li (2008). Browserank : letting web users vote for page importance. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 451–458.

Lopez, V., M. Pasin, et E. Motta (2005). Aqualog : An ontology-portable question answering system for the semantic web. *The Semantic Web : Research and Applications*, 135–166.

Lopez, V., V. Uren, E. Motta, et M. Pasin (2007). Aqualog : An ontology-driven question answering system for organizational semantic intranets. *Web Semantics : Science, Services and Agents on the World Wide Web 5*(2), 72–105.

Page, L., S. Brin, R. Motwani, et T. Winograd (1998). The pagerank citation ranking : Bringing order to the web. *Stanford InfoLab*.

Richardson, M. et P. Domingos (2002). The intelligent surfer : Probabilistic combination of link and content information in pagerank. *Advances in neural information processing systems 2*, 1441–1448.

Robertson, S., S. Walker, S. Jones, M. Hancock-Beaulieu, et M. Gatford (1995). Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 109–126. NIST.

Saracevic, T. (1970). The concept of "relevance" in information science : A historical review. *Introduction to information science*, 111–151.

Saracevic, T. (1996). Relevance reconsidered. In *Proceedings of the 2nd Conference on Conceptions of Library and Information Science*, pp. 201–218.

Simonnot, B. (2002). De la pertinence à l'utilité en recherche d'information : le cas du Web. In *Recherches récentes en sciences de l'information : convergences et dynamiques, Actes du colloque Mics-Lerass*, pp. 393–410.

Tomlin, J. (2003). A new paradigm for ranking pages on the world wide web. In *Proceedings of the 12th international conference on World Wide Web*, pp. 350–355. ACM.

Usunier, N. (2006). *Apprentissage de fonctions d'ordonnancement : une étude théorique de*

la réduction à la classification et deux applications à la Recherche d'Information. Ph. D. thesis, Université Paris-VI.

Volkovs, M. et R. Zemel (2009). Boltzrank : learning to maximize expected ranking gain. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1089–1096. ACM.

Summary

We have developed an information retrieval system based on the automatic annotation using a discursive ontology. The system makes possible information retrieval of specific semantic categories related to bibliographic citations in scientific publications. We propose new ranking criteria which exploit the annotation and the structure of the ontology. These criteria take into account the specificity and the diversity of the annotations, as well as the different types of linguistic markers. We have defined a ranking function and we have evaluated it by a comparison between the system output and human relevance judgments. This evaluation confirms the algorithm efficiency for the ranking of highly relevant results.

Matching Fusion with Conceptual Indexing

Karam Abdulahhad*, Jean-Pierre Chevallet**, Catherine Berrut***

* UJF-Grenoble 1, LIG laboratory, MRIM group
karam.abdulahhad@imag.fr
** UPMF-Grenoble 2, LIG laboratory, MRIM group
jean-pierre.chevallet@imag.fr
*** UJF-Grenoble 1, LIG laboratory, MRIM group
catherine.berrut@imag.fr

Abstract. Many studies have been addressed the term-mismatch problem, which arises when using different terms or words for expressing the same meaning. We also introduce another problem: over-specialized document, which is caused when IR systems prefer documents that have poor query-document intersection, but with high weighting value, to those that have rich query-document intersection with low weighting value. In this study, we propose to use, simultaneously, multiple types of indexing elements: ngrams, keywords, and concepts, instead of only keywords. We followed a late data-fusion technique to achieve that. Through our proposed model, we also try to overcome the over-specialized document problem. Experiments for model validation have been done by using ImageCLEF2011 test collection, UMLS2009 Meta-thesaurus, and MetaMap tool for mapping text into UMLS concepts.

1 Introduction

Two terms or words could have the same meaning in a specific context. For example, (atrial, auricular), (apartment, flat), (air pollution, pollution of the air), etc. This is one of the preferable features of the natural languages, and one of features that gives each author the ability to have her/his own writing style. However, in IR field, it is a problematic feature, because the most of IR systems use a type of query-document intersection. Therefore, by using different terms, in queries and documents, for expressing the same meaning, IR systems will not be able to retrieve relevant documents. This problem is well studied in literatures and is called "**term-mismatch**" problem Woods (1997) Crestani (2000) Baziz (2005) Maisonnasse (2008) Chevallet (2009).

Most of IR systems use a type of weighting for estimating the amount of contribution of an indexing element in the overall matching, and subsequently for ranking the retrieved documents. There are many weighting formulas Harter (1975a) Harter (1975b) Robertson and Walker (1994) Lee (1995) Amati and Van Rijsbergen (2002), each with its properties. However, using element weighting in IR systems poses a problem. In general, IR systems can not warrant that documents, which share more number of distinct elements with queries, will be ranked higher than other documents. In other words, documents that have less shared elements with

queries, but with high weighting values, could be ranked higher than documents that have more shared elements, but with low weighting values. For us, this is inconvenient behavior and it is better to retrieve documents that cover more aspects of a query, even with a low weighting values. We will call this problem an "**over-specialized documents**".

In this study, we address these two problems and we try to find a practical and effective solution for them.

The paper will be organized as follow: in section 2, we present some related works. In section 3, we describe in details our proposed model. Section 4 presents some experiments for model validation, then we discuss our results. We conclude in section 5.

2 Related Works

Before we go forward, we should define terms and concepts. A term is a sequence of words that have a unique meaning in a specific domain Chevallet (2009), whereas concepts could be defined as: "*Human understandable unique abstract notions independent from any direct material support, independent from any language or information representation, and used to organize perception and knowledge*" Chevallet et al. (2007). Practically, each concept is represented by an identifier in an external resource and is associated with a set of synonym terms Baziz (2005) Chevallet (2009).

The term-mismatch problem was heavily studied by multiple researchers. In literatures, several approaches, to solve this problem, could be identified:

1. Dimensionality reduction: reduces the chance that a query and a document use different terms for representing the same meaning. Among the techniques that are used for achieving this mission, we can mention: Stemming Frakes (1992), Latent Semantic Indexing (LSI) Deerwester (1988) Deerwester et al. (1990), and Conceptual Indexing (using concepts instead of terms) Chevallet et al. (2007).
2. Query expansion: extends the query with new terms to increase the chance of matching with documents Efthimiadis (1996).
3. Using term-term semantic similarity measures: this approach presupposes the existence of a measure capable of estimating the similarity between any two terms Crestani (2000).

$$\forall t_i, t_j \in T, \quad 0 \leq Sim\left(t_i, t_j\right) \leq 1 \tag{1}$$

In our previous studies Abdulahhad et al. (2011b) Abdulahhad et al. (2011a), we used concepts as a solution for the term-mismatch problem. For example, the two terms '*Atrial Fibrillation*' and '*Auricular Fibrillation*' correspond to the same concept '*C0004238*' in UMLS [1]. However, using concepts poses another problem: the **concept-mismatch** problem Abdulahhad et al. (2011b). An example about this problem could be: according to UMLS, the two terms '*Dermatofibroma*' and '*Dermatofibrosarcoma*' correspond to two different concepts '*C0002991*' and '*C0392784*', respectively. Therefore, even by using concepts, the mismatch between a document containing '*Dermatofibroma*' and a query containing '*Dermatofibrosarcoma*' still exist.

1. Unified Medical Language System. It is a meta-thesaurus in medical domain. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls

In addition, external resources, e.g. UMLS, that contain concepts are generally incomplete Bodenreider et al. (1998) Bodenreider et al. (2001) Abdulahhad et al. (2011b). For example, the term '*Osteoporotic*' does not map to any concept in UMLS2009.

Many approaches to solve the concept-mismatch problem could be found in literatures:

1. Exploiting semantic relations between concepts: especially the hyponymy-hypernymy relations Baziz (2005) Maisonnasse (2008) Le (2009) Abdulahhad et al. (2011b) Abdulahhad et al. (2011a).
2. Query expansion: by exploiting concepts, their content, and their position in the external resource Aronson and Rindflesch (1997) Baziz (2005).
3. Domain dimensions: indexing documents and queries by domain dimensions, which are more abstract than concepts Radhouani (2008).

Again in our previous studies Abdulahhad et al. (2011b) Abdulahhad et al. (2011a), we exploit semantic relations and enrich the external resource by new relations for solving the concept-mismatch and external resource incompleteness problems, respectively.

In this study, we use concepts, but we tried **data fusion** as another technique to solve, simultaneously, the two problems: concept-mismatch and external resources incompleteness.

Data fusion is the process of combining different result sets of a certain information need. Different result sets of the same information need, in the same corpus, could be produced by Croft (2000):

1. Using different IR systems.
2. Using the same IR system in conjunction with different configurations:
   (a) Different document representations using: different types of indexing elements, different parts of documents, different weighting schemas, etc.
   (b) Different queries of the same information need.
   (c) Different matching formulas.

Actually, in this study, we used three types of indexing elements: ngrams, words, and concepts. Through the following two examples, we will illustrate the reason of these choices.

Example 1: The query number 21 in the Image-based task of ImageCLEF2009[2] is: "*osteoporotic bone*". According to MetaMap[3], there is no concept in UMLS corresponds to the word '*osteoporotic*'. Using only concepts without ngrams and words, and as the word '*bone*' is an indiscriminative, the IR system will not be able to retrieve the relevant documents of this query.

Example 2: The query number 16 in the Image-based task of ImageCLEF2010[4] is: "*images of dermatofibroma*". The word '*dermatofibroma*' does not exist in the corpus, istead the corpus contains the word '*dermatofibrosarcoma*'. Therefore, and as the word '*images*' is an indiscriminative, using only words as indexing elements will not be sufficient. In addition, according to MetaMap, the word '*dermatofibroma*' maps into the concept '*C0002991*' and the word '*dermatofibrosarcoma*' maps into one of the concepts: '*C2697408*', '*C0206647*', or '*C0392784*'. By consulting the UMLS, we did not find any direct relation between the '*dermatofibroma*' concept and one of the '*dermatofibrosarcoma*' concepts. Consequently, and as

2. http://www.imageclef.org/2009
3. is a tool to map text into UMLS concepts. http://metamap.nlm.nih.gov/
4. http://www.imageclef.org/2010

neither the word '*images*' nor the concepts that correspond to it are discriminative, the two types of indexing elements (words and concepts) are not sufficient to retrieve relevant documents. That's what justify the usage of ngrams.

Concerning matching formulas, multiple heuristics could be found in IR literatures. The Fang et al. (2004) recalls some of those heuristics and transforms them to a set of constraints. Any matching formula should satisfy some of these constraints to be effective. Table (1) lists the constraints [5].

TAB. 1 – *The constraints*

Constraints	**Intuitions**
TFC1	to favor a document with more occurance of a query term
TFC2	to favor document matching more distinct query terms
TFC2	to make sure that the change in the score caused by increasing TF (Term Frequency) from 1 to 2 is larger than that caused by increasing TF from 100 to 101
TDC	to regulate the impact of TF and IDF (Inverse Document Frequency): it ensure that, given a fixed number of occurences of query terms, we should favor a document that has more occurences of discriminative terms (i.e. high IDF terms)
LNC1	to penalize a long document
LNC2, TF-LNC	to avoid over-penalize a long document: as it says that if we concatenate a document with itself k times to form a new document, then the score of the new document should not be lower than the original document
TF-LNC	to regulate the interaction of TF and document length: if d_1 is generated by adding more occurences of the query term to d_2, the score of d_1 should be higher than d_2

In our matching formula, we tried to meet some of these constraints, in order to build an effective formula.

3 The Proposed Model

3.1 Three Types of Indexing Elements

After the discussion in the previous section, We believe now that no single type of indexing elements could completely represent the content of documents and queries, because:

1. there is no perfect indexing function Baziz (2005) Aronson (2006) Dozier et al. (2007). It is always an approximate function.
2. concerning concepts, most resources, e.g. UMLS, are incomplete Bodenreider et al. (1998) Bodenreider et al. (2001) Abdulahhad et al. (2011b).

5. The presentation of the constraints is a rendering of the original presentation in Fang et al. (2004)

3. each type of indexing elements covers an aspect of documents and queries Das-Gupta and Katzer (1983). Ngrams cover the morphological aspect, words cover the lexical aspect, and concepts cover the conceptual aspect.

The goal of the indexing function is to convert documents and queries from their original form to another easy to use form. As we have three different types of elements: ngrams (NG), words (W), and concepts (C), three indexing functions are defined (one for each type).

$$Index_{NG} : D \cup Q \rightarrow E^*_{NG} \tag{2}$$

$$Index_{W} : D \cup Q \rightarrow E^*_{W} \tag{3}$$

$$Index_{C} : D \cup Q \rightarrow E^*_{C} \tag{4}$$

Where
D set of documents
Q set of queries
E_{NG} set of ngrams
E_W set of words
E_C set of concepts
E^* the set of all subsets of E

3.2 Matching Function

Our model, as almost all models, depends on some hypotheses. Actually, it depends on the following hypotheses:

1. The more shared elements a document and a query have, the more semantically closer they are. This hypothesis corresponds to the $TFC2$ constraint in Fang et al. (2004) (see Table 1).
2. The descriptive power of an element (local weight): the more frequently an element occurs in a document, the better it describes the document Luhn (1958) Baziz (2005). This hypothesis corresponds to the $TFC1$ constraint in Fang et al. (2004) (see Table 1).
3. The discriminative power of an element (global weight): the less number of documents an element appears in, the more important it is Luhn (1958) Baziz (2005). This hypothesis corresponds to the TDC constraint in Fang et al. (2004) (see Table 1).
4. As we use document length for element frequency normalization, our model is also compatible with the $LNC1$ constraint in Fang et al. (2004) (see Table 1).

One of the future potential works could be to make our model compatible with the other constraints.

By taking these hypotheses into account, our model could be formulated according to each type of elements. The Relevance Status Value (RSV) between a document d and a query q is:

Words

$$RSV_W(d,q) = \|d \cap q\|_W \times \left(\sum_{w \in q} \frac{N}{N_w} \times \frac{f_{d,w}}{\|d\|_W} \times \|w\| \right) \tag{5}$$

Where
$d = \{w | w \in Index_W(d)\}$ a document

$q = \{w|w \in Index_W\,(q)\}$ a query
$\|d \cap q\|_W = \|\{w|w \in Index_W\,(d) \cap Index_W\,(q)\}\|$ the number of shared words between a document d and a query q
N the number of documents in the corpus
$N_w = \|\{d|w \in Index_W\,(d)\}\|$ the number of documents that contain the word w
$f_{d,w}$ the number of occurances of a word w in a document d
$\|d\|_W$ the number of words in a document d
$\|w\|$ the number of characters in a word w

We added the last component $\|w\|$ to express the importance of an element itself, in isolation from the document that contains it and from the corpus. In other words, is it possible to say that a document d_1 containing an element e_1 should have higher retrieval score than another document d_2 containing e_2, in isolation from all statistical aspects of the model? Well, we believe that the existence of such measure will improve the effectiveness of IR models.

By using words as indexing elements, we tried to approximate this measure by simply supposing that the longer a word is, the more important it is. For ngrams and concepts, we supposed that all concepts/ngrams are equally important and we omitted this component from the model. Finding an effective measure for each type of elements, maybe, is a good direction for the future works of this study.

Ngrams

$$RSV_{NG}\,(d,q) = \|d \cap q\|_{NG} \times \left(\sum_{ng \in q} \frac{N}{N_{ng}} \times \frac{f_{d,ng}}{\|d\|_{NG}}\right) \tag{6}$$

Concepts

$$RSV_{C}\,(d,q) = \|d \cap q\|_{C} \times \left(\sum_{c \in q} \frac{N}{N_{c}} \times \frac{f_{d,c}}{\|d\|_{C}}\right) \tag{7}$$

3.3 The Three Types in one Model (Matching Fusion)

As we said earlier, no single type of indexing elements could cover all aspects of documents and queries. Therefore, merging all types (aspects) in one model could improve the performance of our model Croft (1981) Belkin et al. (1993) Shaw and Fox (1994). One of the merging formulas is:

$$RSV_{all}\,(d,q) = RSV_{NG}\,(d,q) + RSV_{W}\,(d,q) + RSV_{C}\,(d,q) \tag{8}$$

We used the SUM formula for combining the result sets of the three types of elements. This is a type of **late-fusion** because we used each type of elements alone, then we merged the result sets. Conversely to late-fusion, **early-fusion** means combining the three types of elements in one index structure, then applying the model as we have one element type.

4 Model Validation

4.1 Validation Context

The proposed model is validated by applying it to the corpus of ad-hoc image-based retrieval task of the Medical Retrieval track of ImageCLEF2011, and by using the UMLS2009

as an external resource. We use MetaMap Aronson (2006) tool to identify concepts from raw text.

ImageCLEF is a part of CLEF[6] (Cross-Language Evaluation Forum), which is a yearly campaign for evaluation of multilingual information retrieval since 2000. ImageCLEFMed concerns searching medical images depending on heterogeneous and multilingual documents that contain text and images.

ImageCLEF2011[7] Kalpathy-Cramer et al. (2011) contains four main tracks: 1) medical retrieval, 2) photo annotation, 3) plant identification, and 4) Wikipedia retrieval. Medical retrieval track contains three tasks: 1) modality classification, 2) ad-hoc image-based retrieval which is an image retrieval task using textual, image or mixed queries, and 3) case-based retrieval: in this task the documents are journal articles extracted from PubMed[8] and the queries are case descriptions.

The corpus that we used contains: about $230,000$ images with their text caption and title written in English and 30 queries written in three languages: English, French, and German.

UMLS is a multi-source knowledge base in the medical domain. It contains three sources of knowledge:

1. Metathesaurus: is a vocabulary database in the medical domain, extracted from many sources, each source of them is called "Source Vocabularies". The Metathesaurus is organized in Concepts, which represent the common meaning of a set of strings extracted from different source vocabularies.
2. Semantic Network: consists of a set of Semantic Types linked together by two different types of Semantic Relations (hierarchical, non-hierarchical). The purpose of the Semantic Network is to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus.
3. SPECIALIST Lexicon: is a set of general English or biomedical terms and words extracted from different sources.

Moreover, UMLS contains many tools to deal with these different sources (e.g. MetamorphSys, UMLS Knowledge Source Server).

MetaMap is a tool to map text into UMLS concepts. This tool is composed of the following components:

1. Morphology and Syntax: extraction of noun phrases from text using NLP techniques.
2. Variation: construction of different forms (variants) of the noun phrase or part of it.
3. Identification: for each noun phrase variant, it retrieves all concepts that possibly correspond to this variant. The set of concepts that possibly corresponds to the noun phrase, is called "Candidates set".
4. Evaluation: ordering the concepts of candidate set according to an evaluation function (f), which determines: "how much the concept represents the noun phrase?".

6. http://www.clef-campaign.org/
7. http://www.imageclef.org/2011
8. http://www.ncbi.nlm.nih.gov/pubmed/

5. Disambiguation: reduction of the size of the candidates set.

In this study, image captions and titles are used as documents and only the English part of queries is taken into account.

4.2 Text Indexing

We extracted three types of indexing elements:

1. 5gram [9]: before extracting 5grams from documents and queries, we deleted all non-ASCII characters. Then we used five-characters-wide window for extracting 5grams with shifting the window one character each time.
2. Words: before extracting words from documents and queries, we deleted all non-ASCII characters. Then we eliminated the stop words and stem the remaining words using Porter algorithm to get finally the list of words that index documents and queries.
3. Concepts: before mapping the text of documents and queries to concepts, we deleted all non-ASCII characters. Then we mapped the text to UMLS's concepts using MetaMap.

4.3 Model Variants

Actually we experimented seven variants of our model in this study, which are:

$$RSV_{5G*}(d,q) = \left(\sum_{5g \in q} \frac{N}{N_{5g}} \times \frac{f_{d,5g}}{\|d\|_{5G}}\right) \tag{9}$$

$$RSV_{5G}(d,q) = \|d \cap q\|_{5G} \times \left(\sum_{5g \in q} \frac{N}{N_{5g}} \times \frac{f_{d,5g}}{\|d\|_{5G}}\right) \tag{10}$$

$$RSV_{W*}(d,q) = \left(\sum_{w \in q} \frac{N}{N_w} \times \frac{f_{d,w}}{\|d\|_{W}} \times \|w\|\right) \tag{11}$$

$$RSV_{W}(d,q) = \|d \cap q\|_{W} \times \left(\sum_{w \in q} \frac{N}{N_w} \times \frac{f_{d,w}}{\|d\|_{W}} \times \|w\|\right) \tag{12}$$

$$RSV_{C*}(d,q) = \left(\sum_{c \in q} \frac{N}{N_c} \times \frac{f_{d,c}}{\|d\|_{C}}\right) \tag{13}$$

$$RSV_{C}(d,q) = \|d \cap q\|_{C} \times \left(\sum_{c \in q} \frac{N}{N_c} \times \frac{f_{d,c}}{\|d\|_{C}}\right) \tag{14}$$

$$RSV_{SUM}(d,q) = RSV_{5G}(d,q) + RSV_{W}(d,q) + RSV_{C}(d,q) \tag{15}$$

9. 5gram is a ngram consists of five characters. We picked out 5grams because they gave the best results using ImageCLEF2010 comparing to the other ngrams.

4.4 Results and Discussion

The following table (see Table 2) contains the obtained results. The first row (Best) is the result of the first ranked run in the ad-hoc image-based retrieval task in the official campaign [10]. We presented the results using four different metrics: 1) MAP: Mean Average Precision, 2) P@10: Precision after 10 documents retrieved, 3) P@20: Precision after 20 documents retrieved, and 4) #rel_ret: total number of relevant documents retrieved over all queries.

TAB. 2 – *The results of ad-hoc image-based retrieval task*

	MAP	P@10	P@20	# rel_ret	Rank
Best	**0.2172**	**0.3467**	0.3017	1471	1
5G* (Formula 9)	0.1123	0.2033	0.1567	1260	
5G (Formula 10)	0.1473	0.2367	0.2017	1290	
W* (Formula 11)	0.1313	0.1900	0.1967	1421	
W (Formula 12)	0.1963	0.3100	0.2850	1501	
C* (Formula 13)	0.1461	0.2333	0.2133	1456	
C (Formula 14)	0.1664	0.2933	0.2633	1463	
5G+T+C (Formula 15)	0.2008	0.3033	**0.3050**	**1544**	8

The first direct deduction from the results (Table 2) is the importance of $\|d \cap q\|$ component. For ngrams, words, and concepts the precision of the system is improved by using $\|d \cap q\|$ component.

– Promoting documents that share more distinct elements with a query, improves system effectiveness.

The other notable thing in the results is the effectiveness of the data fusion technique, especially, in the number of relevant-retrieved documents. Knowing that the formula (SUM formula), which is used for fusion, was very simple.

– Data fusion is effective in retrieving more relevant documents.

The performance of our model was not bad, even with precision degradation by -7.5%, comparing to the best result. Actually, our model is very simple, and even when we used concepts, we did not exploit any relation. Whereas, the best result Vázquez et al. (2011) was obtained by using a type of query expansion, and exploiting the content and relations of the MeSH [11] ontology.

5 Conclusion

We presented in this paper our approach to index and retrieve documents. We used three types of indexing elements (ngrams, words, concepts) for building a multi-facet document representation, and then we used a simple formula based on three hypotheses (the amount of overlap between a document and a query, the descriptive power of an indexing element, and the discriminative power of an indexing element) for retrieving documents, considering all facets

10. To see all results: http://www.imageclef.org/2011/medical
11. http://www.ncbi.nlm.nih.gov/mesh

(elements' types) of documents. We tried, by using three types of elements then merging them together, to solve the concept-mismatch and external resources incompleteness problems.

We obtained good results. The eighth out of 64 runs in the ad-hoc image-based retrieval task. Knowing that, we used a very simple structure for representing documents and queries and also a very simple ranking formula.

Finally, this study still needs some work. We will compare the performance of our model to the performance of some well-known models e.g. DFR Amati and Van Rijsbergen (2002), BM25 Robertson and Walker (1994), etc.

In addition, we verified the effectiveness of our model by using only one corpus ImageCLEF2011. In order to obtain a more reliable and stable deductions, we should check our model using other corpuses.

Concerning data fusion technique, we tried simple formula (SUM formula). We could try other formulas Shaw and Fox (1994).

In section 3.2, we have introduced the notion of the importance of an element itself. Finding effective and expressive measure, according to each type (ngrams, words, and concepts), is not an easy mission and needs a lot of work.

References

Abdulahhad, K., J.-P. Chevallet, and C. Berrut (2011a). Exploiting and Extending a Semantic Resource for Conceptual Indexing. In *Troisième Atelier Recherche d'Information SEmantique (RISE 2011), Avignon, France*.

Abdulahhad, K., J.-P. Chevallet, and C. Berrut (2011b). Solving concept mismatch through bayesian framework by extending umls meta-thesaurus. In *CORIA 2011*, pp. 311–326.

Amati, G. and C. J. Van Rijsbergen (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst. 20*, 357–389.

Aronson, A. R. (2006). Metamap: Mapping text to the umls metathesaurus.

Aronson, A. R. and T. C. Rindflesch (1997). Query expansion using the umls metathesaurus. *Proceedings of the AMIA Annual Fall Symposium*, 485–489.
français

Baziz, M. (2005). *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

Belkin, N. J., C. Cool, W. B. Croft, and J. P. Callan (1993). The effect of multiple query representations on information retrieval system performance. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, New York, NY, USA, pp. 339–346. ACM.

Bodenreider, O., A. Burgun, G. Botti, M. Fieschi, P. L. Beux, and F. Kohler (1998). Evaluation of the unified medical language system as a medical knowledge source. *Journal of the American Medical Informatics Association 5*(1), 76–87.

Bodenreider, O., A. Burgun, and T. C. Rindflesch (2001). Lexically-suggested hyponymic relations among medical terms and their representation. In *in the UMLS, in Proceedings of TIA 2001*.

Chevallet, J.-P. (2009). endogènes et exogènes pour une indexation conceptuelle intermédia. Mémoire d'Habilitation a Diriger des Recherches.

Chevallet, J.-P., J. H. Lim, and T. H. D. Le (2007). Domain knowledge conceptual inter-media indexing, application to multilingual multimedia medical reports. In *ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisboa, Portugal*.

Crestani, F. (2000). Exploiting the similarity of non-matching terms at retrievaltime. *Inf. Retr. 2*, 27–47.

Croft, W. B. (1981). Incorporating different search models into one document retrieval system. *SIGIR Forum 16*, 40–45.

Croft, W. B. (2000). *Combining Approaches to Information Retrieval*, Volume 7. Springer.

Das-Gupta, P. and J. Katzer (1983). A study of the overlap among document representations. *SIGIR Forum 17*, 106–114.

Deerwester, S. (1988). Improving information retrieval with latent semantic indexing. In C. L. Borgman and E. Y. H. Pai (Eds.), *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, Volume 25, Atlanta, Georgia. American Society for Information Science.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 41*(6), 391–407.

Dozier, C., R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. S. Guo (2007). Fast tagging of medical terms in legal text. In *ICAIL*, pp. 253–260.

Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Systems and Technology (ARIST) 31*, 121–187.

Fang, H., T. Tao, and C. Zhai (2004). A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, New York, NY, USA, pp. 49–56. ACM.

Frakes, W. B. (1992). *Stemming algorithms*, pp. 131–160. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Harter, S. P. (1975a). A probabilistic approach to automatic keyword indexing. part i: On the distribution of specialty words in a technical literature. *Journal of the American Society for Informaiton Science 26*(4), 197–206.

Harter, S. P. (1975b). A probabilistic approach to automatic keyword indexing. part ii: An algorithm for probabilistc indexing. *Journal of the American Society for Informaiton Science 26*(4), 280–289.

Kalpathy-Cramer, J., H. Müller, S. Bedrick, I. Eggel, A. G. S. de Herrera, and T. Tsikrika (2011). The clef 2011 medical image retrieval and classification tasks.

Le, T. H. D. (2009). *Utilisation de ressource externes dans un modèle Bayesien de Recherche d'Information: Application a la recherche d'information médicale multilingue avec UMLS*. Ph. D. thesis, Université Joseph Fourier, Ecole Doctorale MSTII.

Lee, J. H. (1995). Combining multiple evidence from different properties of weighting schemes. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, New York, NY, USA, pp. 180–

188. ACM.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev. 2*, 159–165. Français

Maisonnasse, L. (2008). *Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale.* Ph. D. thesis, Université Joseph-Fourier - Grenoble I.

Radhouani, S. (2008). *Un modèle de recherche d'information orienté précision fondé sur les dimensions de domaine*. Ph. D. thesis, Co-tutelle Université Joseph Fourier Grenoble, Université de Genève (Suisse).

Robertson, S. E. and S. Walker (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, New York, NY, USA, pp. 232–241. Springer-Verlag New York, Inc.

Shaw, J. A. and E. A. Fox (1994). Combination of multiple searches. In *Text REtrieval Conference*.

Vázquez, J. M., M. Crespo, and M. J. M. López (2011). Laberinto at imageclef 2011 medical image retrieval task. In V. Petras, P. Forner, and P. D. Clough (Eds.), *CLEF (Notebook Papers/Labs/Workshop)*.

Woods, W. A. (1997). Conceptual indexing: A better way to organize knowledge. Technical report, Mountain View, CA, USA.

Résumé

Analyse formelle et relationnelle de concepts pour la modélisation et l'interrogation d'une collection documentaire

Nada Mimouni*, Adeline Nazarenko*
Sylvie Salotti*

*LIPN, CNRS(UMR 7030), Université Paris Nord, 99, av. Jean-Baptiste Clément
93430 Villetaneuse
Nada.Mimouni, Adeline.Nazarenko, Sylvie.Salotti@lipn.univ-paris13.fr

Résumé. Une collection documentaire est généralement représentée comme un ensemble de documents mais cette modélisation ne permet pas de rendre compte des relations intertextuelles et du contexte d'interprétation d'un document. Le modèle documentaire classique trouve ses limites dans les domaines spécialisés où les besoins d'accès à l'information correspondent à des usages spécifiques et où les documents sont liés par de nombreux types de relations. Cet article propose un modèle permettant de rendre compte de cette complexité des collections documentaires dans les outils d'accès à l'information. En se basant sur l'analyse formelle et relationnelle de concepts appliquée sur des objets documentaires ce modèle permet de représenter et d'interroger de manière unifiée les descripteurs de contenu des documents et les relations intertextuelles qu'ils entretiennent.

1 Introduction

On représente souvent les collections de documents comme des ensembles de documents mais c'est une vue très simplifiée parce que les documents sont en réalité pris dans un ensemble de relations intertextuelles qui conditionnent leur interprétation : un document très souvent ne s'interprète pas isolément mais en référence à l'ensemble des textes qu'il cite, à partir duquel il est construit et parfois même qui en dérivent.

Si le modèle documentaire classique a fait ses preuves dans la recherche d'information généraliste qui se caractérise par le volume de documents appréhendés, la diversité des requêtes des utilisateurs et la redondance de l'information, il trouve ses limites dans les domaines spécialisés comme la médecine ou les domaines réglementaires où les outils d'accès à l'information trouvent des usages professionnels et critiques. C'est en particulier le cas dans le domaine juridique où les documents sont liés les uns aux autres par des relations d'amendements, de dérivation, de transposition, de complémentation, de jurisprudence, etc. Les outils d'accès à l'information juridique doivent tenir compte de cette complexité du matériau juridique (Bourcier, 2011).

Cet article propose un modèle permettant de représenter et d'interroger de manière unifiée les descripteurs de contenu des documents et les relations intertextuelles qu'ils entretiennent. Il repose sur l'analyse formelle et relationnelle de concepts qui est appliquée sur des objets

documentaires. Cette modélisation permet de faire apparaître deux niveaux de sémantique : un niveau sémantique correspondant à la prise en compte des liens entre les documents et un niveau sémantique qui résulte de l'annotation de nos documents par des descripteurs de contenu.

Après une revue de l'état de l'art dans la section 2, nous présentons la manière dont nous proposons de modéliser les collections documentaires sur un exemple détaillé et nous montrons dans la section 4 l'intérêt de cette modélisation pour la recherche d'information.

2 État de l'art

Le modèle classique de la Recherche d'Information (RI) représente les documents comme des sacs de mots auxquels sont assignés des poids mesurant leur importance dans le texte (poids binaire, fréquence, etc.). La recherche est ensuite faite sur cet ensemble de mots pondérés. La RI sémantique, décrite dans (Baeza Yates et R., 1999; Pejtersen, 1998), enrichit la RI classique en généralisant ou en spécifiant la requête à l'aide de ressources sémantiques (thesaurus, ontologies). Mais la RI sémantique, comme la RI classique, retourne comme résultat une liste de documents indépendants sans tenir compte du graphe de documents auquel ils appartiennent (graphe des liens entre les documents).

Quand le graphe de documents est pris en compte, c'est pour améliorer le classement des documents retournés comme dans le cas des algorithmes PageRank (Brin et Page, 1998; Page et al., 1999) et HITS (Kleinberg, 1999).

Un autre ensemble de travaux met l'accent sur l'analyse des graphes de citations (Newman, 2004; Ding, 2011) dans le but d'étudier le comportement d'une communauté en interaction (catégorisation) mais non pas dans une perspective de RI comme nous proposons de le faire dans ce travail en considérant que la prise en compte des liens enrichit sémantiquement l'interprétation et la RI dans une collection de documents.

La méthode que nous proposons repose sur l'Analyse Formelle de Concepts (AFC) et l'Analyse Relationnelle de Concepts (ARC). L'AFC est une méthode de classification conceptuelle qui, à partir d'un jeu de données représenté sous la forme d'un tableau binaire *(objets x attributs)*, construit une hiérarchie de concepts où chaque concept représente un ensemble maximal d'objets $(extension)$ ayant en commun un ensemble maximal d'attributs $(intension)$. Dans cette hiérarchie, appelée treillis de Galois ou treillis de concepts, les concepts sont (partiellement) ordonnés selon l'inclusion ensembliste entre leurs intensions et de façon duale l'inclusion inverse entre leurs extensions. La recherche d'information a été explicitement mentionnée dans (Godin et al., 1995) comme étant l'une des applications possibles des treillis de concepts. La relation de subsumption, qui est une relation d'ordre partiel entre les concepts, permet le passage d'un concept, correspondant à une requête, à un autre plus général ou plus spécifique (Godin et al., 1995).

L'utilisation de l'AFC dans la RI a fait l'objet de plusieurs travaux. Dans (Messai et al., 2006) et (Comparot et al., 2010), les auteurs proposent des techniques de raffinement et d'expansion de requête en s'appuyant sur des ontologies de domaine, ce qui permet d'améliorer le rappel par généralisation ou par spécialisation en se basant sur la structure du treillis de Galois. Sur des données textuelles, (Carpineto et Romano, 2005) propose une méthode de recherche d'information par treillis de concepts. Dans (Messai et al., 2005), les auteurs ont utilisé les treillis de concepts pour la découverte et l'interrogation de ressources génomiques sur le web.

D'autres travaux ont mis l'accent sur la classification et la structuration des résultats fournis par les algorithmes de RI ce qui influe sur les interfaces de navigation (Nauer et Toussaint, 2008; Poshyvanyk et Marcus, 2007; Carpineto et al., 2006; Koester, 2006). L'idée principale est de créer un contexte formel à partir des résultats fournis par les moteurs de recherche sur le web, de construire le treillis correspondant à ce contexte, puis de proposer à l'utilisateur un classement des résultats tel que construit par ce treillis. Ce type d'approche est implémenté dans plusieurs systèmes opérationnels tels que CREDINO (Carpineto et al., 2006), FooCA (Koester, 2006) ou CRECHAINDO (Nauer et Toussaint, 2008). Dans son travail, E. Nauer propose de classifier les résultats de recherche sur le web pour permettre à l'utilisateur de juger la pertinence des résultats qui lui sont fournis. D. Poshyvanyk *et al.* utilisent l'AFC pour classifier les résultats de la RI suite à une requête pour localiser des concepts dans un code source. Ces travaux construisent une classification conceptuelle des documents retournés mais ne tiennent pas compte des liens entre ces documents. De plus, l'exploitation de ces documents est faite *a posteriori* sur la base des résultats qui sont calculés indépendamment.

L'approche que nous présentons ici permet d'intégrer l'intertextualité *a priori* dans le modèle documentaire grâce à l'apport de l'extension relationnelle de l'AFC (ARC). Ce modèle documentaire pourra être exploité par des outils de recherche et de navigation ce qui permettra, entre autres, de répondre à des requêtes qui portent sur les relations entre documents en tant que telles.

L'ARC, proposée par (Rouane et al., 2007), est une extension relationnelle de l'AFC permettant de prendre en compte des relations entre les objets d'un même contexte, les relations entre attributs ou éventuellement la combinaison des deux. L'ARC a été utilisée dans plusieurs domaines d'application comme la classification de services web (Azmeh et al., 2011), l'extraction de patterns d'ontologie (Rouane et al., 2010), la restructuration et la construction d'ontologie, le "refactoring" de diagrammes de cas d'usage UML (Dao et al., 2004) mais jamais, à notre connaissance, pour la RI. Nous nous plaçons dans le cadre de RI dans une collection documentaire où les documents sont inter-reliés.

Nous montrons ici comment l'AFC et l'ARC permettent de représenter une collection documentaire et les perspectives d'interrogation que cela ouvre. Nous utilisons ces techniques pour formaliser un processus de RI qui exploite à la fois le contenu sémantique des documents et leurs relations intertextuelles.

Cette approche, qui n'est pas adaptée à la RI généraliste sur le web, prend tout son sens dans le cadre d'une RI spécialisée portant sur un domaine particulier. Nous nous intéressons ici au domaine juridique où les documents sont fortement liés les uns aux autres et où ces liens jouent un rôle important dans l'interprétation que l'utilisateur fait des documents retournés par un moteur de recherche.

Même dans un domaine restreint, la complexité du calcul peut être rédhibitoire. Même si, dans les applications réelles la complexité théorique maximale n'est pas atteinte (Carpineto et Romano, 2000), la complexité du treillis, mesurée en nombre de concepts et liée à la taille des contextes formels, limite l'utilisation des treillis de concepts pour la recherche d'information. L'ARC est cependant présentée ici comme modèle de représentation de la collection documentaire, sans préjuger du modèle de calcul réel à utiliser sur un corpus de grande taille où les calculs fins pourraient n'être faits que localement.

3 Modélisation d'une collection documentaire

Une collection documentaire est un ensemble de documents d'un domaine spécifique (biologique, scientifique, médical, juridique, etc.) avec des liens entre ces documents. Nous proposons ici un modèle unifié permettant de représenter à la fois le contenu de ces documents et les liens qui existent entre eux.

3.1 Exemple de collection documentaire

Notre étude se place dans le cadre du projet Legilocal [1] dont l'objectif est de faciliter l'accès des citoyens aux documents juridiques des collectivités locales. Le corpus que nous étudions est un ensemble de documents juridiques traitant du bruit. Ces documents sont de plusieurs types : arrêtés municipaux et préfectoraux, décrets, lois, codes et ordonnances. Pour simplifier la présentation du modèle, nous ne distinguons pas ici les différents types de liens entre documents (amendements, dérivation, etc.), mais cette information peut être représentée dans le modèle. Nous illustrons la modélisation sur un ensemble de quelques arrêtés dans lesquels figurent des références à des décrets et des lois. Nous montrons comment l'AFC permet de construire un premier treillis modélisant le contenu des arrêtés, qui peut être ensuite enrichi par la prise en compte, avec l'ARC, d'informations concernant les références aux décrets.

3.2 AFC pour la modélisation du contenu textuel

Dans cette section nous montrons comment l'approche AFC est appliquée pour la formalisation du contenu de notre collection documentaire. Une définition plus détaillée de l'analyse formelle de concepts est donnée dans (Ganter et Wille, 1999).

Le contenu des documents est d'abord modélisé sous la forme d'un contexte formel qui décrit une relation binaire entre un ensemble d'objets et un ensemble d'attributs *(objet x attributs)*. Les objets correspondent aux documents. Les attributs sont des descripteurs sémantiques caractérisant le contenu de ces documents. Ces descripteurs peuvent être des simples mots-clés, mais il peut s'agir de noms de concepts issus d'une ressource sémantique suite à un processus d'annotation sémantique de documents (Kiryakov et al., 2004). Cette phase d'annotation sémantique apparaît dans certains travaux qui se basent sur l'AFC pour faire la RI dans lesquels les objets sont décrits par des attributs qui font référence à des concepts dans des structures sémantiques (e.g. l'annuaire biologique BioRegistery décrit par des concepts de MeSH et NCBI (Messai et al., 2006)). Dans ce travail, nous proposons une description sémantique de nos documents en leurs associant des descripteurs sémantiques du domaine extraits du thesaurus juridique EuroVoc [2].

La formalisation du contenu des documents est donnée par le contexte formel $\mathcal{K}_{arr} = (A, S, I)$, où A est un ensemble de documents (Arrêté préfectoral Paris, Arrêté municipal Strasbourg,...), S est un ensemble de descripteurs sémantiques du domaine (ex. nuisance sonore, bruit) et I une relation binaire entre A et S appelée incidence de $\mathcal{K}_{arr}$ et vérifiant les propriétés : $I \subseteq A \times S$ et $(a, s) \in I$ ou (aIS) où a, s sont tels que $a \in A$ et $s \in S$ signifie que

1. Legilocal est un projet FUI 2010-12. Voir http ://www.mondeca.com/fr/R-D/Projets/LegiLocal-Projet-FUI-9-Cap-digital-2010-2013.
2. http ://eurovoc.europa.eu/

le document a est caractérisé sémantiquement par le descripteur s. Des exemples de contextes formels sont donnés dans la table 1 (arrêtés) et la table 2 (décrets).

	Bruit Anormalement Gênant (bag)	Nuisance Sonore (ns)	Pollution Acoustique (pa)	Sonorisation (son)	Niveau Sonore (nvs)
Arrêté Paris (AP)	x		x		
Arrêté Boulogne Billancourt (AB)	x	x		x	
Arrêté Yvelines (AY)	x	x			x
Arrêté Strasbourg (AS)			x		x

TAB. 1 – *Le contexte formel des arrêtés $\mathcal{K}_{arr}$.*

Un concept formel dans la formalisation des documents de notre collection $\mathcal{K}_{arr}$ est un ensemble de documents partageant un ensemble de descripteurs sémantiques. Un concept formel est défini comme suit.

Definition 1 (Concept formel)**.** *Soit $\mathcal{K}_{arr} = (A, S, I)$ un contexte formel. Un **concept formel** est un couple (X, Y) tel que $X \subseteq A$, $Y \subseteq S$. X et Y sont respectivement appelées $extension$ et $intension$ du concept formel (X, Y).*

Dans la théorie des treillis des relations d'ordre partiel inverse sont définies entre les extensions d'une part, et les intensions d'autre part. On parle de relations de subsumption. Notons par $\mathcal{C}$ l'ensemble des concepts formels de $\mathcal{K}_{arr}$. Soient $C_1 = (X_1, Y_1)$ et $C_2 = (X_2, Y_2)$ dans $\mathcal{C}$. C_1 est subsumé par C_2 si $X1 \subseteq X2$ où de façon duale $Y2 \subseteq Y1$ (noté par $C_1 \sqsubseteq C_2$). $(\mathcal{C}, \sqsubseteq)$ est un treillis complet appelé treillis de concepts correspondant au contexte formel $\mathcal{K}_{arr}$. On notera dans la suite $(\mathcal{C}, \sqsubseteq)$ par $\mathcal{L}(\mathcal{C})$.

La figure1 montre le treillis de concepts $\mathcal{L}(\mathcal{C})$ correspondant au contexte formel des arrêtés $\mathcal{K}_{arr}$ donné par la table 1. De la même façon, à partir du contexte formel des décrets et des lois $\mathcal{K}_{dec}$, on construit le treillis de concepts correspondant $\mathcal{L}(\mathcal{C}')$ (figure 2). Dans ces treillis, nos documents sont structurés sous forme de concepts. Un concept représente une classe de documents (l'extension) caractérisée ou décrite par un ensemble de descripteurs (l'intension).

Pour plus de clarté nous notons dans la suite ai les concepts du treillis des arrêtés et dj les concepts du treillis des décrets. Par exemple, le concept $a4$ dans le treillis des arrêtés (table1) représente l'ensemble des documents qui partagent les descripteurs $b - a - g$ (bruit anormalement gênant) et $n - s$ (nuisance sonore). Cela correspond dans notre exemple aux documents AB (arrêté de Boulogne) et AY (arrêté des Yvelines). Le lien entre les concepts $a3$ et $a4$ peut

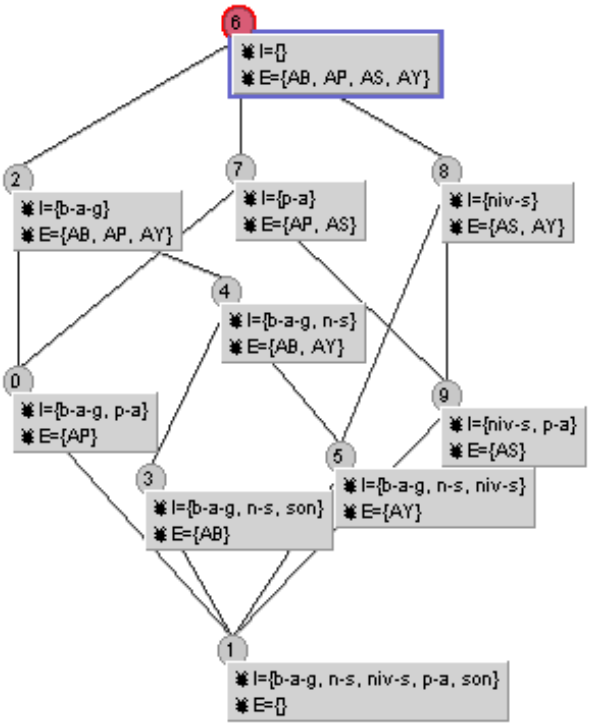

FIG. 1 – *Le treillis de concepts $\mathcal{L}(\mathcal{C})$ correspondant au contexte formel des arrêtés $\mathcal{K}_{arr}$.*

être interprété comme un lien de généralisation/spécialisation entres les classes représentées par ces concepts.

Dans une perspective de RI, le treillis construit par l'AFC regroupe toutes les combinaisons possibles des attributs des documents. Ces combinaisons sont représentées par les intensions des concepts ayant comme extension tous les documents partageant ces propriétés. Pour satisfaire les critères d'une requête en terme de pertinence, la recherche consiste à identifier la classe de documents qui partage le plus d'attributs avec la requête.

3.3 ARC pour la modélisation des liens intertextuels

L'ARC, extension relationnelle de l'AFC, permet de modéliser deux types de relations : relations entre objets et relations entre attributs (propriétés). Nous nous contentons ici de faire l'étude du premier type de relation, qui exprime les relations qui existent entre nos documents. Le deuxième type de relations sera étudié dans des travaux ultérieurs.

L'approche construit à partir de contextes binaires *(objets x attributs)* et d'une relation représentée séparément dans un deuxième contexte, une Famille de Contextes Relationnels. Cette famille constitue le point de départ du processus de formation des structures conceptuelles correspondantes appelées familles de treillis relationnels. Dans notre cas, les références

	Lutte Contre le Bruit (lcb)	Tranquilité du Voisinage (tv)	Activité Bruyante (ab)	Isolation Phonique (ip)
Décret 95 (D95)	x		x	
Code Pénal (CPen)		x		x
Ordonnance 1945 (O45)		x	x	
Loi 1992 (L92)	x			x

TAB. 2 – *Le contexte formel des décrets* $\mathcal{K}_{dec}$.

entre documents sont décrites par un contexte relationnel qui définit les relations entre les objets *(objet x objet)* [3]. Dans notre exemple, nous considérons des références que les arrêtés font aux décrets et autres textes de lois. Un exemple de ces relations de référence est représenté sur la table 3.

L'approche ARC construit un unique treillis unifiant les informations provenant des contextes formels initiaux *(objets x attributs)* et du/des contexte(s) relationnel(s) *(objet x objet)*. Sur notre exemple, le treillis final résultant après enrichissement relationnel est donnée par la figure 3.

	D95	CPen	O45	L92
AP	×			
AB				×
AY		×		
AS			×	

TAB. 3 – *Relation : fait_référence*

4 Résultats et interprétation

4.1 Modèle d'une collection documentaire

Modélisée à l'aide de l'analyse formelle et relationnelle de concepts, la collection documentaire est représentée par un ensemble de classes de documents qui sont caractérisées à la fois par des descripteurs de contenus et par les relations que les documents entretiennent les uns avec les autres. Formellement, la collection documentaire est représentée par un treillis

3. Pour modéliser différents types de liens, il faut créer différents contextes relationnels.

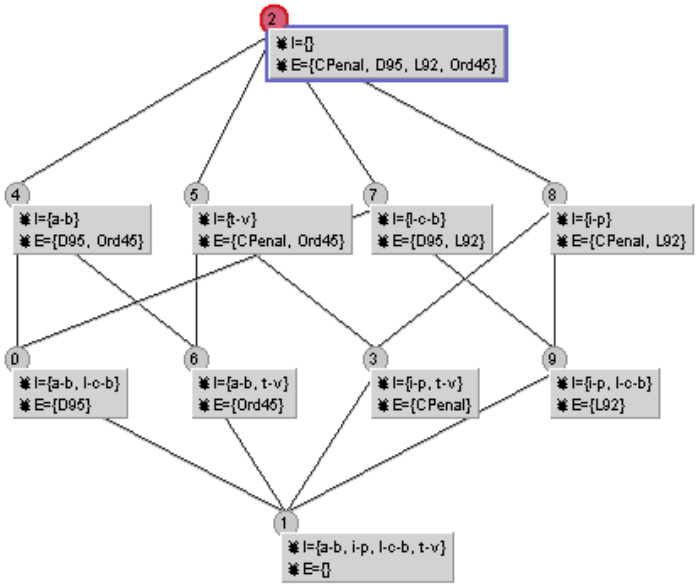

FIG. 2 – *Le treillis de concepts $\mathcal{L}(\mathcal{C}')$ correspondant au contexte formel des décrets $\mathcal{K}_{dec}$*

de concepts formels, dont les extensions sont des classes de documents et les intentions une conjonction d'attributs qui sont des descripteurs de contenu et/ou des relations vers d'autres classes de documents.

La figure 3 montre le treillis obtenu en intégrant au treillis initial de la figure 1 l'information sur les relations que les arrêtés entretiennent avec les décrets. Par souci de lisibilité du résultat et pour faciliter l'interprétation de l'exemple, nous n'avons pris en compte que des relations entre arrêtés et décrets, ce qui correspond à la structure réelle du corpus des documents juridiques que nous traitons dans le cadre de ce travail. Il faut souligner, cependant, que cette structure n'est que partielle. Il faudrait prendre en compte d'autres types de relations, entre arrêtés ou entre décrets, par exemple [4].

Si on compare le treillis de la figure 3 avec celui de la figure 1, la plupart des concepts ont une extension inchangée mais leur intension est enrichie d'attributs relationnels. C'est le cas par exemple du concept $a4$ qui a la même extension $E = \{ABoulogne, AYvelines\}$ dans les deux treillis mais dont l'intension finale combine les descripteurs de contenu de départ ($\{b-a-g, n-s\}$) avec deux descripteurs relationnels ($\{references : c2, reference : c8\}$) qui indiquent que le nouveau concept 4 est lié à deux autres concepts formels, $d2$ et $d8$ [5].

L'introduction des relations fait aussi apparaître de nouveaux concepts. Sur l'exemple jouet présenté, c'est le cas du seul concept $n^o\,10$ qui apparaît dans le treillis de la figure 3 mais qui n'était pas dans le treillis initial. Dans ce cas, l'information relationnelle a conduit

4. Nous considérons que les relations intertextuelles ne sont pas réflexives (un document n'est pas lié à lui-même).
5. Ces deux concepts correspondent à des classes de décrets dans le treillis $\mathcal{L}(\mathcal{C}')$.

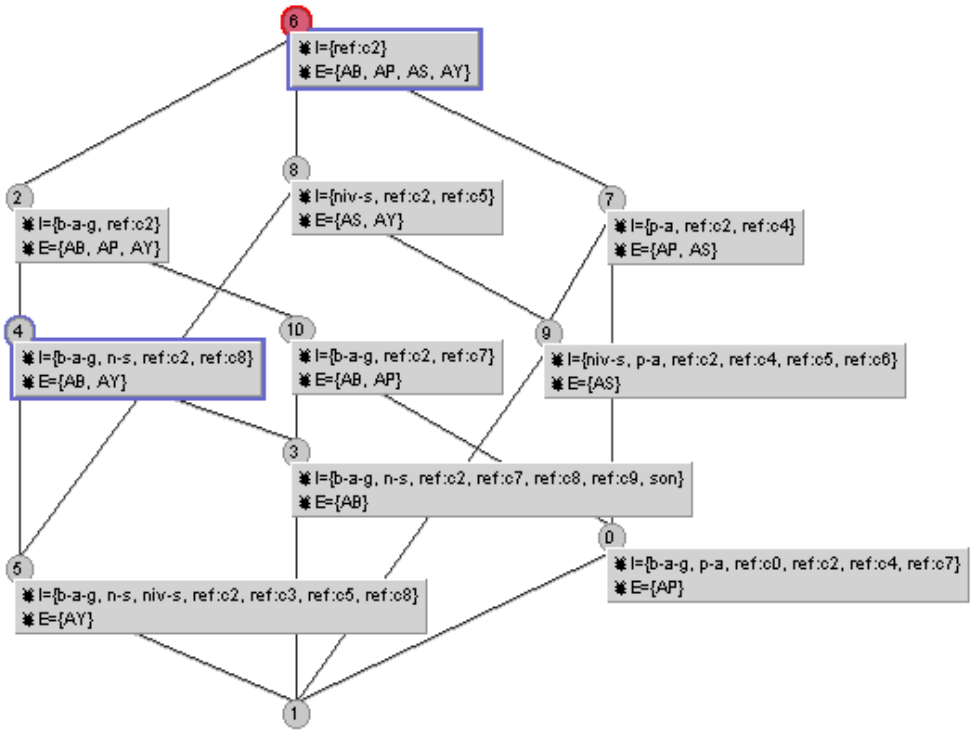

FIG. 3 – *Treillis résultant après enrichissement relationnel entre objets.*

à créer un regroupement intermédiaire ($\{ABoulogne, AParis\}$) entre ceux des concepts $a2$ ($\{ABoulogne, AParis, AYvelines\}$) et $a0$ ($\{AParis\}$) du treillis initial.

L'ajout des attributs relationnels s'interprète comme l'introduction de relations entre différentes classes de documents. Dans notre exemple, la classe $a4$ des arrêtés est ainsi reliée aux classes de décrets $d2$ et $d4$. A noter qu'il y a un processus inductif à ce stade puisque la classe $\{ABoulogne, AYvelines\}$ est reliée à la classe de décrets $\{CPenal, L92\}$ alors que les seules relations intertextuelles explicites au départ étaient entre $AYvelines$ et $CPen$ et entre $ABoulogne$ et $L92$.

4.2 Interrogation

On peut considérer que le treillis initial des arrêtés représente l'ensemble des requêtes (ou combinaisons de descripteurs) qui peuvent être faites sur la collection documentaire des arrêtés et qui sont satisfiables, c'est-à-dire qui permettent de retourner des arrêtés (toutes les combinaisons de descripteurs associées à une extension non nulle). Si la requête correspond à l'intension d'un concept qui a une extension, ce sont les documents de cette extension qui sont retournés en réponse à la requête ; si la requête correspond à une intension sans extension propre, on peut proposer des spécialisations ou au contraire généraliser la requête.

Dans cette perspective de recherche d'information, on peut apprécier l'apport de l'information relationnelle et de la modélisation que nous proposons.

Il faut d'abord souligner que tous les concepts formels initiaux étant conservés dans le treillis final, toutes les requêtes satisfiables sur le premier treillis le restent sur le treillis final.

On peut répondre à davantage de requêtes puisqu'il y a plus de concepts avec une extension propre dans le treillis : l'information relationnelle affine la catégorisation de l'ensemble des documents.

Notre modélisation permet surtout de répondre à de nouvelles formes de requêtes, les requêtes relationnelles :

- On peut retrouver un ensemble de documents associés à un autre ensemble de documents, les premiers constituant en quelque sorte le contexte d'interprétation des seconds :
  Quelles sont les classes de documents qu'un auteur donné cite ou par lequel il est cité ?
  En référence à quels documents un texte doit-il être interprété ?
  Cette forme de requête s'apparente à une requête traditionnelle couplée avec une stratégie de navigation de proche en proche à partir des liens des premiers documents retournés mais s'y ajoute ici un processus inductif qui généralise les liens entre documents individuels à des classes de documents.
- On peut interroger de manière plus globale sur la catégorie de documents qui sont associés à certains textes :
  Étant donné un ensemble d'arrêtés, à quel type de décrets font-ils référence,
  au-delà de liens explicites de référence entre arrêtés et décrets ?
  Sur quoi portent les amendements apportés à un décret particulier
  ou à un ensemble de décrets ?
- On peut finalement faire porter la requête sur les catégories sémantiques ainsi mises en relation pour découvrir par exemple que les décrets portant sur l'isolation phonique (caractérisés par le descripteur $i - p$) ont donné lieu à des arrêtés sur les nuisances sonores ($n - s$, classe $a2$ dans le treillis des arrêtés, figure 1) ou que les directives sur la pollution acoustique sont transposées dans des décrets parlant de bruit.

5 Conclusion

Nous avons présenté une modélisation qui donne une représentation unifiée des descripteurs de contenus et des relations intertextuelles qui caractérisent une collection documentaire. Nous défendons en effet l'idée que la recherche d'information spécialisée doit tenir compte des relations entre documents. C'est notamment critique pour l'accès à l'information juridique. Le modèle que nous décrivons permet de tenir compte de deux types d'informations sémantiques : les descripteurs sémantiques de contenu et les relations intertextuelles. Il permet d'obtenir des réponses plus riches à des requêtes classiques portant sur le contenu des documents, en rendant compte aussi du contexte documentaire dans lequel les documents retournés doivent être interprétés. Il permet également d'exprimer des requêtes plus riches, qui portent directement sur la structure intertextuelle de la collection documentaire.

Références

Azmeh, Z., M. Driss, F. Hamoui, M. Huchard, N. Moha, et C. Tibermacine (2011). Selection of composable web services driven by user requirements. *the Application and Experience Track of ICWS 2011*.

Baeza Yates, R. A. et N. B. R. (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman.

Bourcier, D. (2011). Sciences juridiques et complexité. un nouveau modèle d'analyse. *Droit et Cultures 61*(1), 37–53.

Brin, S. et L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst. 30*, 107–117.

Carpineto, C., A. D. Pietra, S. Mizzaro, et G. Romano (2006). Mobile clustering engine. In *ECIR*, pp. 155–166.

Carpineto, C. et G. Romano (2000). Order-theoretical ranking. *Journal of the American Society for Information Science 51*, 587–601.

Carpineto, C. et G. Romano (2005). Using concept lattices for text retrieval and mining. In *Formal Concept Analysis*, pp. 161–179.

Comparot, C., O. Haemmerlé, et N. Hernandez (2010). Expression de requêtes en graphes conceptuels à partir de mots-clés et de patrons. In *Journées Francophones d'Ingénierie des Connaissances (IC), Nîmes, 08/06/2010-11/06/2010*, http://www.cepadues.com/, pp. 81–92. Cépaduès Editions.

Dao, M., M. Huchard, M. R. Hacene, C. Roume, et P. Valtchev (2004). Improving generalization level in uml models iterative cross generalization in practice. In *ICCS'04: International Conference on Computational Science*, pp. 346–360.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics 5*, 187–203.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis* (Mathematical Foundations ed.). Springer.

Godin, R., W. Mineau, et R. Missaoui (1995). Méthodes de classification conceptuelle basées sur les treillis de galois et applications. *Revue d'intelligence artificielle 9*, 105–137.

Kiryakov, A., B. Popov, D. Ognyanoff, D. Manov, et K. M. Goranov (2004). Semantic annotation, indexing, and retrieval. *Journal of Web Semantics 2*, 49–79.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM 46*, 604–632.

Koester, B. (2006). Conceptual knowledge retrieval with fooca: Improving web search engine results with contexts and concept hierarchies. In *Industrial Conference on Data Mining*, pp. 176–190.

Messai, N., M.-D. Devignes, A. Napoli, et M. Smaïl-Tabbone (2005). Querying a bioinformatic data sources registry with concept lattices. In *ICCS*, pp. 323–336.

Messai, N., M.-D. Devignes, A. Napoli, et M. Smaïl-Tabbone (2006). Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques bioregistry. *Ingénierie des Systèmes d'Information (ISI) 11*(1), 39–60.

Nauer, E. et Y. Toussaint (2008). Classification dynamique par treillis de concepts pour la recherche d'information sur le web. In *CORIA'08: Conférence en Recherche d'Information et Applications*, pp. 71–86.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America 101*, 5200–5205.

Page, L., S. Brin, R. Motwani, et T. Winograd (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Pejtersen, A. M. (1998). Semantic information retrieval. *Commun. ACM 41*, 90–92.

Poshyvanyk, D. et A. Marcus (2007). Combining formal concept analysis with information retrieval for concept location in source code. In *ICPC*, pp. 37–48.

Rouane, M. H., M. Huchard, A. Napoli, et P. Valtchev (2007). A proposal for combining formal concept analysis and description logics for mining relational data. In *Proceedings of the 5th international conference on Formal concept analysis, ICFCA 2007*, LNAI, pp. 51–65. Springer-Verlag.

Rouane, M. H., M. Huchard, A. Napoli, et P. Valtchev (2010). Using formal concept analysis for discovering knowledge patterns. In *CLA'10: 7th International Conference on Concept Lattices and Their Applications*, CEUR, pp. 223–234. University of Sevilla.

Summary

A collection of documents is generally represented as a set of documents but this simple representation does not take into account cross references between documents, which often defines their context of interpretation. This standard document model is less adapted for specific professional uses in specialized domains in which documents are related by many various references and the access tools need to consider this complexity. We propose a unified model based on formal and relational concept analysis applied on documentary objects that represents and queries in a unified way documents content descriptors and documents relations.