Actes de l'atelier recherche d'information sémantique RISE 2010

J.P. Chevallet, Catherine Roussey

Second Atelier Recherche d'Information SEmantique RISE, Marseille 25 mai 2010

Associé au 28ème Congrès INFORSID 2010.

ACTES DE L'ATELIER RECHERCHE D'INFORMATION SEMANTIQUE

Edité par

Catherine ROUSSEY, LIRIS, Lyon, CEMAGREF, Clermont Ferrand (France)

Jean-Pierre CHEVALLET, LIG, Grenoble (France)

Seconde édition de l'atelier Recherche d'Information SEmantique 2010

Atelier Recherche d'Information SEmantique RISE
Associé à la conférence INFORSID 2010

1 Introduction

Les documents produits actuellement sont essentiellement numériques. Une frénésie de numérisation est en passe de rendre accessible les ouvrages les plus anciens. La communication est également massivement numérique et voit l'émergence de nouvelles pratiques (blogs, SMS, réseaux sociaux), en plus des média textuels numériques bien implantés (email). Cette tendance s'intensifie avec la nomadisation de l'accès à l'information (téléphone portable, ultra-portables, iPad). Les objectifs à court terme sont alors une connexion ubiquitaire pour tous au réseau internet. Toutefois, même si cette masse d'informations est disponible, la difficulté majeure réside dans l'accès à de l'information ciblée, c'est à dire réellement en adéquation avec un besoin personnel et ponctuel. Cet accès se fait par filtrage, sélection, navigation ou interrogation.

Les systèmes de Recherche d'Information (RI) ont proposé une première réponse à ce problème d'accès à l'information pertinente. Les modèles développés en RI sont maintenant largement utilisés, par exemple dans les moteurs de recherche du Web. Les technologies actuelles sont basées sur des modèles statiques qui manipulent des informations de bas niveau. Par exemple, la plupart des moteurs de recherche sont basés sur le comptage des mots ou des liens sur les pages. Les dernières avancées de la recherche en RI ont concerné essentiellement l'amélioration des modèles statistiques d'appariement de documents, comme les modèles de langue statistiques. De nouvelles pistes de recherche consistent à ajouter de la sémantique pour obtenir des modèles statistiques intelligents. La sémantique permet d'améliorer la précision des résultats d'un système de RI en évitant les problèmes liés à l'ambiguïté ou au manque d'expressivité des mots simples. Même s'il ne semble pas nécessaire qu'un système de RI "comprenne" le document qu'il indexe, traiter le besoin de l'utilisateur au niveau sémantique permet plus de précision dans les réponses.

Nous pensons donc que l'avenir des systèmes de Recherche d'Information passe par la prise en compte de la sémantique du contenu des documents, permettant à un utilisateur de mieux maîtriser le flux d'information pour cibler l'information dont il a réellement besoin. Une façon d'atteindre cet objectif est de coder explicitement des connaissances associées aux termes, par exemple dans des ontologies. Le but de cet atelier est de discuter de ce nouveau terrain de recherche: les systèmes de "concierge d'information" où le flux d'information est enrichi par une interprétation de son contenu. Nous appellerons ce nouveau paradigme: Recherche d'Information Sémantique. Cet atelier est dédié à tous les types de Recherche d'Information sans

contrainte sur le mode de stockage de cette information. Par exemple la Recherche d'Information peut s'appliquer sur des documents textuels, des images, des vidéos, des flux XML etc...

2 Objectifs

Les travaux sur les ontologies ou les ressources sémantiques existent et sont actifs dans les différentes communautés informatique comme : le Web, la bio-informatique ou les systèmes d'information géographiques. Ainsi, les ressources sémantiques comme les ontologies, les bases de données lexicales, les thésaurii, se développent et sont maintenant disponibles. Cet atelier est spécialement dédié à l'usage des ressources sémantiques dans les systèmes de Recherche d'Information Multimedia et/ou Multilingue.

Des systèmes de Recherche d'Information Multilingue cherchent à retrouver des documents qui correspondent à un thème indépendamment de leur langue d'écriture. Dans le cas de documents non textuels (Multimédia), des données textuelles peuvent être extraites de leur contenu, apparaître dans le voisinage du document ou être issues d'annotations manuelles. Malheureusement, la nature peu structurée et le volume important d'information rendent difficilement accessible l'information pertinente aux utilisateurs. Pour résoudre ce problème, les travaux en Recherche d'Information (RI) se sont orientés vers les technologies issues du Web Sémantique et plus précisément sur l'usage des ressources sémantiques comme les ontologies, les thésaurii ou les bases de données lexicales.

3 Themes

L'atelier RISE a pour but de proposer un lieu de rencontre entre des chercheurs issus de différentes communautés comme la Recherche d'Information, le Web Sémantique, le TALN, le Multimedia, l'Ingéniérie des Connaissances.

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

- Indexation Conceptuelle et Indexation Sémantique,
- Recherche d'Information Multimedia et Multilingue
- Extraction d'Information Multilingue et Multimedia
- Annotation Sémantique
- Web Sémantique
- Ontologies Multilingues et Multimedia,
- Alignement d'Ontologie et Correspondance pour la Recherche d'Information,
- Graphes Conceptuels, Logiques de Description, Langages de Représentation des connaissances pour la Recherche d'Information.

- Utilisation des Distances Sémantiques pour la Recherche d'Information Alignement d'Ontologie et Correspondance pour la Recherche d'Information

4 Comité de Programme

- AUSSENAC-GILLES Nathalie, IRIT Toulouse (France)
- CHEVALLET Jean-Pierre, LIG, Grenoble (France)
- DAMAS Luc, LISTIC, Annecy (France)
- GRAU Brigitte, ENSIIE (France)
- HERNANDEZ Nathalie , IRIT, Toulouse (France)
- MAISONNASSE Loïc, TecKnowMetrix, Lyon (France)
- METAIS Elisabeth, CNAM Paris (France)
- ROCHE Christophe, LISTIC, Annecy (France)
- ROUSSEY Catherine, LIRIS, Lyon (France)
- RUMPLER Béatrice, LIRIS, Lyon (France)
- SCHWAB Didier ,LIG-GETALP, Grenoble (France)
- SERASSET Gilles, LIG, Grenoble (France)
- SIMONET Michel, TIM-C, Grenoble (France)
- ZARGAYOUNA Haïfa , LIPN, Paris (France)
- ZWEIGENBAUM Pierre, LIMSI (France)

La journée de l'atelier a été découpée en deux sessions :

- la première session est dédiée aux mesures de similarité qui peuvent être utilisées soit pour l'alignement d'ontologies soit pour l'appariement entre les représentations des documents et des requêtes.
- la seconde session est centrée sur la classification utilisant des ressources sémantiques.

L'atelier RISE 2010 a réuni une vingtaine de chercheurs francophones venant de différents laboratoires : IRIT, LIG, LIPN, LIRIS, LSIS, MODEME, ONP. Nous remercions tous les membres du comité de programme pour la qualité de leur travail ainsi que les auteurs des articles.

Les organisateurs de l'atelier: Catherine Roussey et Jean Pierre Chevallet

6 Programme

Session : Mesure de similarité et usages

Session : ressources sémantiques et classification

Architecture d'un système d'aide à l'alignement d'ontologies

Mina Ziani1, Danielle Boulanger1, Guilaine Talens1

1 Université Jean Moulin
Equipe MODEME
6 Cours Albert Thomas BP 8242
69355 Lyon
{mina.ziani;db;talens}@univ-lyon3.fr

Abstract. Semantic Interoperability based on ontologies is an important challenge. In Cooperation of shared knowledge, alignment of ontologies requests to find correspondences between semantically related entities of different ontologies. It is a difficult task based on the definition of 'similarity measures'. We classify these methods and introduce some existing tools combining several types of methods to build 'mappings'. Unlike these tools, we propose a computer-aided system to choice similarity measures. We try to define a realistic proposal, applicable to the domain of geotechnics which covers several sub-domains.

Keywords: Ontology, Ontology Alignment, Similarity measure, Semantic Interoperability, Knowledge management.

1 Introduction

La géotechnique est la science qui se préoccupe des interactions entre un sol et un construit [5]. C'est un domaine complexe faisant intervenir plusieurs métiers : physiciens, chimistes… De ce fait, la multiplicité des référentiels géotechniques ainsi que l'hétérogénéité des connaissances pose des problèmes pour le partage et la recherche d'informations.

Dans le cadre d'une convention de recherche (Convention CETU n° 2005_4.69011) avec le centre d'Etude des Tunnels (CETu), nous avons été amenés à re-construire une ontologie destinée à des géotechniciens. En effet, l'ontologie initiale contenant toutes les spécificités des différents métiers devenait de plus en plus difficile à gérer (trop volumineuse). Nous avons relevé les deux problèmes suivants : la masse de documents dans le domaine de la géotechnique est très importante et très redondante. Il faudrait pouvoir représenter toute la connaissance nécessaire, sans aucune redondance, ni aucune ambiguïté (des définitions différentes pour un même terme). De plus, au fur et à mesure que l'ontologie devient importante, la recherche dans un contexte donné (une branche de la géotechnique) est de moins en moins pertinente. Le résultat est une ontologie hybride avec une partie « consensuelle » regroupant les concepts partagés par les experts géotechniciens et un ensemble d'ontologies locales représentant chacune les concepts et instances spécifiques à un métier. Ces ontologies locales sont amenées à coopérer. Pour permettre cette coopération, nous proposons de construire un système d'aide pour l'alignement d'ontologies métier. Son principe repose sur le calcul de mesures de similarité et la création de liens entre les entités appartenant à diverses ontologies. Plusieurs outils et Framework ont été proposés pour permettre ce processus. A partir d'une étude comparative de ces derniers, nous avons proposé un Framework d'aide au choix de mesures de similarité à la demande et permettant la combinaison de plusieurs mesures de similarité. Ce travail est applicable au domaine de la géotechnique.

Le papier est structuré de la façon suivante : la section 2 rappelle quelques types de mesures de similarité, des outils et Frameworks qui permettent une combinaison de méthodes pour construire des mappings et propose une comparaison des différents systèmes décrits. Dans la section 3, nous nous focaliserons sur l'architecture et présenterons quelques caractéristiques du système d'aide. Nous conclurons en faisant le point sur notre travail et mentionnerons quelques perspectives.

2 Alignement : Etat de l'art

L'alignement d'ontologies est le processus de mise en correspondance sémantique des entités qui les composent [4]. Le processus est exécuté selon une stratégie ou une combinaison de techniques de calcul de mesures de similarité et

utilise un ensemble de paramètres (ex : paramètres de pondération, seuils …) et un ensemble de ressources externes (ex : thésaurus, lexique…). Au final, nous obtenons un ensemble de liens sémantiques reliant les entités qui composent les ontologies. Ces derniers comprennent des relations d'équivalence, de généralisation/spécialisation, de chevauchement ou encore d'incompatibilité. De nombreux travaux ont été développés dans le domaine de l'alignement d'ontologies et portent sur les techniques de recherche de similarité et sur les outils ou sur les Framework qui les intègrent.

2.1 Les techniques de mesures de similarité

On retrouve plusieurs méthodes de calcul de la similarité entre les entités de plusieurs ontologies. Des classifications de celles-ci sont données dans [7, 14]. Nous retenons :

- Les méthodes terminologiques [9, 12] : elles sont employées pour calculer la valeur de similitude des entités textuelles, telles que des noms, des méta-données sur les noms, des étiquettes, des commentaires,…
- Les méthodes linguistiques utilisant des ressources externes (dictionnaires, taxonomies,…) : la similarité entre deux entités représentées par des termes est calculée à partir des liens sémantiques déjà existants dans les ressources externes [16].
- Les méthodes structurelles internes [10] : elles calculent la similarité entre deux concepts en exploitant les informations relatives à leur structure interne (restrictions et cardinalités sur les attributs, valeurs des instances,…).
- Les méthodes structurelles externes ou conceptuelles : elles se servent de la structure hiérarchique de l'ontologie et se basent sur des techniques de comptage d'arcs pour déterminer la similarité sémantique entre deux entités [15,19].
- Les méthodes extensionnelles : elles déduisent la similarité entre deux entités qui sont notamment des concepts ou des classes en analysant leurs extensions (leurs ensembles d'instances). Chaque instance peut être représentée par un vecteur de noms et/ou de valeurs. Des calculs de similarités entre vecteurs permettent de comparer les instances [17],
- Les méthodes hybrides : elles combinent plusieurs mesures lorsqu'une seule est insuffisante [8].

Ces méthodes sont intégrées dans des outils permettant la mise en correspondance d'ontologies. Il existe des outils qui combinent plusieurs méthodes de similarité et des Frameworks implémentant plusieurs mesures et permettant ainsi de suggérer à l'expert plusieurs mappings.

2.2 Les outils

Différents outils ont été développés dans le but d'aligner plusieurs ontologies.

PROMPT est un système interactif constituant une aide pour la comparaison, l'alignement, la fusion et l'évolution d'ontologies [13]. Son module d'alignement appelé Anchor-Prompt permet de rapprocher des ontologies de la façon suivante : d'abord, des 'matchers' terminologiques permettent de déterminer un ensemble initial de concepts similaires. A partir de cette liste, un algorithme analyse les chemins dans les sous-graphes délimités par ces concepts et détermine quelles classes apparaissent fréquemment dans les mêmes positions sur des chemins similaires. Cette analyse permet de guider l'utilisateur pour choisir les meilleurs mappings.

OLA (OWL Lite Alignment) est un système implémentant un algorithme d'alignement des ontologies décrites en OWL. OLA mesure la similarité entre deux entités à partir des calculs de similarité entre leurs caractéristiques (leurs types : classe, relation ou instance, leurs liens avec d'autres entités : sous-classes, domaine, …). La valeur de similarité finale est la somme pondérée des valeurs de similarité de chaque caractéristique [4].

AROMA (Association Rule Ontology Matching Approach) est une approche d'alignement pour des ontologies représentées en OWL. Elle permet de découvrir des liens sémantiques de type « subsomption » ou « équivalence » entre deux entités (classes ou propriétés). Le processus d'alignement se déroule en trois étapes : la première procède à l'acquisition des termes contenus dans les descriptions et instances des entités à partir d'outils de Traitement Automatique du Langage (TAL). Ensuite, pour chaque entité, ainsi qu'à ses ancêtres, est associé un ensemble de termes dits représentatifs. La deuxième étape permet de créer des relations de subsomption entre les entités à partir de règles d'association[1]. Des 'matchers' terminologiques sont utilisés pour comparer les différentes descriptions. Enfin, la dernière étape vise à analyser les règles d'associations trouvées afin de : (1) déduire des relations d'équivalence, (2) éliminer les incohérences (cycles) (3), supprimer les relations redondantes, (4) sélectionner le meilleur alignement pour chaque entité [1].

ASMOV (Automated Semantic Mapping of Ontologies with Validation) est un système d'alignement d'ontologies conçu pour faire coopérer des ontologies issues de sources de données hétérogènes. ASMOV permet de produire des mappings entre des concepts et/ou des propriétés et/ou des instances de deux ontologies [6]. L'algorithme implémenté est automatique, il calcule de façon itérative, la similarité entre deux entités appartenant à deux ontologies différentes suivant quatre caractéristiques (les éléments lexicaux, les relations structurelles, la structure interne, les instances de classes et valeurs des propriétés). La similarité finale est

[1] Les règles d'associations sont construites sur le principe qu'une entité X est plus spécifique que ou équivalente à une entité Y si le vocabulaire utilisé dans les descriptions et les instances de X a tendance à être inclus dans celui de Y

calculée à partir de la somme pondérée des quatre mesures et permet d'obtenir un alignement. Le système vérifie ensuite cet alignement afin de s'assurer qu'il ne contient pas d'incohérence sémantique.

2.3 Les Frameworks

Plus récemment, les Frameworks sont apparus dans les systèmes d'alignements d'ontologies. Leur avantage est qu'ils permettent de multiples combinaisons de stratégies de calcul de la similarité. Par exemple :

COMA++ (COmbining MAtching) est un système générique de mise en correspondance de schémas (XML, Schémas relationnels) [2]. L'outil permet l'importation, le stockage, l'édition de schémas ainsi que leurs alignements et ce, afin de les transformer ou de les fusionner. Il fournit une bibliothèque extensible d'algorithmes de mappings, un module pour combiner les résultats obtenus et une plateforme pour l'évaluation des différentes mesures. L'utilisateur peut interagir dans le processus de mise en correspondance en sélectionnant le mode de combinaison de matchers. COMA++ est une évolution de COMA qui améliore ses algorithmes et son interface graphique.

MAFRA (Mapping Framework for distributed ontologies) est un Framework interactif, dynamique et progressif pour l'alignement d'ontologies distribuées dans le cadre du Web sémantique [11]. L'approche de MAFRA se déroule suivant deux dimensions : la dimension horizontale définit les étapes du processus d'alignement : importation et normalisation des ontologies à aligner, calcul des similarités entre les éléments des différentes ontologies à partir d'une combinaison de plusieurs mesures de similarités, formalisation des mappings en établissant des « ponts sémantiques » entre les entités des différentes ontologies, exécution des mappings pour transformer les instances d'une ontologie source vers les instances d'une ontologie cible à partir des ponts sémantiques, créés précédemment, et enfin vérification des résultats obtenus. Quant à la dimension verticale, elle est constituée de quatre modules additionnels (gestion de l'évolution, gestion des ponts sémantiques, construction d'un consensus coopératif et Interface graphique).

FOAM (Framework for Ontology Alignment and Mapping) est un Framework utilisé dans plusieurs systèmes comme QOM[2] , NOM[3] , APFEL[4] , à des fins d'intégration de données, de fusion d'ontologies, d'évolution d'ontologies… L'outil implémente plusieurs mesures et stratégies existantes de recherche de similarités et permet de faire des mappings entre des ontologies décrites en OWL. Le processus général d'alignement est le suivant : on sélectionne les paires d'entités à comparer ainsi que les caractéristiques sur lesquels se fera la comparaison. Le système calcule, pour chaque paire et pour chaque caractéristique, une similarité. Ces résultats sont

2 Quick Ontology Mapping
3 Naïve Ontology Mapping
4 Alignment Process Feature Estimation and Learning

combinés pour obtenir la similarité finale entre chaque paire d'entités. A partir de ces résultats, FOAM permet de proposer à l'utilisateur un ensemble de suggestions d'alignement, qu'il peut accepter ou rejeter [3].

RiMOM (Risk Minimisation based Ontology Mapping) est un Framework interactif implémentant diverses stratégies pour l'alignement d'ontologies [18] et suivant plusieurs étapes : la première consiste à sélectionner les matchers à utiliser selon la similarité supposée entre les ontologies (terminologique ou structurelle). Dans la deuxième phase, plusieurs mesures sont appliquées, indépendamment les unes des autres. Ensuite, les résultats sont agrégés en utilisant une fonction d'interpolation linéaire. La troisième étape consiste à propager les similarités (de concept à concept, de propriété à propriété et de concept à propriété). Enfin, la dernière étape génère des mappings à partir des résultats obtenus précédemment. Le processus est itératif, avec une validation des résultats à chaque itération.

2.4 Comparaison des différents outils et Frameworks

Les outils et Frameworks que nous avons cités précédemment sont, pour la plupart, considérés par l'OAEI[5] parmi les meilleurs systèmes d'alignement. La plupart d'entre eux utilisent des mesures de similarité terminologiques et/ou structurelles et/ou extensionnelles et proposent une stratégie de combinaison pour trouver la similarité finale. Celle-ci représente en général une équivalence ou un lien de subsomption entre deux entités appartenant à deux ontologies différentes.

L'utilisation de plusieurs mesures de similarité donne souvent de meilleurs résultats. Par contre, ces outils ne précisent pas toujours quels matchers ont été utilisés ni comment les similarités ont été agrégées. Par ailleurs, il est à noter que les Frameworks sont plus adaptés pour la réutilisation ainsi que pour la combinaison de mesures de similarité existantes. Ces systèmes diffèrent également au niveau de leur fonctionnement et de l'interaction qu'ils offrent à leurs utilisateurs. L'intervention d'un expert de domaine dans le processus d'alignement d'ontologies s'avère souvent essentielle pour éviter des incohérences. De plus, des outils interactifs tels que PROMPT ou FOAM, qui suggèrent des résultats de mappings à l'utilisateur, donnent souvent de meilleurs résultats. Par contre, ils ne permettent pas de réutiliser des résultats de mappings pour déduire d'autres relations de correspondances.

La comparaison des différents outils et Frameworks est reprise dans le tableau 1.

[5] Ontology Alignment Evaluation Initiative

Outil	Techniques utilisées	Combinaison	Algorithme	Mappings
ASMOV	mesures terminologiques, structurelles internes, conceptuelles et extensionnelles	correspond à la somme pondérée des 4 mesures	automatique	équivalence
AROMA	outils de TAL pour l'extraction de termes, règles d'associations et mesures terminologiques	utilise une fonction d'interpolation linéaire	automatique	subsomption et équivalence
Anchor Prompt	mesures terminologiques et structurelles	pas de combinaison de similarité	semi-automatique	équivalence
COMA	mesures terminologiques et structurelles	est fonction des choix de mesures	semi-automatique	équivalence
MAFRA	mesures terminologiques, structurelles et extensionnelles	pas de combinaison de similarité	semi-automatique	équivalence (ponts sémantiques)
FOAM	mesures terminologiques et structurelles	est fonction de caractéristiques retenues par l'utilisateur	semi-automatique	équivalence
RiMOM	mesures terminologiques, structurelles et extensionnelles	dépend de la similarité supposée entre les ontologies (structurelle ou terminologique)	automatique	équivalence

Tableau. 1. *Comparaison des différents outils et Frameworks.*

Le système d'alignement que nous proposons de construire est un Framework et permet d'interagir avec ses utilisateurs pour trouver des mappings. Le processus d'alignement est basé sur le calcul de mesures de similarité ainsi que sur la réutilisation de liens de synonymie existants (créés durant les opérations d'alignements). Contrairement à la plupart des systèmes qui proposent uniquement des résultats de mappings, notre Framework propose également de guider l'expert pour choisir les méthodes de mesures de similarité les plus appropriées à sa requête, dans le but de lui offrir une meilleure interaction et des suggestions de mappings pertinents.

3 Proposition d'une approche d'alignement

3.1 Contexte

Dans le cadre de la re-conception de l'ontologie initiale construite au CETu, nous avons construit une ontologie hybride constituée d'un ensemble d'ontologies locales, décrivant chacune un métier du domaine de la géotechnique et d'une ontologie globale consensuelle [20]. Dans un premier temps, le système sélectionne l'ontologie « cible ». L'ontologie contenant le plus de concepts communs avec les autres est sélectionnée car elle a le plus de liens avec l'ensemble des ontologies locales. Ensuite, les autres sont intégrées une à une à cette dernière. Le résultat est une ontologie globale consensuelle, contenant tous les concepts communs entre les différentes ontologies ainsi que les relations non conflictuelles qui les lient. Ce travail préliminaire a pour but de permettre aux experts géotechniciens, d'une part de gérer le contenu de la base de connaissances, en manipulant des ontologies métiers représentées sous forme d'arbres lisibles, simples à mettre en œuvre et faciles à appréhender, et d'autre part de concilier les différentes ontologies métiers.

La figure 1 représente une partie de l'ontologie hybride. Pour ne pas alourdir l'exemple, seule une partie des ontologies est représentée. Les ontologies locales représentées sont la maîtrise d'ouvrage (MOA) et la maîtrise d'œuvre (MOE). La première décrit les concepts et relations utilisées par les maîtres d'ouvrage, dont le rôle est l'analyse des besoins et l'expression de la solution fonctionnelle dans les projets géotechniques ; la deuxième décrit ceux utilisés par les maîtres d'œuvres qui sont chargés de réaliser les projets. Seuls, quelques attributs significatifs sont inscrits dans la figure 1 ; en réalité chaque concept en contient beaucoup plus.

Une ontologie globale et cohérente regroupe tous les concepts partagés par les différents métiers. Suivant l'exemple, nous en avons représentés trois : le concept « tunnel » qui est le plus générique dans la hiérarchie, ce qui n'est pas très surprenant étant donné que l'ontologie a été construite par des géotechniciens qui étudient des tunnels. Les deux autres concepts partagés sont les concepts « chantier » et « ouvrage » qui sont parmi les plus courants dans notre domaine, mais qui peuvent être vus différemment selon le métier. En effet, comme le montre l'exemple, les propriétés de ces concepts diffèrent d'un métier à l'autre. L'ensemble de ces ontologies définit une ontologie hybride avec une partie consensuelle partagée par tous, et les concepts et instances spécifiques à la maîtrise d'œuvre et/ou la maîtrise d'ouvrage. Pour faire coopérer les ontologies métiers entre elles, nous proposons un système d'aide pour guider l'expert dans le processus de création de nouveaux liens sémantiques entre les ontologies locales. Dans ce qui suit, nous allons proposer une architecture d'aide à l'alignement, dans le but de permettre la coopération des ontologies métiers.

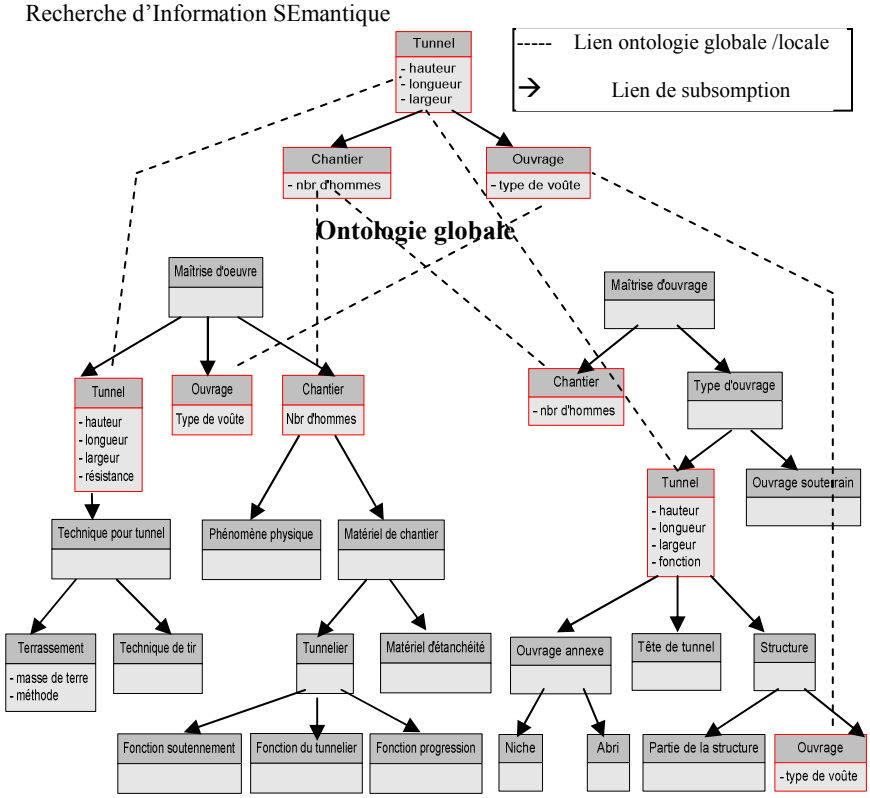

Figure 1. *Représentation d'une partie de l'ontologie hybride.*

3.2 Architecture de notre système

La plupart des systèmes d'alignements exploitent différentes mesures de similarité pour déduire la similarité entre deux entités. La difficulté est de choisir la bonne mesure ou combinaison de mesures pour retrouver des similarités. Le système que nous proposons (cf. figure 3) doit permettre d'aider les experts géotechniciens dans le processus de recherche de similarités entre concepts puis de permettre de générer des mappings. Les ontologies métier (ex : l'ontologie MOA) construites par les experts de domaine ainsi que l'ontologie globale sont représentées en OWL. C'est un langage formel permettant de décrire les ontologies.

Lorsque l'expert d'un domaine souhaite coopérer avec un autre expert, il envoie au système une requête contenant le concept à apparier (1). Cette recherche s'effectue sur les ontologies (2) que l'on souhaite faire correspondre : l'ontologie de départ (celle du métier de l'expert) et une ontologie de recherche (celle avec laquelle il souhaite coopérer). Le but est de découvrir les éléments de l'ontologie de

recherche que l'on peut mettre en correspondance avec l'élément de l'ontologie de départ. Ensuite, le système vérifie l'existence de synonymes (3). Un panel de méthodes de calcul de mesures de similarité est implémenté dans un Framework et peut être proposé à l'expert de domaine (4). L'intérêt du Framework est de (i) permettre de réutiliser toutes les mesures implémentées (terminologiques, conceptuelles,…) sur toutes les ontologies, (ii) permettre la combinaison de plusieurs mesures de similarités pour définir les règles de correspondances entre deux entités. Le système a pour objectif de mettre à la disposition du géotechnicien plusieurs mesures et de l'aider à choisir celle(s) qu'il faut utiliser. Le résultat est un ensemble de similarités entre les concepts initiaux et les concepts trouvés dans l'ontologie de recherche. Celles-ci peuvent être de différents types (équivalence, synonymie,…) et sont stockées dans une base de données des similarités (5), puis proposées à l'expert du domaine (6). Il reste à sa charge le choix de valider ou non ces similarités (7) afin de permettre au système de générer les mises en correspondances et de les stocker dans une base de mappings (8). Lorsque l'expert connait les deux concepts à apparier, il a la possibilité de créer directement un mapping.

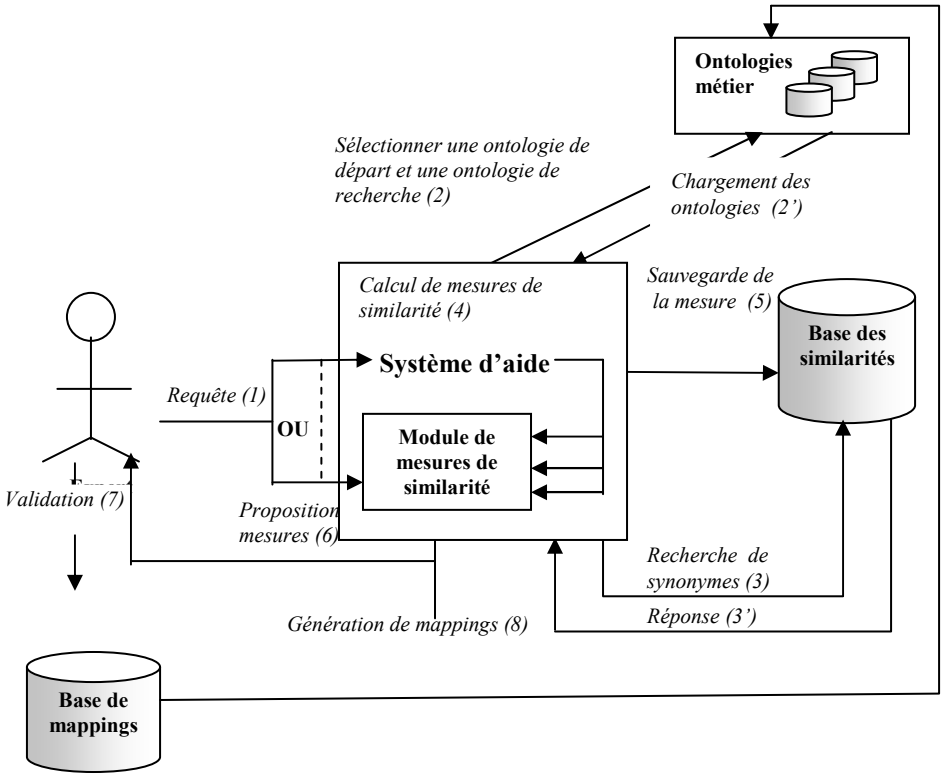

Figure 2. *Architecture du système d'alignements.*

3.3 Quelques caractéristiques du système d'aide

Le système d'aide a pour but de guider l'expert dans son choix d'utilisation d'une méthode.

Les mesures de similarités terminologiques entre concepts sont calculées en priorité, car elles sont les plus simples à gérer. Elles sont calculées entre le nom du concept (et/ou son synonyme) dans l'ontologie de départ et les concepts appartenant à l'ontologie de recherche. Lorsque les mesures de similarités terminologiques donnent un résultat significatif (supérieur à un seuil paramétrable), cela signifie que les concepts comparés sont similaires du point de vue lexical. Cette similarité a un sens s'il existe un lien entre leurs plus petits ancêtres appartenant à l'ontologie globale.

Lorsque les concepts ne sont pas lexicalement similaires ou que les mesures terminologiques sur les concepts ne suffisent pas à déduire la similarité, le système va proposer des méthodes sémantiques pour la déduction des synonymes. Deux méthodes peuvent être utilisées :

Dans le cas, où le nombre d'attributs du concept de départ est significatif (supérieur ou égal à 2, valeur également paramétrable), on préconise de calculer les distances terminologiques entre les attributs du concept de départ (et/ou leurs synonymes) et ceux des concepts de l'ontologie de recherche. Le système donnera en sortie la liste de toutes les similarités trouvées (terminologique entre concepts et terminologique entre attributs).

S'il ne trouve aucune similarité ou que le nombre d'attributs du concept de départ n'est pas significatif (inférieur à 2) c'est la méthode de comptage d'arcs qui est utilisée. Elle consiste à compter le nombre d'arcs (N) entre le concept de départ et le plus petit concept subsumant entre les deux ontologies locales. Ensuite, l'ontologie de recherche est parcourue à partir de l'arc N-2 du concept commun jusqu'à l'arc N+2. Si le nombre de concepts parcourus n'est pas très important (inférieur ou égal à 10), on les propose à l'expert comme étant des concepts probablement similaires au concept de départ.

Enfin, si toutes ces méthodes ne suffisent pas à trouver toutes les similarités existantes, on peut aller plus loin, en comparant les instances de concepts. On dira que deux classes sont équivalentes si elles partagent un sous-ensemble d'instances pour des attributs choisis par l'expert et qui seront représentés par des vecteurs.

3.4 Implémentation

Les ontologies locales sont construites par consensus entre plusieurs experts d'un sous domaine. Quant à l'ontologie globale, elle est construite de façon automatique par une approche d'intégration [20]. Pour manipuler ces ontologies, il existe plusieurs outils permettant l'édition d'ontologies. Nous utilisons Protégé pour les raisons suivantes : d'abord, parce qu'il permet de décrire les ontologies en OWL, ensuite parce qu'il est basé sur un mécanisme de plugins, donc extensible, enfin parce que son interface graphique est très intuitive.

Le développement de notre application est fait en Java car ce langage permet l'utilisation de plusieurs API permettant de manipuler les ontologies : Jena[6] , Sesame[7] , Corese[8] … Les deux premières permettent de manipuler des ontologies décrites en OWL. De plus, il existe également plusieurs API de calcul de mesures de similarité réalisées en Java : Simmetrics [9], SecondString[10] ,… Nous les réutilisons dans notre Framework, en plus d'autres mesures que nous avons développées.

La réalisation du système d'aide avec une interface utilisateur interactive est en cours. Le choix d'une ou de plusieurs mesures de similarité est paramétrable par l'utilisateur. Cependant, il est souhaitable qu'il utilise les mesures qui lui sont suggérées par le système d'aide. L'ensemble des similarités calculées est stocké dans une base de données MySQL. Les mappings générés à partir de ces mesures après leur validation par l'expert du domaine sont aussi conservées.

4 Conclusion

Dans ce papier, nous avons présenté l'architecture d'un système d'aide à l'alignement d'ontologies métiers dans le domaine de la géotechnique. Un expert recherche un des concepts contenus dans son ontologie dans une autre ontologie locale. Afin de réaliser une coopération entre ontologies locales, des mappings horizontaux sont établis.

Nous proposons un système d'aide qui permet de sélectionner des méthodes de similarité tenant compte des caractéristiques de l'ontologie de départ et de l'ontologie de recherche. L'expert est donc guidé dans la sélection des méthodes. Avant de lancer la recherche, la base de similarités est consultée afin de connaître les synonymes potentiels du terme précédemment rencontré dans cette ontologie ou dans une autre. Le prototype est en cours de réalisation. Une perspective de ce travail serait d'étudier l'impact de l'évolution des ontologies locales sur les mappings générés ainsi que sur l'ontologie globale. En effet, l'évolution des

[6] http://jena.sourceforge.net/
[7] http://www.openrdf.org/
8 http://www-sop.inria.fr/acacia/soft/corese/
[9] http://www.dcs.shef.ac.uk/~sam/simmetrics.html
[10] http://secondstring.sourceforge.net/

ontologies peut entraîner la création de mappings verticaux entre une ontologie locale et l'ontologie globale.

5 Bibliographie

1. David J. AROMA : une méthode pour la découverte d'alignements orientés entre ontologies à partir de règles d'associations, thèse de doctorat en informatique, université de Nantes, (2007).
2. Do H., Rahm E. COMA – a system for flexible combination of schema matching approaches, in 28th International Conference on Very Large Data Bases, Hong Kong, (2002) 610–621.
3. Ehring M. Ontology Alignment: Bridging the Semantic Gap : Semantic Web and Beyond, New York, Springer, (2007).
4. Euzenat, J., Valtchev, P. Similarity-based ontology alignment in OWL-lite. In Proceedings 15th European Conference On Artificial Intelligence, Valencia (2004).
5. Faure N., Un système d'aide à la modélisation des connaissances en géotechnique, thèse de doctorat en informatique, université Jean-Moulin, Lyon (2007).
6. Jean-Mary Y., Kabuka, M. ASMOV: Ontology Alignment with Semantic Validation. In SWDB-ODBIS Workshop, Vienna, Austria, (2007) 15-20.
7. Kalfoglou Y., Schorlemmer M. Ontology mapping: the state of the art. The Knowledge Engineering Review Journal, Vol. 18, n°1, (2003) 1–31.
8. Leacock C., Chodorow M. Combining Local Context and WordNet Similarity for Word Sense Identification. In WordNet : An Electronic Lexical Database, MIT Press, (1998).
9. Levenshtein V. Binary codes capable of correcting deletion, insertions, and reversals, In Soviet Physics Doklady, Vol. 10 , n° 8, (1966) 707-710.
10. Madhavan J., Bernstein P., Rahm, E. Generic schema matching with cupid, In Proceedings of the 27th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc, (2001) 49–58.
11. Mädche A, Motik B., Silva N. and Volz R. MAFRA - a MApping FRAmework for distributed ontologies. In Proceedings of the International Conference on Knowledge Engineering and Knowledge Management, Siguenza, Spain, Lecture Notes in Computer Science, Springer, (2002) 235–250.
12. Monge A., Elkan C., The field-matching problem : algorithm and applications, In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, (1996) 267-270.
13. Noy N. et Musen M. Anchor. PROMPT : Using non-local context for semantic matching, In Proceedings workshop on ontology and information sharing (IJCAI), Seattle, (2001) 63–70.
14. Rahm E., Bernstein P. A survey of approaches to automatic schema matching. The International Journal on Very Large Data Bases, Vol. 10, n°4, (2001) 334–350.

15. Resnik P. Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, n° 11, (1999) 95-130.

16. Safar S., Reynaud C., Calvier F. Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire, 1ères Journées Francophones sur les Ontologies, 2007, Sousse, Tunisie, (2007) 21-35.

17. Stumme G., Maedche A. 2001. FCA-MERGE: Bottom-Up Merging of Ontologies, IJCAI'01 Workshop on Ontologies and Information Sharing, Seattle, USA.

18. Tang J., Li J., Liang B., Huang X., Li Y., Wang K. Using Bayesian decision for ontology mapping, Journal of Web Semantics, Vol. 4, n°1, (2006) 243-262.

19. Wu Z., Palmer M. Verb semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, (1994) 133-138.

20. Ziani M., Boulanger D., Talens G. Designing an Hybrid Ontology From Domain Ontologies, 2èmes Journées Francophones sur les Ontologies, Lyon, (2008) 41-47.2. Présentation du texte courant

Une approche de recherche sémantique dans les documents semi-structurés

Rami Harrathi, Sylvie Calabretto

LIRIS UMR 5205 - INSA de Lyon,
7 avenue Jean Capelle,
69621 Villeurbanne cedex, France,
{rami.harrathi, sylvie.calabretto}@insa-lyon.fr

Résumé. Dans cet article, nous proposons une approche de recherche d'information sémantique des documents semi-structurés. L'idée centrale de notre travail est que l'utilisation de ressources sémantiques externes telles que les thesaurus et les ontologies peut améliorer l'efficacité du processus de recherche. Ainsi, nous proposons d'utiliser le modèle vectoriel sémantique où une partie d'un document ainsi que la requête sont représentées par deux vecteurs de concepts. Nous proposons également d'utiliser les mesures de similarité sémantique pour l'évaluation d'une mesure de pertinence. La mesure proposée se base sur la comparaison entre des graphes sémantiques.

Mots clés: recherche d'information, documents (semi-)structurés, XML, ressource sémantique, ontologie.

1 Introduction

La Recherche d'Information (RI) dans les documents semi-structurés (documents XML) consiste à identifier les éléments XML (les nœuds de l'arbre XML) les plus pertinents par rapport à une requête donnée. La majorité des approches proposées dans la littérature sont des adaptations des modèles traditionnels (vectoriel, probabiliste, de langue, etc.).

Ces adaptations visent à tenir compte de la structure et à attribuer des scores de pertinence aux nœuds des documents XML en tenant compte de certaines spécificités des documents XML. Ainsi, les différents types de modèles (vectoriel, probabiliste, de langue, etc.) sont étendus de diverses façons pour tenir compte de la structure. Ainsi, on ajoute des paramètres supplémentaires pour ajuster les formules classiques comme : le nombre d'enfants d'un élément [1], son type, la fréquence de ce type d'élément dans la collection [2, 3], l'importance d'un terme dans les autres éléments du même type [2].

La majorité des approches proposées dans la recherche des documents semi-structurés (documents XML) reposent sur des systèmes d'indexation à base de mots clés ou encore sur les termes. Les seules informations utilisées concernant ces termes sont leurs fréquences d'apparition dans les documents, ou dans les éléments du document (en fonction du niveau de granularité). Ainsi, ces approches ne prennent pas en considération le sens du mot (terme).

L'indexation par des mots clés est généralement imprécise [4]. Cette imprécision est due au fait que les termes d'indexation présentent une forte ambigüité. En effet, le sens d'un mot clé peut varier selon le contexte dans lequel il apparait (phénomène de polysémie). Aussi, ces approches ne prennent pas en compte la synonymie. Par conséquent, dans ces systèmes, il est impossible de trouver des parties des documents représentées par un mot M_1 synonyme d'un mot⟦ M⟧_2, où M_2 représente une requête. Par conséquent, il se peut qu'un système de RI basé sur les mots ne renvoie pas un élément pertinent, c'est-à-dire un élément qui satisfait la requête.

Un moyen pour améliorer les performances des systèmes de RI sur les documents semi-structurés [5] est la prise en compte de la sémantique des termes d'indexation. Ce type d'indexation passe du niveau des mots au niveau des concepts (les sens des mots) pour mieux décrire le contenu du document et de la requête. Ces approches utilisent des ressources sémantiques (thésaurus, ontologies, etc.) dans les phases d'indexation et de recherche.

Dans cet article nous proposons une approche de recherche d'information sémantique des documents semi-structurés. Nous présentons cet article de la manière suivante : dans la section 2 nous décrivons les travaux connexes ; dans la section 3, nous présentons notre approche pour la recherche d'information sémantique des documents semi-structurés. Dans la section 4, nous présentons un

plan d'expérimentation que nous comptons mettre en pratique prochainement, et nous concluons dans la section 5.

2 Travaux connexes

La recherche d'information sémantique dans les documents semi-structurés s'intéresse principalement à la représentation des documents et des requêtes par des taxonomies de concepts. Les systèmes d'indexation et de recherche par les concepts proposés dans la littérature nécessitent de disposer de ressources sémantique afin d'extraire des concepts à partir des textes, et un modèle de mesure de similarité entre concepts [6].

2.1 Recherche d'information sémantique dans les documents semi-structurés

Les documents semi-structurés sont caractérisés par la présence d'une structure organisant leurs contenus textuels. Ainsi, les systèmes d'indexation et de recherche de documents semi-structurés par les concepts se divisent en trois approches correspondant à trois manières de tenir compte de la structure et du contenu textuel lors de l'indexation.

2.1.1 Approches orientées structure

Les approches orientées structure proposent d'indexer uniquement la structure. Le processus d'indexation consiste à représenter les noms des éléments (les noms des balises) par des concepts en utilisant une ontologie de noms. Par exemple les balises comme "university" et "school" ou "car" et "automobile" sont indexées par le même concept. L'intérêt de cette approche est de supporter les requêtes vagues, par exemple si on veut chercher un élément dont le nom est "university", le système de recherche peut retourner les éléments "university" et "school".

Parmi ces approches, on peut citer le système CXLEngine [7]. Ce système utilise une ontologie de noms (ontology label) pour faire correspondre les noms des balises aux concepts dans la hiérarchie de l'ontologie.

Dans [8], les auteurs proposent d'utiliser une ontologie pour gérer des documents de structures hétérogènes. Dans cette approche, la grammaire DTD (Document Type Definition) associée à un document XML est indexée par une DTD de concept (Ontology DTD).

2.1.2 Approches orientées structure et contenu

Les approches orientées structure et contenu consistent à indexer la structure et le contenu textuel par des concepts en utilisant une ontologie. Dans [9], un document XML est considéré comme un ensemble de paires (concept élément, valeur) où "valeur" désigne l'index du contenu textuel qui est représenté par un ensemble de concepts pondérés. Le score de pertinence attribué à un élément est

calculé par une fonction de similarité entre l'ensemble de concepts de la requête et la valeur de l'élément.

Le système de recherche XXL [10], permet l'interrogation de documents XML. Le moteur de recherche XXL présente une architecture s'appuyant sur 3 structures d'index [10] :

– Index des noms des éléments et des attributs : les noms sont indexés par des concepts. Cet index permet l'accès aux nœuds parents, descendants et ancêtres d'un nœud donné. Il permet de calculer la distance entre ces deux nœuds.

– Index du contenu d'élément : permet de retrouver les éléments dans lesquels un terme apparaît. La pertinence des termes est calculée par le TF-IEF (Term Frequency - Inverse Element Frequency).

– Index ontologie : permet de retrouver des mots reliés sémantiquement à un mot donné. Il calcule pour cela une similarité qui peut être restreinte à un certain type de liens. A partir de cette valeur une mesure de similarité peut être calculée entre deux concepts.

Le langage XXL permet d'interroger les documents XML avec une syntaxe proche de la syntaxe SQL. En effet, il est basé sur les langages de requêtes tels que XML-QL et XQuery auxquels il ajoute un opérateur de similarité sémantique noté « ~ ». Cet opérateur permet d'exprimer des conditions de similarité sémantique sur les éléments ainsi que sur leur contenu textuel. L'évaluation de la requête se base sur un calcul de similarité dans une ontologie.

2.1.3 Approches orientées contenu

Dans les approches orientées contenu, on indexe uniquement le contenu textuel. Dans [11], les auteurs proposent d'utiliser les graphes conceptuels pour indexer les nœuds feuilles (porteuses du contenu textuel). Les index des autres nœuds sont obtenus par l'agrégation des index (graphes conceptuels) des nœuds fils en utilisant l'opérateur de jointure maximale entre les graphes conceptuels. Le mécanisme d'interrogation proposé se base sur l'opérateur de projection. L'approche proposée présente des limites. La jointure entre deux graphes conceptuels nécessite la présence d'un concept plus spécifique commun entre les deux graphes à joindre. Par la suite il est impossible de construire les index. Dans ce modèle, le traitement des résultats se fait de manière booléenne. Par conséquent, il est impossible d'attribuer un score de pertinence à un élément. L'indexation par les graphes conceptuels consiste à extraire les concepts et relations entre concepts à partir du texte, ce qui est très difficile. Cette difficulté est essentiellement due à l'absence d'une ressource sémantique riche en termes de relations.

[12] propose une indexation sémantique des documents XML. Ainsi le contenu textuel (nœuds feuilles dans l'arbre XML) est indexé par un ensemble de concepts en utilisant la ressource sémantique WordNet[11] . Cette approche présente une

[11] Wordnet :http://wordnet.princeton.edu/

extension de mesure de similarité entre deux concepts en se basant sur la mesure de Wu-Palmer [15]. La mesure de similarité entre concepts définie précédemment est utilisée pour désambiguïser le sens des termes en favorisant le sens rattaché au concept qui maximise la densité du réseau sémantique. L'originalité de l'approche consiste principalement dans la mesure de similarité utilisée pour enrichir la méthode de pondération des termes. Le score de pertinence d'un élément est calculé en utilisant le modèle vectoriel.

XOntoRank [13] est un système de recherche de documents médicaux. Ce système utilise le thesaurus sémantique SNOMED[12] pour l'extraction et la pondération des concepts à partir du contenu textuel. L'évaluation du score de pertinence d'un élément XML se base sur le principe de propagation de pertinence. Ainsi, le score est calculé à partir des scores des éléments fils en utilisant une fonction d'agrégation.

2.2 Mesures de similarité sémantique

L'objectif des mesures de similarité sémantique est d'évaluer la proximité sémantique entre les concepts (auxquels les termes des requêtes et documents sont rattachés). En recherche d'information, les mesures de similarité jouent un rôle important, en particulier dans le processus de désambiguïsation des concepts, la pondération des concepts et l'évaluation de la pertinence. De nombreuses approches ont été proposées pour évaluer la similarité sémantique entre deux concepts. Ces approches se divisent [14] en trois catégories : les approches basées sur les arcs, les approches basées sur le contenu informationnel et les approches hybrides.

2.2.1 Approches basées sur les arcs

Ce type de mesure s'appuie sur la structure de la ressource sémantique en proposant un comptage plus ou moins élaboré du nombre d'arcs séparant deux concepts. Ces mesures se servent de la structure hiérarchique de l'ontologie pour déterminer la similarité sémantique entre les concepts. Parmi les travaux classifiés sous cette bannière on peut citer :

La mesure de Wu-Palmer [15]. Dans une ontologie, la similarité est définie par rapport à la distance qui sépare deux concepts dans la hiérarchie et également par leur position par rapport à la racine. La similarité entre c_1 et c_2 est :

$$\mathrm{sim}_{\mathrm{WPalmer}}(c_1, c_2) = \frac{2*\mathrm{prof(c)}}{\mathrm{dist}(c_1,c)+\mathrm{dist}(c_2,c)+ 2*\mathrm{prof(c)}}. \quad [1]$$

[12] Snomed: http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html

Où *c* est le concept le plus spécifique qui subsume les deux concepts c_1 et c_2, *prof(c)* est le nombre d'arcs qui sépare c de la racine et $dist(c_i,c)$ le nombre d'arcs qui séparent c_i de *c*.

La mesure de Zargayouna [12]. Cette mesure de similarité est inspirée de celle de [15]. Le lien père-fils est ainsi privilégié par rapport aux autres liens de voisinage en adaptant la mesure de Wu-Palmer. L'adaptation de la mesure est faite au travers de la fonction de calcul du degré de spécialisation d'un concept (spec) qui mesure sa distance par rapport à l'anti-racine.

$$\mathrm{sim}_{\mathrm{Zargayouna}}(c_1, c_2) = \frac{2*\mathrm{prof(c)}}{\mathrm{dist}\,(c_1,\mathrm{c}) + \mathrm{dist}\,(c_2,\mathrm{c}) + 2 * \mathrm{prof}\,(\mathrm{c}) + \mathrm{spec}(c_1,c_2)} \quad [2]$$

$$\mathrm{spec}(c_1, c_2) = \mathrm{prof_b(c)} * \mathrm{dist}\,(c_1,\mathrm{c}) * \mathrm{dist}\,(c_2,\mathrm{c})$$

Où $prof_b(c)$ correspond au nombre maximum d'arcs qui séparent le plus petit ancêtre commun du concept « virtuel » représentant l'anti-racine.

La mesure de Resnik [16]. La similarité est définie par rapport à la longueur des chemins qui relient deux concepts dans la hiérarchie. La similarité entre c_1 et c_2 est :

$$\mathrm{sim}_{\mathrm{ResnikEdge}}(c_1, c_2) = 2\mathrm{D} - \mathrm{len}(c_1, c_2) \quad [3]$$

Où *D* est le maximum des longueurs des chemins possibles qui relient c_1 et c_2 et $len(c_1, c_2)$ le plus petit chemin entre c_1 et c_2.

2.2.2 Approches basées sur le contenu informationnel

La notion de contenu informationnel (CI) a été pour la première fois introduite par Resnik [17]. Le contenu informationnel d'un concept traduit la pertinence d'un concept dans le corpus en tenant compte de sa spécificité ou généralité. La fréquence de concepts dans le corpus est calculée pour retrouver le contenu informationnel. Cette fréquence regroupe la fréquence d'apparition du concept lui-même ainsi que des concepts qu'il subsume. La formule est la suivante :

$$\mathrm{CI(c)} = -\log\big(\mathrm{P(c)}\big) \quad [4]$$

Parmi les mesures basées sur le contenu informationnel on peut citer :

La mesure de Resnik [17]. La similarité entre c_1 et c_2 est définie de la façon suivante :

$$\mathrm{sim}_{\mathrm{Resnik}}(c_1, c_2) = \mathrm{CI}(\mathrm{c}) \quad [5]$$

Où c est le concept le plus spécifique qui subsume les deux concepts c_1 et c_2.

La mesure de Lin [18]. La similarité entre deux concepts est mesurée par le ratio du contenu d'information nécessaire pour mesurer la "communalité" des deux concepts sur le montant du contenu d'information nécessaire pour décrire chacun des deux concepts. La communalité entre deux concepts dépend du contenu d'information (CI) de leur concept commun le plus spécifique (LCS)

$$\mathrm{sim}_{\mathrm{Lin}}(c_1, c_2) = \frac{2 \times \mathrm{CI}(\mathrm{LCS}(c_1, c_2))}{\mathrm{CI}(c_1) + \mathrm{CI}(c_2)} \quad [6]$$

2.2.3 Approches hybrides

Ces approches sont fondées sur un modèle mixte qui combine des approches basées sur les arcs (distances) en plus du contenu informationnel qui est considéré comme facteur de décision.

La mesure de Jiang [19]. Cette mesure combine le contenu informationnel du concept le plus spécifique (dénoté par c) à ceux des concepts et le nombre d'arcs. Ainsi la similarité est :

$$\mathrm{sim}_{\mathrm{Jiang}}(c_1, c_2) = \frac{1}{\mathrm{CI}(c_1) + \mathrm{CI}(c_2) - 2 \times \mathrm{CI}(\mathrm{c})} \quad [7]$$

3 Une approche de recherche d'information sémantique dans les documents semi-structurés

3.1 Modélisation d'un document semi-structuré

Dans notre approche, nous adoptons le modèle DOM [20] où la structure d'un document est modélisée par un arbre de nœuds. Les nœuds de cet arbre sont typés (éléments, attributs, texte) et sont reliés par des relations de structure (parent-fils, ancêtre–descendant). Les nœuds feuilles représentent le contenu textuel du document, ils sont de type texte. Les autres nœuds sont des nœuds internes, ils sont de type élément.

Dans la suite on dénotera par :

- N_T : un nœud d'arbre de type texte ;
- N_E : un nœud d'arbre de type élément.

La figure ci-dessous donne un exemple de document XML sous une forme arborescente :

Document XML

```
<doc >
<titre> European Museum Guide
</titre>
<sect1 >
 <par1> The Louvre Museum is
located in Paris city and Paris is the
Capital City of France Country…
 < par1>
<par2> African Museum is located
in Lyon city... </par2>
</sect1>
<sect2 >
<par3> The Alte Nationalgalerie
Museum is located in Berlin City ….
</par3>
 </sect2>
</doc>
```

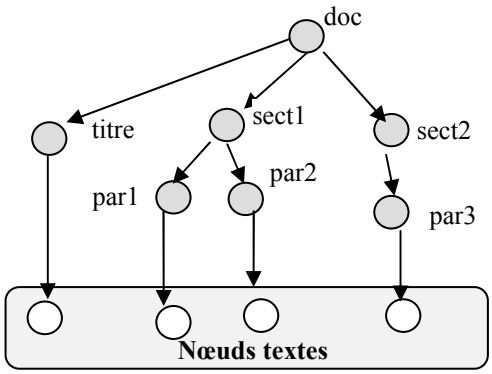

Figure 1. *Exemple de document XML sous forme arborescente.*

3.2 Vers une indexation conceptuelle du contenu textuel

Nous proposons d'utiliser le modèle vectoriel sémantique [21, 22] pour indexer le contenu textuel d'un document semi-structuré par un ensemble des concepts plutôt que par des termes. Les nœuds textes ainsi que la requête sont représentés par des vecteurs dans l'espace d'indexation. Les dimensions de l'espace d'indexation sont l'ensemble des concepts d'une ontologie Ω. Ainsi, dans un espace conceptuel d'indexation $C_\Omega=\{c_1,...c_n\}$ où les c_i sont les concepts d'indexation, un nœud texte N_T^j est représenté par un vecteur de poids des concepts.

$$\overrightarrow{\mathrm{N_T^J}} = (w_{1j}, \dots, w_{kj}, \dots w_{nj}) \quad [8]$$

Où w_{kj} est le poids du concept c_k dans le nœud texte N_T^j. De la même façon une requête q est représentée dans l'espace d'indexation C_Ω par un vecteur des poids des concepts qui composent la requête.

$$\vec{q} = (w_1, \dots, w_k, \dots w_n) \quad [9]$$

3.2.1 Extraction des concepts

Le processus d'extraction des concepts consiste à détecter les concepts dans un contexte documentaire. Un contexte documentaire est défini comme une unité textuelle à l'intérieur d'un document, il peut représenter une phrase, un paragraphe ou un élément logique de la structure logique (les nœuds texte dans les documents XML).

Afin d'extraire les concepts, on analyse les textes à l'aide d'un analyseur morphosyntaxique [31]. Cet analyseur fournit des mots segmentés, étiquetés syntaxiquement et lemmatisés. L'énumération des termes candidats vise à repérer les séquences des mots susceptibles d'être des labels de concepts dans l'ontologie. Dans cette étape des patrons peuvent être utilisés pour extraire seulement les syntagmes nominaux. Après cette étape seuls les termes qui ont une correspondance dans l'ontologie sont retenus. Dans ce cas, on considère un terme comme étant le label d'un concept. La dernière étape de l'extraction des concepts consiste à identifier ces derniers à partir des termes. Un terme pouvant dénoter plusieurs concepts, cette étape nécessite une désambiguïsation des termes. La figure ci-dessous décrit les étapes du processus d'extraction des concepts.

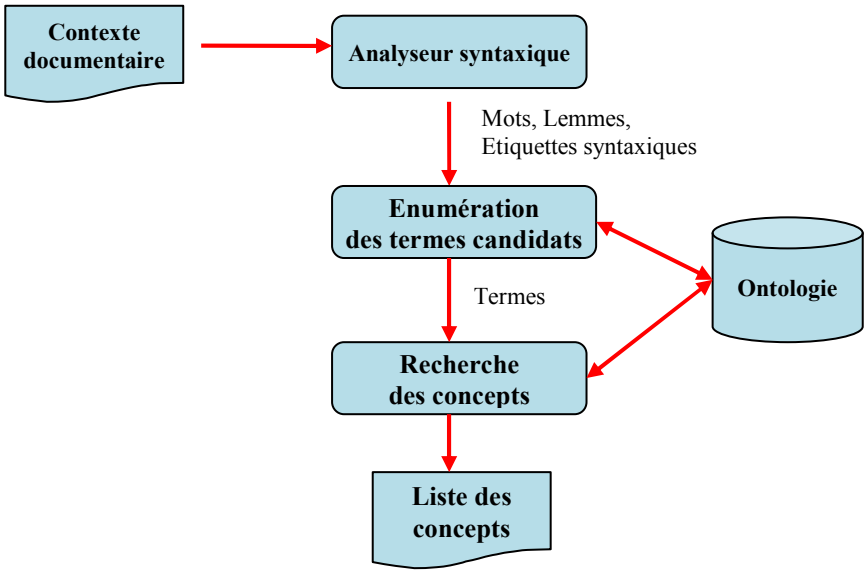

Figure 2. *Processus d'extraction des concepts [31]*.

3.2.2 Désambiguïsation des termes

On définit un contexte d'apparition CA, comme le contexte documentaire dans lequel les termes apparaissent ensemble. Le contexte d'apparition peut être une phrase ou un nœud texte.

$$CA = \{t_1, \dots, t_k, \dots t_n\} \quad [\mathbf{10}]$$

On dénote par $C_\Omega(t_k)$ l'ensemble des concepts de l'ontologie Ω ayant comme label le terme t_k.

$$C_\Omega(t_k) = \{c_k^1, \dots, c_k^i, \dots c_k^m\} \quad [\mathbf{11}]$$

Où c_k^i est le ième concept dénoté par le terme t_k et m est la cardinalité de l'ensemble $m = | C_\Omega(t_k) |$.

La désambiguïsation vise à sélectionner un seul concept parmi l'ensemble des concepts dénotés par un terme. Autrement dit, il faut sélectionner une seule combinaison des concepts (notée CB_{CA}) parmi les combinaisons possibles des concepts du contexte d'apparition CA .

$$CB_{CA} = \left\{ c_1^{i_1}, \dots, c_k^{i_k}, \dots c_n^{i_n} \right\}, \; avec \; 1 \leq i_k \leq | C_\Omega(t_k) | \quad [\mathbf{12}]$$

Le nombre de combinaisons possibles est : $\prod_1^{k=n} | C_\Omega(t_k) |$.

En prenant exemple sur les travaux de Baziz [23], nous avons utilisé le contexte d'apparition pour la désambiguïsation ainsi que l'hypothèse de Harris [24] selon laquelle les mots qui apparaissent dans des contextes similaires tendent à avoir des sens proches. De cette façon, on sélectionne la combinaison des concepts dans laquelle les concepts sont très proches. La proximité sémantique entre les concepts peut être évaluée par l'utilisation des mesures de similarités sémantiques. La mesure de similarité est généralement une fonction à deux paramètres : les deux concepts considérés. Dans notre travail, nous proposons la définition suivante pour la similarité sémantique : soient une ontologie Ω, C_Ω l'ensemble des concepts de cette ontologie et c_1, c_2 deux concepts de C_Ω. Une fonction sim_Ω est une fonction de similarité définie sur C_Ω de la façon suivante :

$$sim_\Omega : \begin{cases} C_\Omega \times C_\Omega \longrightarrow [0,1] \\ (c_1, c_2) \longrightarrow sim_\Omega(c_1, c_2) \end{cases}$$

$\forall c \in C_\Omega, \; sim_\Omega(c,c) = 1$

sim_Ω est symétrique : $\forall c_1, c_2 \in C_\Omega,\ sim_\Omega(c_1, c_2) = sim_\Omega(c_2, c_1)$

$sim_\Omega(c_1, c_2)=0$ signifie que c_1 n'est pas similaire à c_2

$sim_\Omega(c_1, c_2)=1$ signifie que c_1 est fortement similaire à c_2

Pour sélectionner une combinaison des concepts, on doit calculer la similarité globale entre les concepts de cette combinaison. Soit $CB_{CA}=\{c_1,\ldots, c_k,\ldots c_n\}$ une combinaison de concepts d'un contexte d'apparition CA, la similarité globale est définie comme la moyenne des similarités *MS* entre les concepts.

$$MS(CB_{CA}) = \frac{2.\sum_{i=1}^{i=n}\ \sum_{j=i+1}^{n} sim_\Omega(c_i, c_j)}{n.(n-1)} \quad [\mathbf{13}]$$

La désambiguïsation consiste à sélectionner la combinaison des concepts CB_{CA}^{max} dont la moyenne de similarités entre les concepts est maximale ($max=ArgMax(MS(CB_{CA}^{i}))$, où CB_{CA}^{i} est la ième combinaison de concepts du contexte d'apparition *CA*.

3.2.3 Pondération des concepts

Dans la recherche de documents semi-structurés, le poids d'un terme tend à rendre compte de son importance de manière locale au sein de l'élément et de manière globale au sein de la collection. Le poids d'un terme est évalué selon trois dimensions : la fréquence d'un terme dans le nœud texte (TF); la fréquence inverse de document pour le terme (IDF) et la fréquence inverse de l'élément pour le terme (IEF).

Une étude sur la pondération des termes [25] a montré que la combinaison de TF et IEF donne la meilleure performance. Ainsi, nous adoptons cette mesure pour calculer les pondérations des concepts. Le poids d'un concept est évalué selon deux dimensions:

- CF_i^j : la Fréquence du Concept c_j dans le nœud texte N_T^i
- $IECF_j$: la Fréquence Inverse d'Elément pour le Concept c_j

$$IECF_j = log\left(\frac{|N_T|}{|N_T^{c_j}|}\right) \quad [\mathbf{14}]$$

Où $|N_T|$ est le nombre total de nœuds textes de la collection, et $|N_T^{cj}|$ est le nombre total de nœuds textes contenant le concept c_j.

Le poids d'un concept c_j dans un nœud texte N_T^i (dénoté par W_{ij}) est donné par la formule suivante :

$$W_{ij} = CF_i^j * IECF_j \quad [\mathbf{15}]$$

3.3 Appariement nœud/requête basé sur un graphe sémantique

L'appariement nœud/requête vise à attribuer des scores de pertinence aux éléments d'un document (les nœuds de type texte et les nœuds de type élément dans l'arbre XML).

3.3.1 Score de pertinence d'un nœud de type texte

Dans notre approche, nous utilisons le modèle vectoriel sémantique où dans un espace conceptuel d'indexation, un nœud texte et une requête sont représentés par deux vecteurs de poids des concepts. Généralement, on mesure la proximité entre documents et requêtes grâce au cosinus Salton [26]. Le problème du cosinus est qu'il considère comme indépendantes des dimensions proches [27, 28]. Cependant, les concepts ont des relations sémantiques entre eux. Aussi la mesure de cosinus est incapable de détecter si la représentation sémantique d'une requête est proche de la représentation du nœud texte. Par exemple si on a une requête indexée par un seul concept $\{ c_q \}$ et de la même façon pour le nœud $\{ c_n \}$, il est impossible de retrouver ce nœud dans le cas où c_q est différent de c_n, même si les deux concepts c_q et c_n sont sémantiquement très proches.

Dans notre approche, deux représentations sémantiques sont considérées comme proches si et seulement si les concepts de la requête sont proches des concepts du nœud. Pour illustrer notre approche, on représente un nœud texte ainsi qu'une requête par un graphe (voir Figure 3) pondéré dont les nœuds sont les concepts. Chaque arête de ce graphe est affectée d'un poids représentant la similarité sémantique entre les concepts.

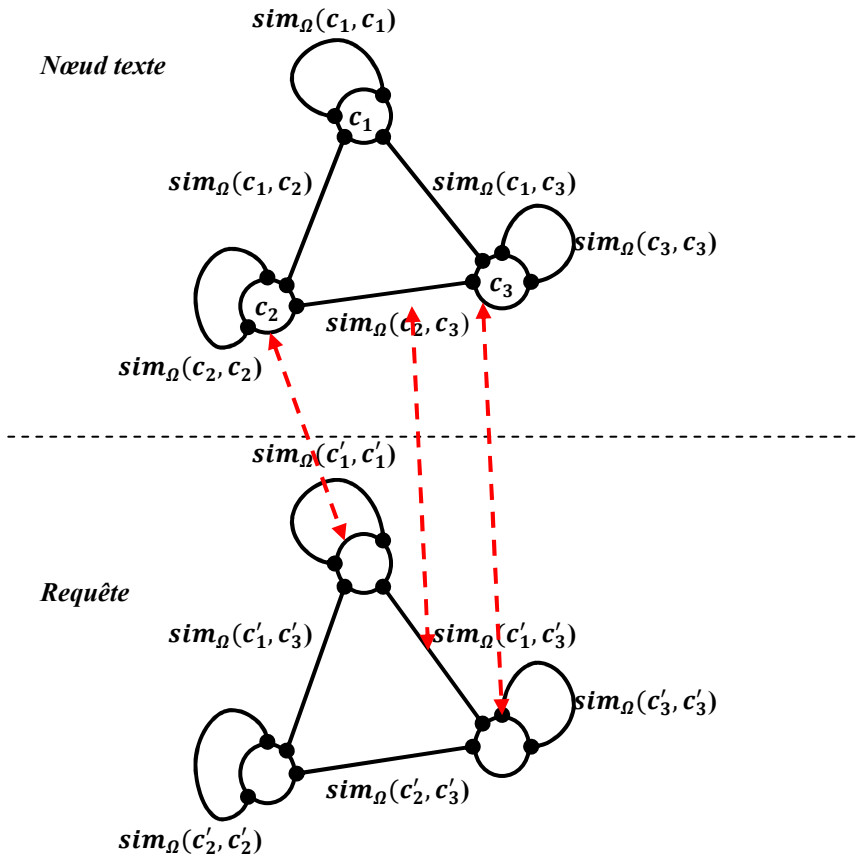

Figure 3. *Graphe sémantique d'une requête et d'un nœud texte*.

Ainsi, le calcul de score de pertinence d'un nœud texte vis-à-vis d'une requête revient à mesurer à quel point le graphe sémantique du nœud est proche du graphe sémantique de la requête. Afin d'évaluer la formule de calcul du score, on utilise la représentation classique d'un graphe sous forme de matrice. Etant donné un vecteur sémantique d'un nœud texte $N_T^j = (w_{1j}, \dots, w_{kj}, \dots w_{nj})$, la matrice (notée $M_T{}^j$) représentant le graphe sémantique du nœud est une matrice carrée d'ordre n qui est définie de la façon suivante :

$$M_T^j[a,b] = w_{aj} * w_{bj} * sim_\Omega(c_a, c_b), pour\ 1 \leq a, b \leq n \qquad [16]$$

Où w_{aj} et w_{bj} sont les poids respectifs des concepts c_a et c_b dans le nœud texte N_T^j, $sim_\Omega(c_a, c_b)$ est la mesure de similarité sémantique définie sur l'ontologie Ω entre les deux concepts c c_a et c_b. Nous avons tenu compte des poids des concepts dans la matrice. En effet, si un concept admet un poids nul, alors il n'a pas une similarité avec les autres concepts (ce concept n'existe pas dans le graphe car son poids est nul).

De la même façon, le graphe sémantique associé au vecteur requête

$\vec{q} = (w_1, \ldots, w_k, \ldots w_n)$ est représenté par une matrice carrée d'ordre *n* qui est définie de la façon suivante :

$$M_q[a,b] = w_a * w_b * sim_\Omega(c_a, c_b), pour\ 1 \leq a, b \leq n \qquad [17]$$

L'évaluation de la pertinence entre le graphe sémantique d'un nœud texte et le graphe sémantique d'une requête revient à mesurer la similarité entre les deux matrices représentant respectivement le nœud et la requête. La similarité entre les deux matrices est obtenue en utilisant la mesure de cosinus.

Définition 1. Le cosinus entre deux matrices carrées A et B de même ordre n est [29] :

$$cosinus\ (A,B) = \frac{\langle A\ , B\rangle_F}{\|A\|_F \|B\|_F} \qquad [18]$$

Où $\|A\|_F$ et $\|B\|_F$ sont les normes de Frobenius de *A* et *B*. La norme de Frobenius [29] d'une matrice carrée *A* d'ordre *n* est définie de la façon suivante :

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{i,j}|^2} \qquad [19]$$

$\langle A, B\rangle_F$ est le produit interne de Frobenius de A et B. Le produit interne de Frobenius [29] entre deux matrices carrées A et B de même ordre n est défini de la façon suivante :

$$\langle A, B\rangle_F = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j} * b_{i,j} \qquad [20]$$

Le cosinus entre deux matrices est considéré comme une mesure de similarité [29]. Cette mesure est équivalente à la mesure du cosinus entre les vecteurs dans le modèle vectoriel.

Le score de pertinence d'un nœud texte $N_T{}^j$ vis-à-vis une requête q est obtenu en utilisant la mesure de cosinus (Formule 18) entre les deux matrices $M_T{}^j$ et M_q.

$$Score\left(N_T^j, q\right) = cosinus(M_T^j, M_T^j) = \frac{\langle M_T^j\ ,\ M_q\rangle_F}{\left\|M_T^j\right\|_F \left\|M_q\right\|_F} \qquad [\mathbf{21}]$$

3.3.2 Score de pertinence d'un nœud de type élément

Dans notre approche on indexe seulement les nœuds de type texte par des vecteurs sémantiques de concepts. Comme les documents semi-structurés possèdent une structure arborescente, les index des nœuds sont imbriqués les uns dans les autres et par conséquent, l'index d'un nœud de type élément contient les index de ses nœuds descendants de type texte [11, 30]. Ainsi, les concepts des nœuds de type texte sont propagés dans l'arbre des documents. La construction des index se base sur deux hypothèses:

– **Hypothèse 1:** les concepts apparaissant près de la racine d'un sous-arbre paraissent plus porteurs d'information pour le nœud associé que ceux situés plus bas dans le sous-arbre. Autrement dit, plus la distance entre un nœud de type texte et son ancêtre est importante, moins il contribue à sa représentation.

– **Hypothèse 2:** les concepts apparaissant plusieurs fois dans les nœuds descendants sont plus porteurs d'information pour le nœud ancêtre. Autrement dit, plus un concept apparait souvent dans tous les nœuds descendants, plus il contribue à sa représentation, même si sa fréquence dans chaque nœud est faible.

Nous modélisons l'hypothèse 1 par l'utilisation dans la fonction de propagation du paramètre $dist(N_E, N_T{}^k)$, qui représente la distance entre le nœud de type élément N_E et de ses nœuds descendants de type texte et $N_T{}^k$ dans l'arbre du document, c'est à dire le nombre d'arcs séparant les 2 nœuds.

Comme nous utilisons le modèle vectoriel sémantique pour la représentation interne des index, le vecteur d'un nœud de type élément est construit à partir des vecteurs de ses nœuds descendants de type texte en utilisant l'opérateur somme entre les vecteurs. Etant donnés un nœud de type élément N_E et un ensemble des_{NE} de ses nœuds descendants de type texte : $des_{NE} = \{ N_T{}^1, \ldots, N_T{}^k, \ldots N_T{}^m \}$, le vecteur sémantique représentant le nœud N_E en tenant compte de l'hypothèse 1 est calculé de la façon suivante :

$$\overrightarrow{\mathrm{N_E}} = \sum_{k=1}^{m} \lambda^{1-\,dist(\,N_E, N_T^k)} \times \overrightarrow{\mathrm{N_T^k}} \qquad [\mathbf{22}]$$

Où $N_T{}^k$ est le vecteur sémantique représentant le k-ième nœud descendant du nœud élément $N_T{}^k$ *et* $\lambda \in]0,1]$ est un paramètre permettant de quantifier l'importance de la distance séparant les nœuds dans la formule de propagation. Ainsi, le poids w_j du concept c_j dans le vecteur N_E est :

$$\mathrm{w_j} = \sum_{k=1}^{m} \lambda^{1-dist(N_E, N_T^k)} \times \mathrm{w_{kj}} \ , pour\ 1 \leq j \leq \mathrm{n} \qquad [\mathbf{23}]$$

Où w_{kj} est le poids du concept c_j dans le vecteur $N_T{}^k$

Nous modélisons l'hypothèse 2 par l'utilisation dans la fonction de propagation du paramètre $|c_E^j|$, qui représente le nombre des nœuds texte descendants de N_E contenant le concept c_j. Plus le nombre $|c_E^j|$ est grand, plus le concept c_j contribue dans la représentation du nœud N_E. La formule de calcul de poids en tenant compte de l'hypothèse 1 et l'hypothèse 2 est :

$$\mathrm{w_j} = \sum_{k=1}^{m} \left|\mathrm{c_E^j}\right|^{\beta} \times \lambda^{1-dist(N_E, N_T^k)} \times \mathrm{w_{kj}} \ , pour\ 1 \leq j \leq \mathrm{n} \qquad [\mathbf{24}]$$

Où $\beta \in]0,1]$ est un paramètre permettant de quantifier l'importance du nombre des nœuds texte descendants contenant le concept c_j dans la formule de propagation.

Le score de pertinence d'un nœud élément vis-à-vis d'une requête est obtenu facilement en utilisant la formule 20.

$$Score(N_E, q) = cosinus(\ M_E\ , M_q\ ) = \frac{\langle M_E\ ,\ M_q \rangle_F}{\|M_E\|_F \left\|M_q\right\|_F} \qquad [\mathbf{25}]$$

Où M_E et M_q sont les matrices représentant les graphes sémantiques associés au vecteur du nœud $\mathrm{N_E}$ et au vecteur du requête q .

4 Projet d'évaluation

A court terme nous proposons de valider notre approche de recherche d'information sémantique dans les documents semi-structurés. Nous utilisons dans nos expérimentations la collection de la campagne d'évaluation INEX[13]. INEX fournit une collection de documents, un ensemble de requêtes et des jugements de pertinence, c'est-à-dire les estimations humaines des éléments pertinents concernant chaque requête. La collection actuelle d'INEX est composée de 2666190 documents en anglais extraits de l'encyclopédie en ligne Wikipedia et représente un volume d'environ 50.7GB.

[13] INitiative for Evaluation of XML retrieval : http://www.inex.otago.ac.nz/

Dans notre proposition le choix d'une ressource sémantique adaptée constitue un point déterminant pour les performances de l'approche. Il est nécessaire de disposer de mesures de similarité s'appliquant sur cette ressource. Il faut que la ressource sémantique soit généraliste. En effet la collection de test fourni par INEX est de domaine général. Ainsi, nous proposons d'utiliser le thesaurus sémantique Wordnet[14] .

La première évolution à court terme consiste dans l'implémentation d'un prototype opérationnel en mesure de nous permettre d'évaluer notre approche sur la collection de documents de la campagne d'évaluation INEX. Actuellement, nous avons développé un module en Java (voir Figure 4) permettant d'extraire les concepts à partir du texte. Ce module fait appel à l'analyseur morphosyntaxique Stanford POS Tagger et au thesaurus Wordnet. Pour la désambiguïsation des termes, des mesures de similarité sémantique sur WordNet sont disponibles dans la librairie Java WordNet::Similarity .

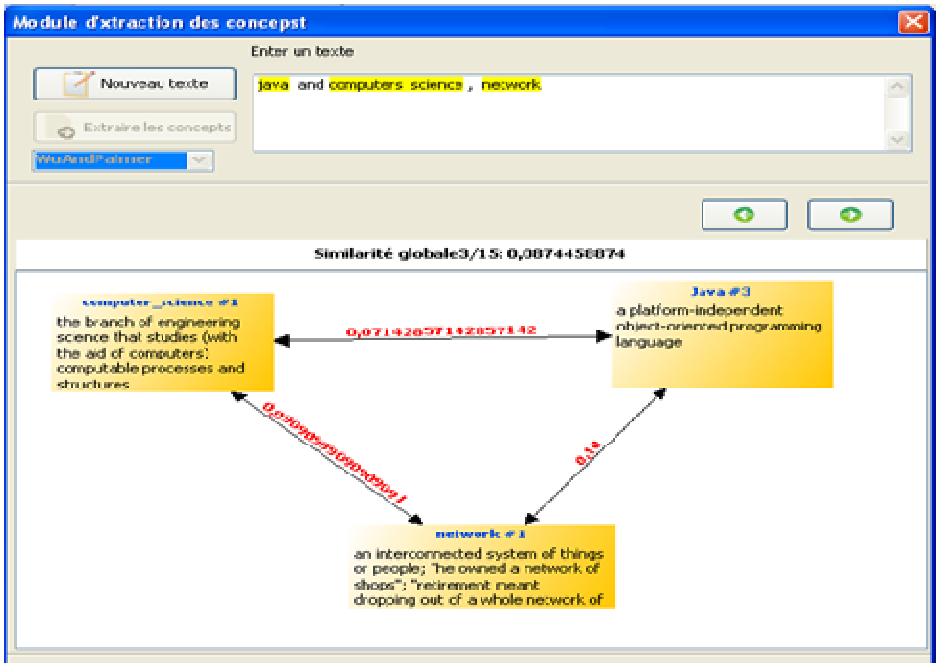

Figure 4. *Module d'extraction des concepts*.

[14] Wordnet :http://wordnet.princeton.edu/

Les résultats obtenus par ce module sont encourageants. Nous signalons que la qualité de la désambiguïsation des termes dépend fortement de la qualité de la mesure de similarité utilisée.

5 Conclusion

Nous avons présenté dans cet article un état de l'art de différentes approches de la recherche d'information sémantique dans les documents semi-structurés. Cela passe notamment par l'emploi de ressources sémantiques externes à la collection de documents, sur lesquelles il est nécessaire de disposer de mesures de similarité sémantique pour pouvoir effectuer des comparaisons entre concepts.

Nous avons proposé une représentation des nœuds de l'arbre d'un document ainsi que la requête par des vecteurs sémantiques de concepts. L'extraction des concepts se base sur un analyseur morphosyntaxique et des mesures de similarité sémantiques pour la désambiguïsation. La pondération des concepts est évaluée selon deux dimensions : la fréquence d'un concept dans un nœud et la fréquence inverse d'élément pour le concept.

La mesure de pertinence d'un nœud proposée se base sur l'évaluation de la similarité entre les graphes sémantiques du nœud et de la requête. La représentation du contenu sous forme de graphe sémantique permet d'évaluer la mesure de pertinence en utilisant la représentation matricielle des graphes. Le modèle vectoriel sémantique nous permet de rendre notre modèle d'indexation flexible du fait que le vecteur d'un nœud est construit facilement à partir des vecteurs de ses nœuds descendants.

6 Bibliographie

1. Fuller M., Mackie E., Sacks-Davis R., and Wilkinson R: Structured answers for a large structured document collection. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 204–213, (1993)

2. Grabs T. , Schek H.-J, Zurich ETH: Flexible Information Retrieval from XML with PowerDB-XML.. INEX Workshop 2002: 141-148 (2002)

3. Schlieder T. and Meuss H: Querying and Ranking XML Documents. Journal of American Society for Information Science and Tecnology (JASIST), Special Topic Issue on XML and Information Retrieval, 53(6):489–503, Apr. (2002)

4. Genest D. and Chein, M: A content-search information retrieval process based on conceptual graphs", Knowledge and Information Systems ,Volume 8 , Issue 3(September 2005), pp. 292-309 , Springer-Verlag, (2005).

5. Rosso, P., Ferretti, E., Jimenez, D., Vidal, V.: Text categorization and information retrieval using wordnet senses. Proceedings of the 2nd Global Wordnet Conference (GWC 2004), Czech Republic 299-304 (2004)

6. Bellia, Z., Vincent, N., Kirchner, S., Stamon, G.: Assignation automatique de solutions à des classes de plaintes liées aux ambiances intérieures polluées. 8èmes journées d'Extraction et de Gestion des Connaissances (EGC 2008), Sophia-Antipolis (2008)

7. Taha K. and Elmasri R: CXLEngine: A Comprehensive XML Loosely Structured Search Engine, In Proceedings of the EDBT workshop, Nantes, France 2008. ACM International Conference Proceeding Series, New York, USA; Vol. 261, ISBN:978-1-59593-966-1, pp 37-42.(2008)

8. Kim M. S. and. Kong Y.-H: Ontology-DTD Matching Algorithm for Efficient XML Query, in FSKD (2), ser. Lecture Notes in Computer Science, L. Wang and Y. Jin, Eds., vol. 3614. Springer, 2005, pp.1093-1102, (2005)

9. Weikum G., Theobald M. and Schenkel R.: Exploiting structure, annotation and ontological knowledge for automatic classification of xml data, In WebDB, San Diego, CA. 2003.

10. Schenkel R., Theobald A., and Weikum G, : Semantic similarity search on semistructured data with the XXL search engine, Information Retrieval, 8(4): pp. 521–545, (2005)

11. Chiaramella Y.: Information Retrieval and Structured Documents. In Proceedings of the Third European Summer-School on Lectures on Information Retrieval (ESSIR 2000)-Revised Lectures, pp. 286-309, (2000).

12 Zargayouna, H., Salotti, S.: Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Actes de la conférence IC'2004 (2004)

13. Farfán, F. , Hristidis V., Ranganathan A., Weiner M.,: XOntoRank: Ontology-Aware Search of Electronic Medical Records. In Proc. ICDE 2009: pp. 820-831,(2009)

14. Slimani T. Yaghlane B. B., and Mellouli K.: A new similarity measure based on edge counting. In Proceedings of world academy of science, engineering and technology, (2006).

15. Wu Z. and Palmer M.: Verb semantics and lexical selection. In 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 133 –138, (1994).

16. Resnik P.: Using information content to evaluate semantic sirnilarity. In: Proc. 14th Int. Joint Conf. Artificial Intelligence, Montreal, pp. 448-453, (1995).

17. Resnik P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research, pages 95–130,(1999).

18. Lin D.: An information-theoretic definition of similarity. In Proc. 15th International Conf. on Learning, Morgan Kaufmann, San Francisco, CA, pp. 296–304, (1998).

19. Jiang J. J and Conrath D. W.:Semantic similarity based on corpus statistics and lexical taxonomy. In International Conference Research on Computational Linguistics (ROCLING X) (1997).

20. Apparao,V., Byrne, S., Champion,M., Isaacs, S., Jacobs, I., Le Hors,A., Nicol,G., Robie,J., Sutor,R., Wilson,C., Wood,L. :Document Object Model (DOM). W3C recommendation, Technical Report REC-DOM-Level-1-19981001, (1998)

21. Woods W. Conceptual Indexing : a better way to organize knowledge. Technical Report SMLI TR-97-61 : SUN Micosystems, Lab. Mountain View Canada, (1997)

22. Berry, M. W., Z. Drmac, et E. R. Jessup : Matrices, vector spaces, and information retrieval. SIAM Rev. 41(2), 335–362(1999).

23. Baziz M, : Indexation Conceptuelle Guidée par Ontologie pour la Recherche d'Information .Thèse., Institut de Recherche en Informatique de Toulouse , Toulouse,2005.

24. Harris, Z., Gottfried, M., Ryckman, T., Mattick, P., Daladier, A., Harris, T.N., Harris, S.: The form of Information in Science: Analysis of an immunology sublanguage. Dordrecht : Kluwer Adademic Publishers (1989)

25. Sauvagnat k, Boughanem.M., Propositions pour la pondération des termes et l'évaluation de la pertinence des éléments en recherche d'information structurée. Dans : Actes de CORIA 2006, Lyon, 15-17 mars (2006).

26. Salton, G.; Wong, A. and Yang, C. S. A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620. (1975).

27. Ventresque A, Cerqueus T, Celton L, Hervouet G, Levin D., Lamarre P, Cazalens S: Mysins : make your semantic INformation system. EGC 2010: 629-630 (2010)

28. Ventresque A: Espaces vectoriels sémantiques : enrichissement et interprétation de requêtes dans un système d'information distribué et hétérogène. Thèse,Université de Nantes (2008)

29. Bing . S.: Sense Matrix Model and Discrete Cosine Transform., In Proceedings of AIRS 2004 (the first Asia Information Retrieval Symposium). Oct 18-20, Beijing, CHINA; LNCS AIRS Proceedings, Springer Verlag, (2004).

30. Abolhassani M. and Fuhr N :Applying the divergence from randomness approach for content-only search in XML documents. In Proceedings of ECIR 2004, Sunderland, pages 409-419, (2004).

31 Maisonnasse L: Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale. Thèse,Université Joseph Fourier – Grenoble I (2008)

Classification supervisée sémantique d'articles de presse en français

Samuel Gesche[1], Elöd Egyed-Zsigmond[1], Sylvie Calabretto[1], Guy Caplat[2], Jean Beney[2]

[1] Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
20, avenue Albert Einstein, 69621 Villeurbanne Cedex
prenom.nom@insa-lyon.fr

[2] INSA-Lyon, LCI, F-69621, France
20, avenue Albert Einstein 69621 Villeurbanne Cedex
prenom.nom@insa-lyon.fr

Résumé. *La classification supervisée est un champ de recherche fertile, qui ne dispose pas encore d'une théorie figée. Elle se nourrit donc régulièrement d'avancées dans d'autres domaines. Nous présentons ici nos récentes expériences en classification supervisée de textes courts de presse francophone. Ces expériences tirent parti des bases sémantiques généralistes de grande taille qui ont été récemment mises en place dans le cadre du Web sémantique. Nous effectuons un enrichissement sémantique de nos documents afin de pallier à leur petite taille qui rend les approches classiques inefficaces. L'utilisation de ressources multilingues nous permet de tirer partie des ressources disponibles en anglais.*

Mots-clés: *classification supervisée, enrichissement sémantique, textes courts*

Abstract. *Supervised document classification is a research field where a unified theory has yet to be found. Therefore, advances in other fields can often be used in order to get better results in some case or another. In this context, we present our latest experiences with large-scale semantic databases. We use these to semantically enrich small French texts for which statistical methods show a poor performance. Besides, the use of multilingual resources allows us to circumvent the fact that most resources are in English language.*

Keywords: *supervised classification, semantic enrichment, small texts*

1 Introduction

La classification supervisée de documents est un champ de recherches ancien, mais toujours actif : il n'existe pas encore de théorie unifiée permettant de définir l'algorithme optimal pour ranger des documents dans les classes suivant une problématique donnée. Cependant, de nombreuses et diverses approches sont disponibles, de l'optimisation par interface du rangement manuel aux algorithmes par apprentissage et au traitement automatique de la langue.

Les avancées récentes du Web Sémantique ont par ailleurs conduit à la compilation de grandes bases de données, souvent structurées sous formes d'ontologie. Or ces bases de données sont désormais multilingues, et peuvent être utilisées sur des matériaux en français là où la plupart des bases classiques sont limitées à l'anglais. Cela nous a amené à explorer l'utilisation de ces ressources comme catalyseurs d'une classification supervisée. Plus précisément, nous partons de documents de petite taille en français (titre et résumé de quelques dizaines de mots), à ranger dans des classes décrites de manière similaire (étiquette et descriptif) ; nous utilisons donc ces ressources sémantiques pour enrichir le contenu de ces documents et exemples afin de pouvoir raisonnablement escompter des recoupements.

Nous présenterons donc ici une approche de classification supervisée par enrichissement sémantique. Nous commencerons par présenter le contexte de notre recherche. Ensuite, nous présenterons le concept de nuage pondéré de lemmes, qui forme le centre de notre approche, ainsi que le processus d'enrichissement sémantique qui permet de les constituer. Enfin, nous présenterons nos premiers résultats, avant donner nos conclusions sur cette étude.

2. Contexte

2.1 Le projet IPRI

Le projet IPRI[15], auquel nous apportons notre contribution en tant que chercheurs en informatique, a pour objectif l'analyse du pluralisme et de la redondance dans la presse en ligne.

En effet, il existe des initiatives nationales en France qui garantissent le pluralisme de la presse écrite (comme le contrôle du temps de parole par le CSA[16]), mais rien n'existe de comparable pour la presse en ligne. La position officielle est

[15] Internet : Pluralisme et Redondance de l'Information, projet ANR jeunes chercheuses et jeunes chercheurs de 2009 à 2012, site web hébergé sur http://liris.cnrs.fr/IPRI

[16] Conseil Supérieur de l'Audiovisuel

que l'Internet est un lieu naturel de pluralisme ([1], [2]). Cependant, il s'agit aussi d'un lieu naturel de copie [3], et pas seulement en ce qui concerne les œuvres artistiques. Cette dernière thèse est renforcée par des études qualitatives qui ont pointé vers un risque de redondance du fait du métier de rédacteur Web [4]. Afin de comprendre le phénomène, nous menons actuellement une étude complète, et notamment avec un volet quantitatif pour donner du poids aux analyses plus qualitatives.

A cette fin, il nous est nécessaire, en tant que partenaires informaticiens, de fournir les outils d'analyse quantitative et qualitative nécessaires à cette étude. Nous avons donc mis en place un mécanisme de collecte d'un échantillon à peu près exhaustif de la production éditoriale d'actualités générales et politiques[17] sur Internet, échantillon défini par nos partenaires en Communication[18]. Nous travaillons désormais à la structuration de cette collection, qui est une étape nécessaire. Cette structuration passe par une classification des articles au sein de thématiques.

2.2 Positionnement

L'étude de la presse en ligne est devenue ces derniers temps un enjeu de recherche phénoménal. En effet, les agrégateurs classiques laissent de côté beaucoup de problématiques, à commencer par la fouille de données. L'enjeu est réel, puisqu'il consiste à créer la consommation journalistique de demain. En effet, les agrégateurs ont déjà amené un effacement progressif du concept de média au profit de celui de sujet. Il est donc important pour les médias qui veulent conserver leur visibilité d'offrir un service supplémentaire à celui de l'actualité (et la visite des archives est un bon départ), et pour les futurs remplaçants des agrégateurs d'offrir ce que voudra l'internaute de demain. Ainsi, on peut voir aujourd'hui un certain nombre de projets basés sur l'utilisation de ressources sémantiques pour la fouille de corpus d'archives comme ceux de l'AFP [5], la BBC [6] ou le New York Times [7]. Parallèlement, des initiatives telles que MediaCloud[19] ou des entreprises comme Linkfluence [8] proposent une approche transversale, multi-sites, mais centrée plus particulièrement sur l'actualité. L'INA ayant obtenu le dépôt légal du Web en France, leurs équipes de recherche travaillent elles aussi sur la question de la structuration d'un tel corpus.

Notre approche est plus modeste : nous ne comptons pas offrir à l'internaute de demain l'actualité, mais offrir des réponses fiables à la question du pluralisme. Nous ne voulons pas explorer en profondeur un média donné, mais acquérir une vision exhaustive de l'offre –et de la consommation, mais c'est un autre sujet–

[17] Cette catégorie, qui correspond à une labellisation officielle, est le principal champ d'investigation des études concernant le pluralisme.
[18] Notamment, les laboratoires ELICO de l'Université de Lyon et LERASS de Toulouse.
[19] http://www.mediacloud.org/

d'information. De ce fait, nous nous concentrons sur l'acquisition de la publication de l'ensemble des sites d'information générale et politique, mais nous ne pouvons nous offrir ni l'accès à toutes les bases de données, ni un nombre inconsidéré de partenariats. Nous travaillons donc à une troisième alternative, celle de tirer le maximum de ce qui est disponible directement du le Web –et donc à l'internaute. Nous avons donc des textes courts issus des flux RSS, des articles sous forme HTML pour lesquels l'hétérogénéité structurelle est la règle. Nous exposons ici comment nous traitons l'information de ces flux RSS.

3 Approche

L'enrichissement sémantique implique trois entités différentes :

- Les articles sous forme d'entrées RSS ;
- La taxonomie selon laquelle classer ces articles ;
- Les ressources sémantiques qui permettent d'effectuer la classification.

L'approche en elle-même consiste à prendre les textes courts de part et d'autre (ceux des articles et ceux de la taxonomie) et à les enrichir pour former ce que nous appelons des nuages pondérés de lemmes (figure 1). Le degré d'appartenance d'un article à une thématique de la taxonomie est ensuite calculé à partir de l'intersection de ces nuages, par analyse vectorielle.

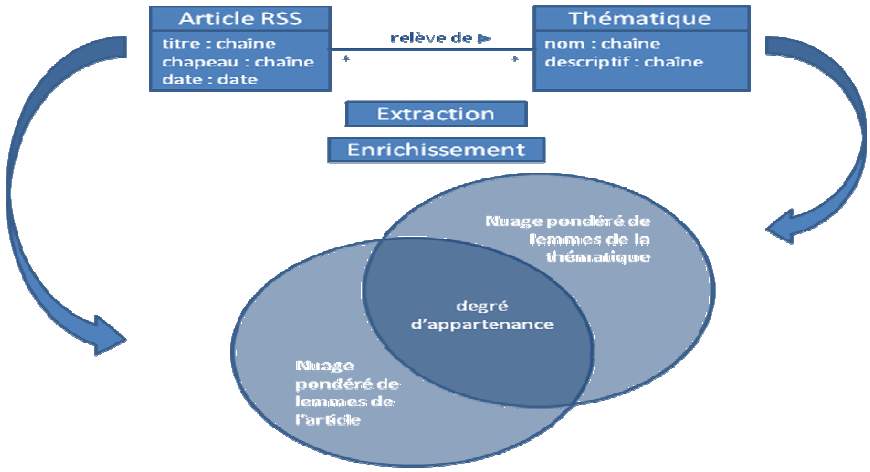

Figure 1. *Schéma général du procédé de classification par enrichissement sémantique*

3.1 Lemmatisation

Puisque nous manipulons, dans les trois entités, des termes au sein de textes construits, nous avons pris le parti d'effectuer une lemmatisation et un étiquetage lexical de ces termes. Cela nous permet d'effectuer une comparaison des termes sans devoir recourir à une analyse de similarité coûteuse en termes de temps. La lemmatisation est effectuée à la collecte pour les articles, et une fois pour toutes en ce qui concerne la taxonomie et les ressources sémantiques.

La lemmatisation diminue grandement le nombre de mots que nous avons à manipuler tout en n'introduisant que peu de bruit (les ambiguïtés, polysémies et homonymies). L'étiquetage lexical des termes nous permet de nous focaliser, en plus de cela, sur certaines natures de mots (par exemple les noms), ce qui réduit d'autant plus la charge.

Puisque nous n'effectuons nos calculs de catégorisation que sur des formes lemmatisées et étiquetées, dans la suite de notre propos, nous parlerons de *lemmes* et non plus de *termes*.

Notre approche se fonde sur la provenance des lemmes et sur leur nature. Nous faisons abstraction de tout lien que des lemmes pourraient avoir entre eux (par exemple, faire partie d'un même texte, ou être en succession l'un de l'autre). Nous bâtissons donc, à mesure que l'enrichissement progresse, des ensembles, ou nuages, de lemmes.

Un article ou une thématique est de ce fait décrite par un nuage contenant non seulement ses lemmes propres, mais également les lemmes obtenus par enrichissement.

3.2 Enrichissement

La présence de ressources sémantiques colossales est une donne relativement récente dans le monde de la recherche. En effet, depuis quelques années, nous assistons à la concrétisation du Web Sémantique par la constitution de gigantesques ressources. Cela fait longtemps qu'existent des ressources de petite taille, comme des ontologies de domaine ; en revanche, la récupération toujours plus exhaustive de l'encyclopédie Wikipedia [9] au sein de bases sémantiques généralistes comme DBPedia [10] et Yago ([11], [12]) forme une grande avancée. Ces ressources peuvent être interrogées en ligne (pour des usages modérés) ou téléchargées sous forme de bases de triplets (pour une utilisation intensive).

L'intérêt majeur de ces bases, dans notre cas, est qu'elles indexent une encyclopédie multilingue et sont donc, par essence, multilingues. Elles nous permettent donc de nous affranchir de la problématique de la traduction. Par ailleurs,

la gratuité de ces sources, et leur disponibilité directe par téléchargement ou interrogation, nous permet de nous tenir à jour relativement facilement[20].

Nous avons testé quatre procédés d'enrichissement sémantique différents à partir de ces ressources sémantiques.

Enrichissement par généralisation. Les taxonomies, même quand elles ont un descriptif, utilisent des termes généraux, alors que les articles utilisent des termes spécifiques de la problématique traitée. Ainsi, l'actualité parlera de « *Michaël Jackson* » ou de « *fuite de produit chimique toxique* » tandis que la taxonomie disposera des étiquettes « *Rock* » et « *Accident industriel* ».

L'idée est donc, puisque les ontologies commencent à fleurir sur Internet, et en particulier celles qui sont liées aux préoccupations des internautes, de remonter dans l'ontologie jusqu'à trouver le concept optimal pour le classement.

DBPedia, par exemple, fournit non seulement un ensemble colossal de concepts tirés des pages Wikipedia, mais les organise en plus au sein d'une ontologie. Cependant, cette ontologie est encore embryonnaire et orientée principalement sur les personnes, les lieux et les œuvres artistiques, qui sont les préoccupations majeures du Web Sémantique. Cette approche est donc peu efficace avec la base DBpedia elle-même.

Enrichissement par catégorisation. Une deuxième manière d'effectuer l'enrichissement de lemmes, si l'on trouve un concept dans une ressource sémantique tirée de Wikipedia, consiste à utiliser le réseau de catégories que fournit l'encyclopédie. En effet, dans un certain nombre de cas, les pages Wikipedia ont été regroupées au sein de listes thématiques. Par exemple, on pourra relier une œuvre artistique à la liste des œuvres de son auteur, de son année, de son mouvement artistique, et ce ne sont que quelques exemples. Toutes ces données peuvent être adjointes aux données initialement contenues dans le texte.

Enrichissement par spécialisation. De manière symétrique, dès lors que les catégories thématiques de la taxonomie sont suffisamment étoffées, nous pouvons décider de rester aux thématiques générales (politique, économie, société, sport etc.) au lieu de chercher la thématique précise. Dans ce contexte, il est intéressant d'enrichir ces thématiques générales par le contenu sémantique additionnel des sous-thématiques. Ainsi, le champ du sport, par exemple, contiendra le nom de tous les sports médiatiques.

Dans ce cas, il est important de tenir compte des disparités d'enrichissement. Supposons en effet, comme c'est le cas dans la taxonomie de l'IPTC, que le domaine du sport contienne quelques dizaines de sous-thématiques, alors qu'un autre domaine, comme les faits divers, n'en contiennent que quelques unes. Beaucoup d'articles relevant du fait divers risquent d'être catégorisés comme sport,

[20] Cependant, le coût de la lemmatisation de ces bases limite quelque peu l'intérêt de mises à jour fréquentes.

à cause du seul nombre des termes sportifs qui peuvent s'y trouver. Cette préoccupation diminue fortement en cas de multi-classification.

Enrichissement par champ sémantique. Cette dernière méthode consiste à utiliser la ressource sémantique comme un vecteur de champ sémantique. On retrouve ici une approche qui est déjà utilisée avec des réseaux de termes comme WordNet [13]. L'idée est, une fois que l'on a pu lier un terme du texte à un concept de la ressource, d'utiliser les termes propres de la ressource en renfort.

En particulier, quand à la ressource est adjointe une définition, nous avons presque invariablement une généralisation et une différenciation.

Dans le cas de ressources structurées (comme une base lexicale, un thésaurus voire une ontologie), il est possible de pondérer ce champ sémantique en fonction d'une distance sémantique. Cependant, dans notre cas, le champ sémantique est constitué d'un texte non structuré.

3.3 Pondération

Au final, en combinant ces méthodes d'enrichissement, nous allons nous retrouver avec des « nuages[21] » de lemmes gravitant autour des articles et des thématiques.

Tous ces lemmes n'ont pas le même poids. Certains sont présents directement dans le titre ; d'autres font partie du champ sémantique du chapeau ; d'autres encore sont le fruit d'une généralisation. Il y a des noms, des adjectifs, des verbes et des articles. Certains mots sont présents une fois, d'autres reviennent à plusieurs endroits. Toutes ces informations sont exprimées sous la forme de poids.

Nous avons opté pour une pondération multiplicative : un lemme aura un certain nombre de caractéristiques (mode d'accès, nature et nombre d'occurrences). Ces caractéristiques seront traduites en coefficients. Le poids du lemme sera le produit de ces coefficients.

Mode d'accès au lemme	Coefficient
Lemme du titre	1.0
Lemme de la description	0.75
Opération de généralisation	0.5
Opération de recherche des catégories	0.5
Opération de récupération du champ sémantique	0.33

Tableau 1. Exemple de coefficients attachés au mode d'accès

[21] Sans relation aucune avec l'artefact de visualisation, le 'nuage de mot'.

Nature du lemme	Coefficient
Nom commun	1.0
Nom propre	1.0
Verbe	0.5
Adverbe	0.0
Adjectif	0.5

Tableau 2. Exemple de coefficients attachés à la nature des lemmes

Les tableaux 1 et 2 présentent des exemples de coefficients. Pour prendre un exemple, supposons que nous obtenions un lemme en généralisant un terme du champ sémantique de l'un des lemmes du titre d'un article : le coefficient attaché à son mode d'accès sera, selon le tableau 1, de 1.0*0.33*0.5, soit 0.17. Il aura donc un poids six fois moindre que le lemme du titre lui-même dans le nuage.

4 Premières expériences

4.1 Ressources

Notre objectif est de classifier des articles de presse généralistes tirés de flux RSS (donc des textes courts comprenant un titre et une description) au sein de la taxonomie généraliste élaborée par l'IPTC (donc des classes comprenant une étiquette et une description). Pour effectuer l'enrichissement sémantique, nous utilisons la ressource multilingue DBPedia, proposant à la fois :

– l'encapsulation des pages Wikipedia, dans toutes les langues disponibles, au sein de concepts translingues étiquetés par une URI ;

– une ontologie minimale en OWL, comprenant environ 200 concepts, à laquelle sont liés ces concepts ;

– l'ensemble des catégories de Wikipedia regroupant les pages par thème (ainsi, les concepts-pages de DBPedia sont liés à des concepts-catégories) ;

– le titre et un court résumé en Français d'environ 500000 pages (contre plus de 1,2 millions en anglais).

La mise en relation de ces différentes ressources se fait de la manière suivante :

– Entre les nuages pondérés de lemmes des articles et des thématiques (comprenant les lemmes originaux auxquels ont été rajoutés les lemmes issus de l'enrichissement sémantique), la mise en relation se fait par identité : la lemmatisation est censée nous affranchir des contraintes de recherche de similarité ; deux lemmes représentés par la même chaîne de caractère, et ayant la même nature

lexicale, sont identiques ; deux lemmes présentant une différence d'un côté ou de l'autre sont différents.

– Entre un lemme et un concept de DBPedia, la question est moins immédiate : le rapprochement est effectué par minimisation de la distance d'édition entre le lemme et le titre lié au concept. Plus précisément, nous utilisons l'algorithme suivant : pour un lemme L donné, nous cherchons tous les concepts contenant un lemme représenté par la même chaîne de caractère dans leur titre ; parmi ces concepts, nous choisissons celui dont la distance d'édition entre le titre et la chaîne de caractères représentant le lemme L est minimale.

4.2 Résultats

Nous avons mené des expériences sur un échantillon de 200 articles triés à la main, à catégoriser selon la taxonomie de l'IPTC. Nous n'avons pas enrichi la taxonomie à l'aide d'exemples (cela sera fait dans de prochaines expériences).

Notre première expérience a consisté à tester le nombre de lemmes que pouvait nous fournir le champ sémantique des lemmes des thématiques de plus haut niveau de la taxonomie IPTC. Les résultats sont présentés dans la figure 2.

Comme nous pouvons le voir, l'enrichissement est relativement conséquent, autour de 20 à 30 fois le nombre de lemmes initial, à l'exception d'une thématique qui n'obtient que 13 fois ce nombre. Les résultats sont similaires en ce qui concerne les articles.

Par ailleurs, la distance de similarité (la distance d'édition) donne une excellente performance, d'autant meilleure bien sûr qu'il y a beaucoup de pages Wikipedia candidates pour un lemme.

En revanche, l'étude de l'ontologie, comme nous l'avions pressenti, donne peu de résultats, et ces résultats sont généralement toujours les mêmes. Les catégories, de même, regroupent le même type d'informations que l'ontologie (lieux, personnes et œuvres) et donne des résultats similaires. La plupart des lemmes ne sont tout simplement pas enrichis.

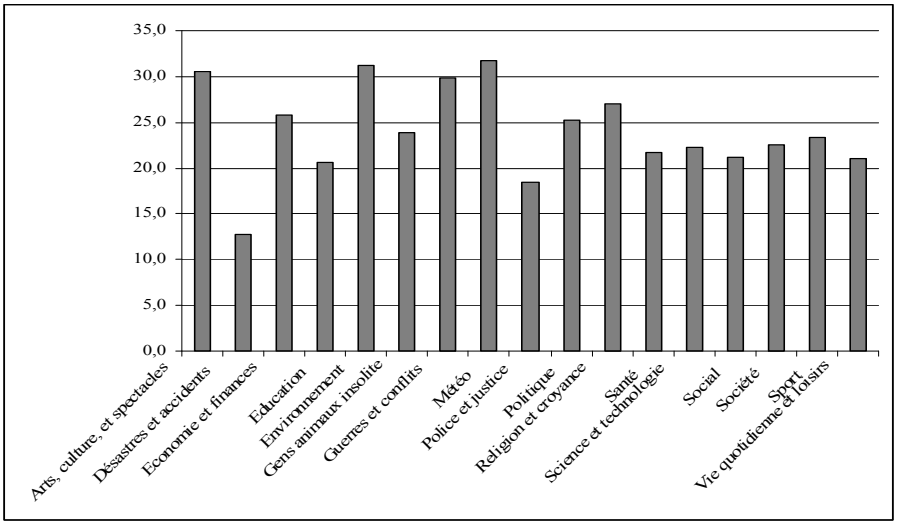

Figure 2. *Impact de l'enrichissement sémantique des thématiques de l'IPTC (en abscisse : les différentes thématiques IPTC ; en ordonnées : le facteur d'accroissement du nombre de lemmes*

Une tentative de catégorisation sans aucun enrichissement donne un résultat de l'ordre de 10% de pertinence (très peu de lemmes sont communs entre les articles et les classes) sans utiliser d'exemples. L'utilisation de l'ontologie et des catégories reste pauvre (aucun gain en performance notable n'est obtenu). En revanche, la spécialisation des thématiques par les sous-thématiques, ainsi que l'utilisation du champ sémantique, donnent des résultats meilleurs (bien que toujours limités en ce qui concerne le problème lui-même de classification). Le tableau 3 résume ces résultats. L'impact de la pondération n'est pas significative dans les expériences que nous avons faites jusqu'ici.

Procédé employé	Pertinence
Aucun	10%
Spécialisation	17%
Généralisation	10%
Catégorisation	10%
Champ sémantique	17%

Tableau 3. Présentation des résultats suivant la méthode d'enrichissement

4.3 Discussion

Contrairement à ce que nous avions espéré, l'enrichissement sémantique ne permet pas de se passer des exemples, et donc d'une classification préliminaire à la main d'une partie du corpus.

Nous n'avons pas pu étudier la pondération de manière satisfaisante. Nos premières expériences n'ont pas montré d'influence mesurable, cependant il est probable que la médiocrité des résultats impacte ce paramètre.

La question du temps réel n'était pas centrale dans nos expériences, mais elle l'est dans notre problématique. L'accès au champ sémantique d'un lemme se fait en 10 secondes environ. Cela nous amène à des temps de calculs de plusieurs minutes par article ; le temps souhaitable est de l'ordre de la seconde, voire moins (nous avons 2500 articles par jour).

Le caractère multilingue (surtout français/anglais) de notre approche est en revanche un réel apport. Nous sommes en mesure d'enrichir les textes français avec des lemmes indifféremment français et anglais en utilisant les concepts comme pivots, ce qui nous aidera à mesure que nous intégrerons d'autres ressources sémantiques.

5 Conclusion

Nous avons présenté notre méthode d'enrichissement sémantique, et nous l'avons appliqué au cas de textes courts, pour lesquels les méthodes classiques sont moins efficaces. Nous avons noté, en l'absence d'exemples, un apport dans deux cas : l'utilisation de ressources sémantiques dédiées au corpus (comme les sous-thématiques de la taxonomie IPTC sont dédiées à leurs thématiques respectives), et l'utilisation du champ sémantique d'une ressource généraliste, entendu comme un ensemble textuel tenant lieu de définition. En revanche, nous n'avons pas noté d'amélioration lors de l'emploi de ressources sémantiques spécialisées dans peu de domaines (même lorsque ces domaines sont pertinents mais non exhaustifs, comme les lieux, personnes et œuvres dans notre cas).

Le coût temporel de l'enrichissement sémantique, qui plus est, n'est pas négligeable dans le coût total de la classification. Cela n'est pas forcément rédhibitoire dans le cas d'expériences ponctuelles ; cependant, la classification en temps réel des articles collectés n'est pas envisageable de cette manière.

Afin de parfaire notre analyse, il nous reste à combiner plusieurs de ces approches entre elles, et à les utiliser en présence d'exemples.

6 Bibliographie

1. Lancelot A., Les problèmes de concentration dans le domaine des médias, Rapport pour le Premier ministre, 2005.
2. Tessier M., Baffert M., La presse au défi du numérique, Rapport pour le Ministre de la culture et de la communication, 2007.
3. Olivennes D., Le développement et la protection des œuvres culturelles sur les nouveaux réseaux, Rapport au ministre de la Culture, 2007.
4. Rebillard F., « Du traitement de l'information à son retraitement. La publication de l'information journalistique sur l'internet », Réseaux, n°137, 2006, p. 29-68.
5. Troncy R., « Explorer des actualités multimédia dans le Web de Données », Actes des 10èmes journées Ingéniérie des Connaissances IC'2009, Hammamet, Tunisie, 25-29 mai 2009, p. 181-192
6. Kobilarov G., Scott T., Raimond Y., Oliver S., Sizemore C., Smethurst M., Bizer C., LeeMedia R., « Media meets Semantic Web - How the BBC uses DBpedia and Linked Data to make Connections », In European Semantic Web Conference, Heraklion, Grèce, 31 mai-4 juin 2009.
7. Bizer C., The Emerging Web of Linked Data, IEEE Intelligent Systems, vol. 24, n° 5, 2009, p. 87-92.
8. Fouetillou G., Jacomy M., Pfaender F., « Two visions of the Web : from globality to locality », In. 2nd IEEE International Conference on Information and Communication Technologies ICTTA '06, Damas, Syrie , 24 – 28 avril 2006.
9. Wu F., Weld D., « Autonomously Semantifying Wikipedia », In 16th Conference on Information and Knowledge Management (CIKM-07), Lisbonne, Portugal, 6-8 novembre 2007, p. 41-50.
10. Becker C., DBpedia – Extracting structured data from Wikipedia. Presentation à Wikimania, Buenos Aires, Argentine, août 2009.
11. Suchanek F M., Automated Construction and Growth of a Large Ontology, Thèse de doctorat, Saarland University, 2008.
12. Suchanek F M, Kasneci G., Weikum G., « YAGO: A Large Ontology from Wikipedia and WordNet », Elsevier Journal of Web Semantics, vol. 8, n°3, 2009, p. 203-217.
13. Miller G A., WordNet: A Lexical Database for English. Communications of the ACM, vol. 38, n°11, 1995, p 39-41.

Classification multilingue et multimédia pour la recherche d'images dans le projet OMNIA

Projet OMNIA

David Rouquet— Achile Fallaise — Didier Schwab— Hervé Blanchon— Vallérie Bellynck— Christian Boitet— Emmanuel Dellandréa — Ningning Liu— Liming Chen — Alexandre Saidi — Sandra Skaff— Luca Marchesotti— Gabriela Csurka

XRCE, LIRIS, LIG-GETALP
http ://www.omnia-project.org

RÉSUMÉ. *Cet article expose différents traitements pour l'indexation automatique d'images accompagnées de textes multilingues. La combinaison de ces traitements a pour but d'obtenir des descripteurs multi-facettes d'images (thématique, affectif, esthétique, etc.). Cette approche est évaluée dans le cadre du projet ANR OMNIA qui rassemble des équipes de XRCE, du LIRIS et du LIG.*

ABSTRACT. *This article present various treatments for automatic indexation of images and multilingual texts. These treatments extract descriptors of images (thematic, emotional, aesthetic, etc.). This work is part of the OMNIA project which gathers teams of several Laboratories :XRCE, LIRIS and LIG.*

MOTS-CLÉS : *classification, sémantique, multimédia, multilingue, images, masse de données*

KEYWORDS: *classification, semantic, multimedia, multilingual, images, data base*

1. Introduction

Un des buts du projet OMNIA, financé par l'ANR de 2008 à 2011, est l'indexation automatique d'images, accompagnées de courts textes en langue naturelle, issues de grands entrepôts de données sur le web. Un aspect original est la prise en compte d'attributs affectifs et émotionnels. L'indexation est réalisée grâce aux résultats de différentes analyses visuelles des images, ainsi qu'au traitement des textes multilingues accompagnant les images. Cet article présente les traitements successifs que subissent une image et ses textes compagnons dans le système OMNIA.

Dans une première partie, nous décrivons le Classificateur Visuel Générique et l'identification de caractéristiques esthétiques développés au laboratoire XRCE. Nous verrons ensuite les travaux du LIRIS qui permettent l'utilisation de descripteurs visuels pour attribuer une émotion dans un modèle dimentionnel. Nous présentons enfin les travaux de l'équipe GETALP du LIG pour l'extraction de contenu dans des textes multilingues. Le défi que présente la fusion des résultats de tels processus hétérogène dans un descripteur unique et les travaux en cours sont décrits dans la conclusion.

2. XRCE : Analyse des images pour la classification thématique et esthétique

Nous présentons d'abord les résultats obtenus au XRCE. Les traitements ont été testés sur la base MIRFLICKR [1]. Les images sont classées selon leur contenu et leurs attributs esthétiques.

2.1. *L'analyse contextuelle*

La Catégorisation Visuelle Generique (GVC - Generic Visual Categorization), est un processus qui catégorise automatiquement les images dans un ensemble discret de classes sémantiques. Les classes pourraient être intérieur ou extérieur, naturel ou artificiel, plages ou couchers de soleil ou montagnes. On attribue à l'image des étiquettes (labels) correspondant aux objets ou concepts présents dans l'image. La GVC est un problème multi-classe, ce qui signifie que plusieurs étiquettes peuvent être attribuées à la même image. En ce sens, la GVC est souvent appelé annotation automatique d'image, car les catégories pertinentes peuvent être considérées comme des annotations attribuées automatiquement (tags, labels). Contrairement à la plupart des méthodes de détection et de reconnaissance qui sont spécifiquement développées pour une classe donnée (ex. détection de visage), la technologie GVC est indépendante des classes ou des types d'objets. Elle peut donc être qualifiée de générique et applicable sans modification spécifique des paramètres à des categories très variées comme des classes d'objets, des scènes ou des événements, des peintures, etc.

1. http ://press.liacs.nl/mirickr/

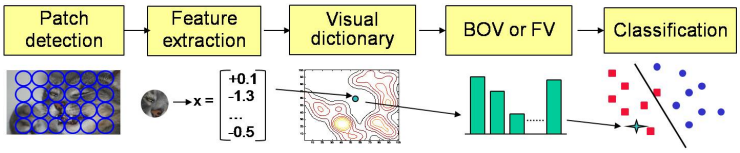

Figure 1. *Generic Visual Categorization (GVC) - pipeline*

L'approche sac de mots visuels (BOV - Bag Of Visual words) (la figure 1) est constituée par les étapes suivantes :

Détection de patch : Premièrement, les régions d'intérêt dans l'image (patches) sont détectées.

Extraction de caractéristiques : A partir de chaque patch, les caractéristiques sont extraites. Elles peuvent être ou ne pas être invariantes par transformation géométrique simple (translation, rotation, etc.).

Vocabulaire visuel : Toutes les caractéristiques extraites sont mappées à l'espace de caractéristiques et regroupés pour obtenir le vocabulaire visuel. Souvent, un K-means simple est utilisé [CSU 04], mais les Gaussian Mixture Models [PER 06] peuvent également être utilisés pour obtenir une classification douce.

Estimation d'histogramme : Un sac de mots visuels est construit en comptant le nombre de patches assignés à chaque groupe : chaque patch est attribué au mot visuel le plus proche ou à tous les mots visuels de manière probabiliste dans le cas d'un modèle de vocabulaire visuel stochastique. L'histogramme est calculé en accumulant les occurrences de chaque mot visuel.

Classement : L'histogramme ainsi obtenu peut être vue comme un vecteur de caractéristiques haut niveau de l'image et classifié par une série de OVA (one-versus-all) classificateurs (e.g. SVMs comme dans [CSU 04] une ou multi classificateurs multi classes KNN [BOS 06], pLSA [BOS 06], pour déterminer quelles catégories à attribuer à l'image.

En résumé, l'entrée de l'analyseur visuel est une image et la sortie est un ensemble de mots-clés (les concepts d'ontologie de OMNIA) associés à des degrés de confiance. Ces degrés représentent la vraisemblance qu'un concept soit effectivement représenté dans l'image.

2.2. *L'étape d'annotation*

Nous avons testé avec succés notre système GVC dans plusieurs compétitions internationales tel que Pascal VOC [EVE 06] ou ImageCLEF09 Large Scale Visual

Concept Detection and Photo Annotation Task [2]. Dans le démonstrateur de ce papier nous avons utilisé les 53 concepts de cette dernière compétition.

Figure 2. *Extrait de l'ontologie des concepts*

Ces concepts sont organisés dans une ontologie comme présentée dans la Figure 2. Une partie de la base MIRFLICKR a été renommée et manuellement annotée par plusieurs annotateurs. La compétition consistait à entrainer les classificateurs sur les 5000 images d'entrainement et à les tester sur les 13000 images de test. Le but était de décider pour chaque image quels concepts sont présents et absents tout en assurant une consistance avec les ontologies.

Nous avons entrainé notre système GVC d'une manière OVA (One Versus All) non hiérarchique, puis pour assurer les constraintes ontologique, nous avons post-traités les degrés de confiances afin d'assurer ces contraintes.

La Figure 3 montre les résultats de la compétition. Deux types de mésures sont utilisés pour évaluer les systèmes. D'un côté les mesures classiques comme EER (Equal Error Rate) et AUC (Area under Curve) mesurent la performance de chaque classificateur individuellement. De l'autre côté, la mesure hiérarchique proposée évalue la performance d'auto annotation des images en considérant l'accord entre les annotations manuelles et les annotations automatiques. Cette dernière mesure est calculée

2. http ://www.imageclef.org/2009

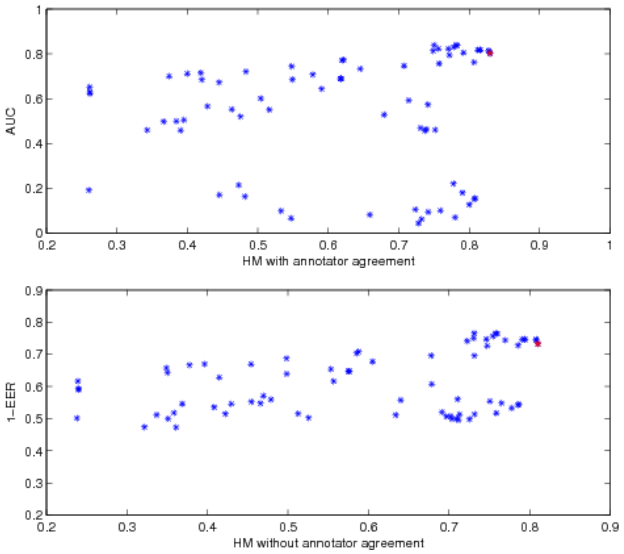

Figure 3. *En rouge : notre resultats ; en bleu : les resultats des autre systèms participants. Les mesures AUC et EER contre HM (Hierarchical Measure). Le dernier est une mesure hiérarchique devloppée par [IMA].*

sur chaque image test et moyenne à la fin. Comme le montre la Figure 3 notre système obtient la meilleur selon la mesure de EER.

Dans notre démonstrateur, nous avons appliqué ces classificateurs sur l'ensemble des images MIRFLICKR pour obtenir les annotations de ces images.

En plus de la classification basée sur le contenu que nous venons de décrire, nous considérons les aspects esthétiques suivant : luminosité, contraste, netteté, teinte et taille.

3. LIRIS : Reconnaissance de la sémantique émotionnelle portée par les images

3.1. *Introduction*

Un des buts de l'informatique, et particulièrement de l'intelligence artificielle est d'élaborer des ordinateurs intelligents qui ont la capacité d'interagir avec des êtres humains de façon naturelle. Dès lors, une des questions essentielles est de permettre aux ordinateurs de reconnaitre, de comprendre et d'exprimer des émotions [PIC 97]. Plusieurs travaux ont été faits depuis plusieurs années sur ces aspects en informatique mais également en robotique. Quand il s'agit de reconnaitre des émotions (voir [ZEN 09] pour un tour d'horizon très complet), les recherches portent principalement

sur la reconnaissance d'affects dans des données audio (parole ou musique) et sur la reconnaissance visuelle d'expressions faciales. Très peu de contributions traitent de la reconnaissance de la sémantique émotionnelle portée globalement par les images que ce soit par ses couleurs, sa composition ou tout autre élément qui peut provoquer une émotion. Face à ce sujet de recherche émergeant, un grand nombre de questions doivent être abordées concernant principalement les trois problèmes suivants : la représentation des émotions, l'extraction de caractéristiques visuelles nécessaire à la reconnaissance des émotions et les modèles de classification pour traiter les différentes propriétés des émotions [WAN 05, WEI 06, WAN 08]. En effet, comme dans tous les autres problèmes de vision par ordinateur, la principale difficulté consiste à franchir le fossé sémantique qui existe entre les descripteurs bas-niveau extraits des images et les concepts sémantiques de haut-niveau qui sont dans notre cas les émotions.

Dans cet section, nous nous proposons d'étudier l'efficacité de différents types de descripteurs visuels ainsi que les classificateurs nécessaires à la reconnaissance d'émotions dans les images. De plus, nous proposerons d'utiliser la théorie des fonctions de croyance de Dempster-Shafer [DEM 68, SME 90], qui permet la manipulation de connaissances ambigües et incertaines comme celles relatives aux émotions.

3.2. *Représentation des émotions*

Plusieurs modèles ont été étudiés dans la littérature pour représenter les émotions [ZEN 09]. Les deux principales approches sont le modèle discret et le modèle dimensionnel. Le premier modèle consiste à choisir des noms ou des adjectifs pour décrire les émotions, tels que le bonheur, la tristesse, la peur, la colère, le dégout et la surprise. Le second modèle décrit les émotions selon une ou plusieurs dimensions où chacune représente une caractérisation de l'émotion, les plus utilisées étant l'appréciation, l'activité ou le contrôle. Ce deuxième modèle permet de représenter un plus large éventail d'émotions que le premier.

Le choix de la représentation émotionnelle est généralement guidé par l'application. Ainsi, les deux approches sont utiles et peuvent même être combinées, car elles peuvent apporter des informations complémentaires. Dans cet section, nous proposons une représentation hybride comme l'illustre la figure 4. Chaque image est ainsi représentée comme un point de l'espace constitué des deux dimensions que sont l'appréciation (variant de très déplaisante à très plaisante) et l'activité (variant de très calme à très dynamique). Cet espace est divisé en quatre quadrants permettant d'obtenir quatre types d'émotions distinctes afin de caractériser la charge émotionnelle de chaque image.

3.3. *Descripteurs d'images pour la reconnaissance des émotions*

L'extraction des caractéristiques propres d'une image est une question clé pour la reconnaissance de concepts dans des images, et en particulier, les émotions. Ces

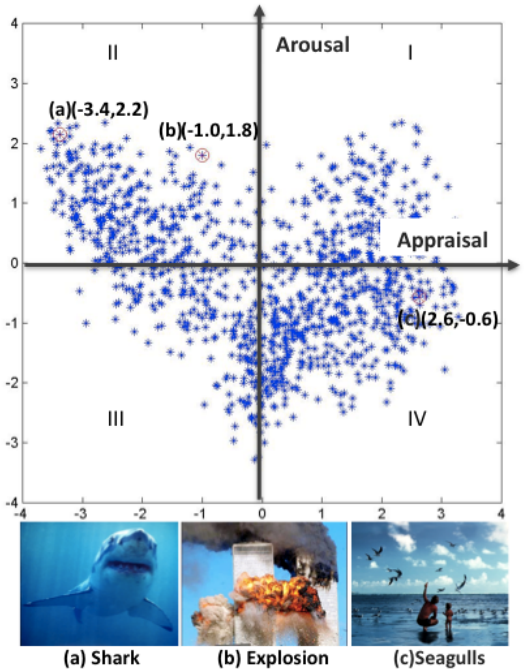

Figure 4. *Représentation des images de la base IAPS selon des critères d'activité (arousal) et d'appréciation (appraisal).*

caractéristiques doivent porter les informations nécessaires pour permettre la reconnaissance des différents concepts. Comme la reconnaissance des émotions dans les images est un domaine de recherche émergent, très peu de travaux ont été réalisés pour identifier les caractéristiques de l'image qui sont les plus efficaces dans ce contexte.

3.3.1. *Descripteurs d'images traditionnels*

La plupart des travaux traitant de la reconnaissance des émotions utilisent les mêmes descripteurs que ceux généralement exploités pour d'autres problèmes de vision par ordinateur. Les trois principales catégories de descripteurs d'images sont basées sur la couleur, la texture et la forme. En ce qui concerne la couleur, des études ont montré que l'espace HSV (Hue, Saturation, Value) est un espace de couleur qui est mieux adapté à la perception réelle des couleurs par l'homme que d'autres espaces tels que l'espace RGB traditionnel. Ainsi, sur la base de cet espace de couleur, plusieurs façons de décrire le contenu couleur des images peuvent être considérées tels que les moments de couleurs, les corrélogrammes et histogrammes de couleur ainsi que les histogrammes relatifs à la température de la couleur [DUN 09, LI 07]. En ce qui concerne la texture, la principale caractéristique demeure les matrices de cooccurrences [LI 07, WU 05]. Toutefois, les descripteurs de Tamura [WU 05] peuvent égale-

ment représenter une alternative intéressante. En effet, des descripteurs tels que la granularité, le contraste ou la directionnalité se sont avérés fortement corrélés avec la perception visuelle de l'homme. Enfin, la description des formes peut être envisagée grâce à l'extraction des contours permettant l'obtention de l'histogramme d'orientation des lignes [COL 99, WU 05] ou encore les descripteurs de Haar [DUN 09, CHO 02].

3.3.2. *Descripteurs d'images pour la reconnaissance des émotions*

Certaines tentatives ont été faites pour identifier des descripteurs de plus haut-niveau liés aux émotions. En effet, les études sur les peintures ont mis en évidence la portée sémantique des couleurs et des lignes qui y apparaissent, comme cela est rappelé dans les travaux de [COL 99] où sont proposés des descripteurs d'images plus corrélés aux émotions grâce à l'exploitation de ces informations. Ainsi, en utilisant la théorie des couleurs d'Itten, une signification émotionnelle des couleurs peut être dégagée. Tout d'abord, comme mentionné plus haut, les couleurs sont décrites en terme de teinte, de luminance et de saturation grâce à l'espace de couleur HSV, afin de se rapprocher de la perception humaine des couleurs. Ces couleurs sont ensuite projetées sur un cercle chromatique, appelé cercle d'Itten où les couleurs fortement contrastées ont des coordonnées opposées par rapport au centre du cercle. Itten a montré que les combinaisons de couleurs peuvent produire des effets tels qu'une harmonie, une disharmonie, du calme ou de l'excitation. Ainsi, l'harmonie sera détectée sur le cercle d'Itten si les positions des couleurs connectées entre elles constituent un polygone régulier. Le descripteur correspondant à cette hypothèse est obtenu en mesurant la distance entre le centre du cercle d'Itten et le centre du polygone reliant les couleurs dominantes de l'image. Ces dernières sont préalablement obtenues par un algorithme basé sur les k-means.

Les lignes portent également une information sémantique importante sur les images. En effet, des lignes obliques suggèrent le dynamisme et l'action tandis que les lignes horizontales ou verticales communiquent plutôt le calme et la détente. Pour exprimer cela en terme de descripteurs d'images, les lignes sont d'abord extraites grâce à une transformée de Hough, puis le rapport entre le nombre de lignes obliques et le nombre total de lignes dans une image est calculé.

3.4. *Modèles de classification pour la reconnaissance des émotions*

3.4.1. *Classificateurs traditionnels*

La plupart des travaux traitant de la classification des émotions dans les images reposent sur des approches traditionnelles de classification largement utilisée dans d'autres problèmes de vision par ordinateur. Malheureusement, elles ne sont pas toujours appropriées pour traiter de la spécificité des émotions. Parmi ces approches, on peut citer les réseaux de neurones [KUR 02], les machines à vecteurs supports (SVM) [DUN 09, WU 05] ou les modèles par mélange de gaussiennes[DUN 09].

3.4.2. *La théorie des fonctions de croyance*

Les émotions sont des concepts de haut-niveau sémantique qui sont, par nature, hautement subjectifs et ambigus. Ainsi, afin de s'acquitter efficacement de cette tâche de reconnaissance, il est nécessaire de traiter des informations qui peuvent être incertaines, incomplètes, équivoques et pouvant conduire à des conflits. C'est la raison pour laquelle nous proposons de faire usage de la théorie des fonctions de croyance qui gère naturellement ces difficultés.

Aperçu de la théorie des fonctions de croyance

La théorie des fonctions de croyance de Dempster-Shafer [DEM 68, SME 90] propose un cadre permettant un raisonnement sur des connaissances qui peuvent être incertaines, incomplètes et conduisant à des conflits. Cette théorie s'appuie sur des fonctions de masse qui sont une généralisation des probabilités et des mesures de possibilité. Soit $\Theta = \{\theta_1, \theta_2, ..., \theta_K\}$ l'ensemble fini des K hypothèses possibles. Cet ensemble est nommé cadre de discernement. Les concepts de base de la théorie sont les suivants :

Fonction de masse de croyance élémentaire : la fonction de masse m, associée à une source d'information donnée (un type de descripteur dans notre cas), attribue une valeur comprise dans l'intervalle $[0, 1]$ pour toute partie A de Θ (proposition) et remplit les conditions suivantes : $m(\emptyset) = 0$ et $\sum_{A \subseteq \Theta} m(A) = 1$.

$m(A)$ représente la confiance, ou croyance, que nous pouvons avoir dans la réalisation d'une proposition A. Les éléments focaux sont des sous-ensembles A tels que $m(A) > 0$. Si $m(\Theta) = 1$ alors la source est totalement incertaine alors que si $m(\theta_1) = 1$ alors la source est parfaite pour l'hypothèse θ_1.

Règle de combinaison : l'une des propriétés les plus intéressantes de la théorie des fonctions de croyance réside dans sa capacité à combiner les fonctions de masse différentes issues de plusieurs sources d'information. Considérons $m_1(.)$ et $m_2(.)$ deux fonctions de masse provenant de deux sources d'information indépendantes S_1 et S_2 respectivement. Dès lors, $m_1(.)$ et $m_2(.)$ peuvent être combinées pour obtenir la masse de la croyance engagée sur $C \subseteq \Theta$, $C \neq \emptyset$ selon la formule de combinaison suivante (Shafer, 1976) :

$$m(C) = \frac{\sum_{B \cap A = C} m_1(B).m_2(A)}{1 - \sum_{B \cap A = \emptyset} m_1(B).m_2(A)} \quad (1)$$

Une fois que les fonctions de masse des différentes sources d'informations à notre disposition sont combinées en une seule fonction de masse, une décision finale peut être prise en considérant l'hypothèse qui est associée à la valeur la plus élevée.

Construction de la croyance élémentaire

Une des principales difficultés rencontrées lors de l'élaboration d'une méthode de classification basée sur la théorie des fonctions de croyance concerne la manière dont les fonctions de masse de croyance élémentaire sont construites à partir des descrip-

teurs d'images. Dans ce travail, nous avons utilisé l'approche proposée dans [ALA 02] qui estime les fonctions de masse à partir de classificateurs en minimisant l'Erreur Quadratique Moyenne entre les résultats de la classification et les sorties attendues.

3.5. *Expérimentations*

Dans nos expérimentations, nous avons utilisé la base de données d'images IAPS qui est une base de référence en psychologie pour l'étude des émotions communiquées par les images [LAN 08]. Elle fournit une caractérisation des images selon trois critères en fonction de l'émotion produite : l'appréciation, l'activité et le contrôle. Cette base comporte 1192 images qui peuvent donc être représentées dans un espace dimensionnel des émotions, selon les axes d'appréciation et d'activité. Par commodité, cette représentation des émotions n'est pas utilisée directement, mais est utilisée pour définir 4 classes d'émotions correspondant aux 4 quadrants de la figure 4. Le corpus IAPS est partitionné aléatoirement en un ensemble d'apprentissage (80% des données, 953 images) et un ensemble de test (20% des données, 239 images). Toutes les expériences sont répétées 10 fois pour obtenir un pourcentage moyen de classification correcte. Pour évaluer la performance des différents classificateurs pour la reconnaissance des émotions dans les images, nous avons examiné quatre classificateurs représentatifs : machines à vecteurs supports (SVM), réseaux de neurones (Feed-Forward Neural Networks), Adaboost et K-plus proches voisins. Le schéma de classification que nous avons retenu consiste à utiliser deux classificateurs binaires. Le premier est entrainé pour identifier l'activité, et le second sert à identifier l'appréciation. Les résultats sont ensuite combinés pour identifier l'une des 4 classes d'émotion.

Les caractéristiques d'entrée sont générées en utilisant les techniques décrites dans la partie 3.3 et alignées en un seul vecteur, ce qui correspond à une fusion précoce. Les résultats de classification sont donnés dans la figure 5. Nous pouvons observer que les classificateurs SVM et Adaboost sont les plus efficaces avec des performances très proches, leur pourcentage moyen de classification correcte étant respectivement de 62,6% et 63,3%.

	NN (%)	*SVM (%)*	*Adaboost(%)*	*Knn(%)*
I	57.21	61.55	**65.02**	51.33
II	63.42	60.34	62.53	**64.42**
III	58.21	**62.61**	61.31	51.52
IV	61.72	**65.75**	64.30	61.71

Figure 5. *Pourcentages moyens de classification correcte pour 4 classes d'émotion obtenus par les 4 classificateurs.*

Un autre aspect intéressant consiste à comparer la capacité des différents types de descripteurs d'images à porter l'information relative aux émotions. Ainsi, le système

de classification basé sur SVM décrit précédemment a été appliqué indépendamment pour chaque type de descripteurs. Les résultats sont donnés dans la figure 6. Cette figure présente également le pourcentage de bonne classification obtenu avec la fusion de tous les descripteurs en s'appuyant sur l'approche fondée sur la théorie des fonctions de croyance à la section 3.4.2. La première remarque est que la performance entre les différents descripteurs est très similaire, variant de 53,3% pour les corrélogrammes jusqu'à 58,2% pour LBP (Local Binary Patterns). Toutefois, parmi les différents descripteurs, il semble que la texture (LBP et la matrice de cooccurrences) soit le type de descripteurs le plus efficace. En outre, les descripteurs de plus haut-niveau (dynamisme et harmonie) même s'ils peuvent paraître moins performants au premier abord ne reposent que sur une seule valeur et donc, leur efficacité est tout à fait remarquable. Enfin, il faut mentionner que l'approche proposée pour la fusion de l'ensemble des descripteurs basée sur la théorie des fonctions de croyance, et dont la matrice de confusion est donnée dans la figure 7, donne les meilleurs résultats avec un pourcentage moyen de classification correcte de 64,6%. Cette valeur montre la capacité de la théorie des fonctions de croyance à combiner différentes sources d'information et à exploiter leurs complémentarités.

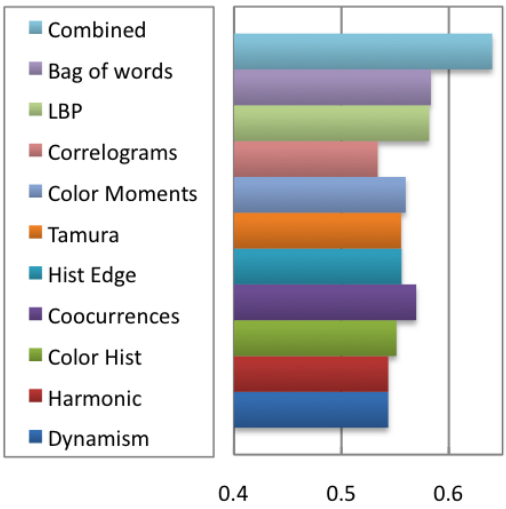

Figure 6. *Taux de reconnaissance moyens obtenus pour chaque type de descripteurs et par fusion (combined).*

Ainsi, dans le cadre du projet OMNIA, ces techniques offrent la possibilité non seulement d'attribuer une ou des étiquettes aux images correspondant aux émotions qui leur sont associées mais également de fournir pour chaque image un score selon les dimensions d'appréciation et d'activité afin de faciliter et d'améliorer la recherche dans les collections d'images en intégrant l'émotion comme critère de recherche.

Réel \ Prédit	I	II	III	IV
I	**63.32**	12.25	11.23	11.15
II	11.05	**61.42**	12.27	11.82
III	16.21	12.53	**66.19**	10.52
IV	10.42	13.80	10.31	**67.51**
Total	100	100	100	100

Figure 7. *Matrice de confusion pour les 4 classes d'émotion en utilisant la théorie des fonctions de croyance.*

4. GETALP : Extraction de contenu dans des textes multilingues

4.1. *Introduction*

Le contenu visuel des images, auquel on souhaite accéder via le système OMNIA, est indépendant de la langue utilisée dans les textes compagnons. Par exemple, une image de chien accompagnée d'une légende "dog", "perro" ou "cane" reste une image de chien [POP 07]. D'autre part, lors du processus de recherche, un utilisateur doit pouvoir formuler librement des requêtes dans une langue naturelle préférée (si possible sa langue maternelle). Ainsi, un traitement mutilingue des textes est nécessaire, tant pour la recherche que pour l'indexation.

Lors du processus d'indexation, l'analyse des textes vise à extraire le contenu pertinent pour la description des images associées. Lors du processus de recherche, les requêtes en langue naturelle doivent être transformées en requêtes formelles ne contenant que les informations pertinentes. Ces deux tâches relèvent de l'extraction de contenu (un cas particulier d'extraction d'information) et nécessitent une approche différente de celles employées en traduction automatique. Il est montré dans [DAO 06] que l'annotation de mots ou locutions avec des items sémantiques (ou "présémantiques") est une approche valide pour commencer une extraction de contenu. En particulier, ce processus ne requiert pas d'analyse syntaxique, une tâche coûteuse et dont la qualité est limitée.

Sans entrer dans le détail des modules qui composent traditionnellement un système d'extraction d'informations [GRI 97], nous pouvons dire qu'ils sont développés en fixant un certain nombre de paramètres (e.g. le domaine ou le type des entrées). La langue des entrées est bien sûr l'un de ces paramètres et il est montré dans [HAJ 07] que le portage linguistique de tels systèmes est grandement facilité si l'on dispose d'une représentation interne formalisée du contenu des textes.

Ainsi, en vue de permettre une extraction de contenu dans des textes multilingues, nous proposons de les représenter et d'effecter les traitements initiaux avec le formalisme des Systèmes-Q [COL 70] et d'annoter les mots ou locutions de ces textes avec les lexèmes interlingues (Universal Words, UW) du langage pivot UNL (Universal

Network Language). Ce processus peut être vu comme une lemmatisation interlingue qui n'est aujourd'hui proposée par aucun logiciel de lemmatisation. On peut également le qualifier d'annotation présémantique des textes.

Notre méthode est testée sur une base de données, contenant 500K images accompagnées de textes d'une cinquantaine de mots (environ 2,5M mots au total), fournie par l'agence de presse Belge *Belga News* pour la campagne *CLEF09*. Ce corpus n'est disponible qu'en anglais mais permet de tester le passage à l'échelle de notre méthode. Il a été traduit automatiquement dans 5 langues pour de futures expériences sur le multilinguisme.

4.2. *Ressources et structures de données*

4.2.1. *UNL (Universal Networking Language)*

[Uch 09] fait référence à trois choses différentes :

1) un projet international impulsé en 1996 par l'UNU (Université des Nations Unies) à Tokyo ;

2) un langage pivot dans lequel les phrases sont représentées sous forme de graphes sémantiques fondés sur l'anglais ;

3) un format de document multilingue (aligné au niveau des phrases), intégré à HTML.

Le langage UNL [BOI 09] représente le sens d'une phrase par une structure sémantique abstraite (un hyper-graphe). Chaque arc de l'hyper-graphe est étiqueté avec une relation sémantique parmi 41 disponibles (agt, obj, aoj, pos, pls, mea, cag, etc.). Chaque noeud contient, soit un lexème interlingue appelé UW (Universal Word) et des attributs sémantiques (cardinal, aspect, intonation, flexion, etc.), soit un sous-graphe (ce qui explique la notion d'hyper-graphe).

Un UW est composé de :

1) un *mot vedette*, si possible tiré de l'anglais, qui peut être un mot, des initiales, une expression ou même une phrase complète. C'est une étiquette pour le concept qu'il représente ;

2) une liste de *restrictions* dont le but est de spécifier précisément le concept auquel l'UW fait référence.

Exemples :

– book(icl>do, agt>human, obj>thing) et book(icl>thing)

– ikebana(icl>flower_arrangement)

– go_down

Un ensemble d'UW constitue un lexique pour UNL. Idéalement, un UW fait référence de façon non ambigüe à un concept partagé par plusieurs cultures. Cependant,

les UW sont faits pour représenter les acceptions d'une langue ; nous nous trouvons ainsi des UW différentes (e.g. affection et disease) qui font référence au même concept (maladie). Les UW peuvent ainsi être qualifiées de présémantiques ou préconceptuelles.

Nous utilisons actuellement 207K UW construites automatiquement à partir des synsets du Princeton Wordnet dans le cadre du consortium U++ [Jes 09]. Ces UW sont reliés aux langues naturelles par des dictionnaires bilingues dont le stockage et la manipulation en ligne sont assurés par la plateforme PIVAX [NGU 07].

4.2.2. *PIVAX*

est une plate-forme de gestion de dictionnaires en ligne basée sur JIBIKI, une plate-forme générique pour la gestion et l'édition collaborative de ressources lexicales [SER 05, SER 09].

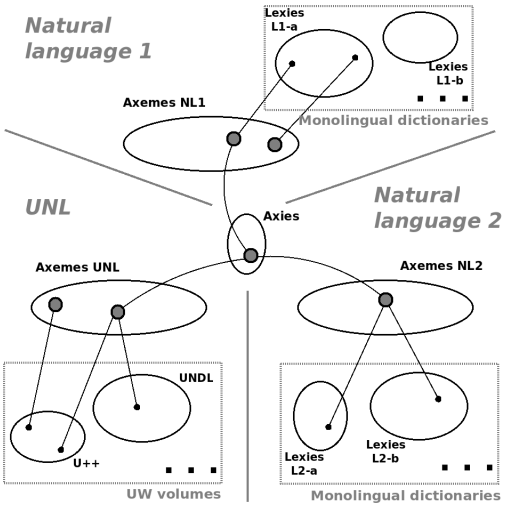

Figure 8. *Stockage de dictionnaires*

Pour chaque langue supportée, une instance de PIVAX dispose d'un espace dédié contenant :
- un ou plusieurs volumes de *lexies* correspondant aux sens des mots dans un dictionnaire ;
- un unique volume d'*axèmes* (acceptions monolingues) qui sont des liens entre les lexies synonymes d'une langue ;
- un volume partagé d'*axies* (acceptions interlingues) qui sont des liens entre les axèmes synonymes de différentes langues.

La figure ci-contre illustre le stockage de dictionnaires multilingues dans PIVAX.

4.2.3. *Les Systèmes-Q :*

Il est possible d'insérer des annotations directement au fil du texte avec des balises (e.g. XML) comme dans la table 1. Cependant, cette approche n'est pas adéquate pour représenter des ambiguïtés de segmentation (dans l'exemple suivant, il est possible de lister les différentes interprétations pour "in", mais pas de représenter "waiting", "room" et "waiting room" comme trois unités lexicales possibles).

in a waiting room
<tag uw='in(icl-sup-how), in(icl-sup-adj), in(icl-sup-linear_unit, equ-sup-inch)'>**in**</tag> <tag uw='unk'>**a**</tag> <tag uw='waiting_room(icl-sup-room, equ-sup-lounge)'>**waiting room**</tag>

Tableau 1. *Annotation naïve du fragment de texte "in a waiting room"*

Pour permettre la représentation des ambiguïtés (et notamment des ambiguïtés de segmentation), nous proposons d'utiliser les Systèmes-Q. C'est une représentation des textes dans une structure de graphe de chaà®nes (les Q-graphes), dont les arcs sont décorés par des expressions parenthésées (des arbres). De plus, les Systèmes-Q permettent des traitements à l'aide de règles de réécriture (un ensemble de telles règles est appelé système-Q).

Un exemple de ce formalisme est donné dans la figure 11 de la section 4.2.3. La figure présente successivement le code décrivant un graphe-Q, une règle de réécriture et un schéma du graphe-Q obtenu après application d'un système-Q contenant la règle de l'exemple.

Les Systèmes-Q ont été développés par Alain Colmerauer à l'Université de Montréal [COL 70]. Nous utilisons actuellement une réimplémentation élaborée par Hong-Thai Nguyen lors de sa thèse dans l'équipe LIG-GETALP en 2007 [NGU 09].

4.3. *Etapes du processus d'annotation*

4.3.1. *Vue d'ensemble :*

Le processus d'annotation est composé des étapes suivantes :

1) Lemmatisation avec un logiciel adéquat

2) Transcription des textes lemmatisés en graphes-Q ;

3) Création par PIVAX de dictionnaires locaux bilingues, sous forme de systèmes-Q dont chaque règle est de la forme lemme → lemme + UW ;

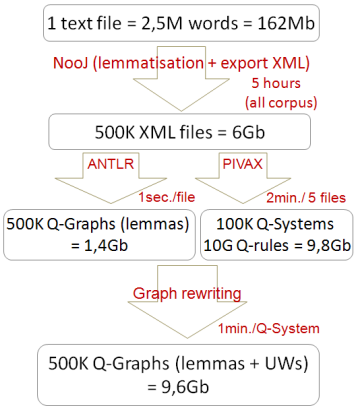

Figure 9. *Taille des données et durées des traitements (2,1GHz CPU, 2Gb RAM)*

4) Exécution de ces systèmes-Q (dictionnaires) sur les graphes-Q (textes) ;

5) Extraction de contenu sur les graphes-Q obtenus.

Dans la suite, nous décrivons en détail les quatre premières étapes. Nous présentons également des résultats expérimentaux obtenus sur la base de test Belga (500K légendes d'une cinquantaine de mots, soit environ 2,5M mots au total). La taille des données et la durée des traitements sont récapitulées dans la figure ci-contre.

4.3.2. *Lemmatisation :*

Nous utilisons des dictionnaires dont les entrées sont des lemmes ; la première étape est donc de lemmatiser le texte d'entrée. Cette étape entraine deux types d'ambiguïtés : d'une part des ambiguïtés de segmentation pour déterminer les unités lexicales, d'autre part la multitude des interprétations possibles pour une unité lexicale.

Pour l'extraction ou la recherche d'information, il est plus judicieux de conserver les ambiguïtés que de mal les résoudre. Nous avons donc besoin de lemmatiseurs qui conservent les ambiguïtés (éventuellement un pour chaque langue). Pour chacun, nous proposons d'utiliser des grammaires ANTLR [Ter 09] pour transformer les résultats en graphes-Q.

Dans notre première expérience sur le corpus Belga, nous avons utilisé le système NooJ. Il représente toutes les ambiguïtés dans un *réseau de confusion* comme illustré dans la figure 10.

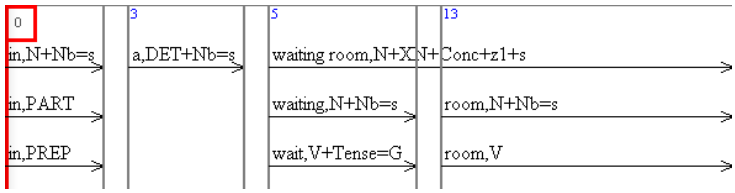

Figure 10. *Sortie de NooJ pour l'exemple"waiting room"*

4.3.3. *Dictionnaires locaux sous forme de Systèmes-Q :*

Une fois les textes annotés avec des lemmes, sous forme de graphes-Q, nous utilisons les possibilités de réécriture des Systèmes-Q pour les annoter avec des UW. Gràçe à PIVAX, nous exportons les entrées des dictionnaires bilingues sous forme de règles-Q (lemme → lemme + UW). Afin que les systèmes-Q produits soient exécutables avec des ressources raisonnables, nous créons un système-Q par texte (appelé dictionnaire local), qui ne contient que les entrées de ce texte (une cinquantaine d'entrées au maximum pour les textes de Belga). La création de ces dictionnaires locaux est la tâche la plus coûteuse en temps dans notre processus d'annotation et devra être optimisée (voir 9). Cependant, la première expérience a montré que notre méthode passe déjà bien à l'échelle.

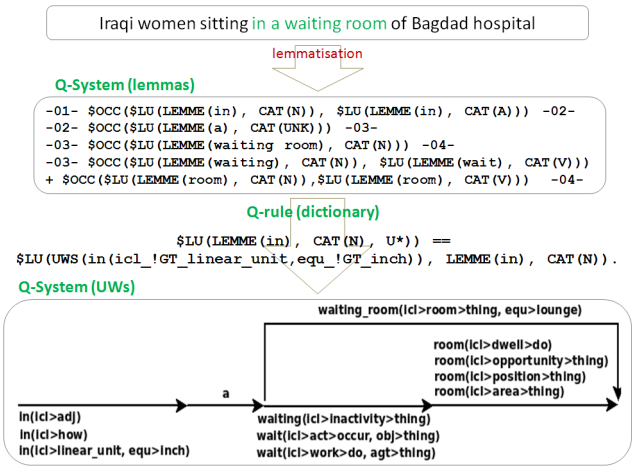

Figure 11. *Creation et execution d'un Système-Q*

4.4. *Vers une extraction de contenu interlingue*

Notre processus d'annotation donne lieu à de nombreuses ambiguïtés qui viennent tant de la lemmatisation que de l'interprétation des lemmes comme UW. Par exemple,

la figure 11 montre que l'unité lexicale "waiting room" peut être interprétée de 13 faà§ons différente avec des UW. Nous travaillons donc sur un processus de désambiguïsation. Il est basé sur l'utilisation de *vecteurs conceptuels* [SCH 05] et associe des scores de vraisemblance à chaque UW pouvant correspondre à une unité lexicale du texte.

L'extracteur de contenu que nous développons dans le cadre du projet OMNIA est guidé par une base de connaissances pour déterminer quelles sont les informations pertinentes à extraire des textes, en vue de classifier les images associées. Cette base de connaissances prend la forme d'une ontologie avec une faible expressivité logique. Comme l'extraction de contenu est faite sur des graphes-Q étiquetés par des UW (une représentation interlingue des textes), il est nécessaire que notre base de connaissances soit reliée aux lexiques d'UW. La construction et le maintien d'une correspondance entre une ontologie et un lexique interlingue est un défi, intéressant également pour la multilinguïsation d'ontologies [ROU 09].

5. Conclusion

Nous avons vu dans cet article plusieurs traitements, visuels ou textuels, permettant de recueillir des informations variées et complémentaires pour décrire différentes facettes d'une image (thématique, esthétique, affective, etc.). Chaque traitement donne lieu à un descripteur spécifique de l'image qui n'a pas été décrit en détail ici. Grossièrement, ces descripteurs ont la forme de vecteurs donc chaque composante est associée à un aspect ou une classe pour l'image (mer, montagne, active, etc.). Les valeurs scalaires stockées dans les vecteurs peuvent avoir différentes interprétations : score de vraisemblance, intensité, etc. La fusion des descripteurs spécifiques au sein d'un descripteur unique constitue la prochaine tache difficile du projet OMNIA. Pour relever ce défi, nous devrons notamment spécifier précisément la sémantique des scalaires contenus dans un descripteur et déterminer des stratégies de résolution quand les informations issues de différents descripteurs s'avèrent contradictoires.

Un autre travail important en cours est l'intégration des différents processus dans une chaine de traitement (*workflow*) pour l'indexation des images. Nous souhaitons enfin exploiter les résultats de l'indexation dans une interface graphique permettant à l'utilisateur d'exprimer des requà[a]tes spontanées dans sa langue maternelle.

6. Remerciements

Les auteurs remercient l'ANR et le projet OMNIA qui leur permettent de développer ces recherches.

7. Bibliographie

[ALA 02] AL-ANI A., DERICHE M., « A new technique for combing multiple classifier using the Dempster Shafer theory of evidence », *J. Artif. Intell. Res*, vol. 17, 2002, p. 333-361.

[BOI 09] BOITET C., BOGUSLAVSKIJ I., CARDENOSA J., « An Evaluation of UNL Usability for High Quality Multilingualization and Projections for a Future UNL++ Language », *Computational Linguistics and Intelligent Text Processing*, p. 361–373, 2009.

[BOS 06] BOSCH A., ZISSERMAN A., MUNOZ X., « Scene classification via pLSA », *ECCV*, 2006.

[CHO 02] CHO S.-B., LEE J.-Y., « A human-oriented retrieval system using interactive genetic algorithm », *IEEE Transactions on systems, man and cybernetics*, vol. 32, n° 3, 2002, p. 452-458.

[COL 70] COLMERAUER A., « Les systèmes-q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur », *département d'informatique de l'Université de Montréal, publication interne*, vol. 43, 1970.

[COL 99] COLUMBO C., BIMBO A. D., PALA P., « Semantics in visual information retrieval », *IEEE Multimedia*, vol. 6, n° 3, 1999, p. 38-53.

[CSU 04] CSURKA G., DANCE C., FAN L., WILLAMOWSKI J., BRAY C., « Visual Categorization with Bags of Keypoints », *ECCV Workshop on Statistical Learning for Computer Vision*, 2004.

[DAO 06] DAOUD D., « Il faut et on peut construire des systémes de commerce électronique à interface en langue naturelle restreints (et multilingues) en utilisant des méthodes orientées vers les sous-langages et le contenu », PhD thesis, UJF, septembre 2006.

[DEM 68] DEMPSTER A. P., « A generalization of Bayesian inference », *Journal of the Royal Statistical Society, Series B*, vol. 30, 1968, p. 205-247.

[DUN 09] DUNKER P., NOWAK S., BEGAU A., LANZ C., « Content-based mood classification for photos and music », *ACM MIR*, , 2009, p. 97–104.

[EVE 06] EVERINGHAM M., ZISSERMAN A., WILLIAMS C., GOOL L. V., « The PASCAL Visual Object Classes Challenge 2006 », 2006.

[GRI 97] GRISHMAN R., « Information Extraction : Techniques and Challenges », *International Summer School on Information Extraction : A Multidisciplinary Approach to an Emerging Information Technology*, Springer-Verlag, 1997, p. 10–27.

[HAJ 07] HAJLAOUI N., BOITET C., « Portage linguistique d'applications de gestion de contenu », *TOTh07*, Annecy, 2007.

[IMA] IMAGCLEF, « http ://ir.shef.ac.uk/imageclef/ ».

[Jes 09] JESUS CARDENOSA ET AL., « The U++ Consortium (acces en septembre 2009) », http ://www.unl.fi.upm.es/consorcio/index.php, septembre 2009.

[KUR 02] KURODA K., HAGIWARA M., « An image retrieval system by impression words and specific object names IRIS », *Neurocomputing*, vol. 43, 2002, p. 259-276.

[LAN 08] LANG P. J., BRADLEY M. M., CUTHBERT B. N., « International affective picture system (IAPS) : Affective ratings of pictures and instruction manual », *Technical Report A-8. University of Florida, Gainesville, FL*, , 2008.

[LI 07] LI C.-T., SHAN M.-K., « Emotion-based impressionism slideshow with automatic music accompaniment », *ACM Multimedia*, , 2007, p. 839-842.

[NGU 07] NGUYEN H., BOITET C., SERASSET G., « PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot », *SNLP*, Bangkok, Thailand, 2007.

[NGU 09] NGUYEN H.-T., « EMEU_w, a simple interface to test the Q-Systems (acces en septembre 2009) », http ://sway.imag.fr/unldeco/SystemsQ.po ?localhost= /home/nguyenht/SYS-Q/MONITEUR/, septembre 2009.

[PER 06] PERRONNIN F., DANCE C., CSURKA G., BRESSAN M., « Adapted Vocabularies for Generic Visual Categorization », *ECCV*, 2006.

[PIC 97] PICARD R., « Affective Computing », *MIT Press, Cambridge*, , 1997.

[POP 07] POPESCU A., « Image Retrieval Using a Multilingual Ontology », Pittsburg PA, USA, juin 2007.

[ROU 09] ROUQUET D., NGUYEN H., « Multilinguïsation d'une ontologie par des corespondances avec un lexique pivot », *TOTh09*, vol. à paraitre, Annecy, mai 2009.

[SCH 05] SCHWAB D., « Approche hybride - lexicale et thÃ©matique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte », PhD thesis, Université Montpellier 2, juillet 2005.

[SER 05] SERASSET G., « Multilingual Legal Terminology on the Jibiki Platform : The LexALP Project », *SNLP05*, Thailand, 2005.

[SER 09] SERASSET G., MANGEOT M., « The Jibiki project on LIGforge (acces en septembre 2009) », http ://ligforge.imag.fr/projects/jibiki/, septembre 2009.

[SME 90] SMETS P., « The combination of evidence in the transferable belief model », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, n° 5, 1990, p. 447-458.

[Ter 09] TERENCE PARR ET AL., « ANTLR Parser Generator (acces en septembre 2009) », http ://www.antlr.org/, septembre 2009.

[Uch 09] UCHIDA HIROSHI ET AL., « The UNDL Foundation (acces en septembre 2009) », http ://www.undl.org/, septembre 2009.

[WAN 05] WANG S., WANG X., « Emotion semantics image retrieval : a brief overview », *ACII*, , 2005, p. 490-497.

[WAN 08] WANG W., HE Q., « A survey on emotional semantic image retrieval », *ICIP*, , 2008, p. 117-120.

[WEI 06] WEINING W., YINGLIN Y., SHENGMING J., « Image retrieval by emotional semantics : A study of emotional space and feature extraction », *IEEE ICSMC*, vol. 4, 2006, p. 3534–3539.

[WU 05] WU Q., ZHOU C., WANG C., « Content-based Affective Image classification and retrieval using support vector machines », *ACII*, , 2005, p. 239-257.

[ZEN 09] ZENG Z., « A survey of affect recognition methods : audio, visual and spontaneous expressions », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, n° 1, 2009, p. 39-58.