Annotation de documents par le contexte de citation basée sur une ontologie

Lylia Abrouk

▶ **To cite this version:**

Lylia Abrouk. Annotation de documents par le contexte de citation basée sur une ontologie. Interface homme-machine [cs.HC]. Université Montpellier II - Sciences et Techniques du Languedoc, 2006. Français. NNT : . tel-00142568

HAL Id: tel-00142568

https://theses.hal.science/tel-00142568

Submitted on 19 Apr 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Numéro d'identification :

Académie de Montpellier

Université Montpellier II
— Sciences et Techniques du Languedoc —

Thèse

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de DOCTORAT

Spécialité	: **Informatique**
Formation Doctorale	: **Informatique**
École Doctorale	: **Information, Structures, Systèmes**

Annotation de documents par le contexte de citation basée sur une ontologie

par

Lylia ABROUK

Soutenue le 27 novembre 2006 devant le Jury composé de :

Rose DIENG-KUNTZ, Directrice de recherche, INRIA Sophia antipolis,.....Rapporteur
Mohand-Said HACID, Professeur, LIRIS, Université Lyon 1,.....Rapporteur
Pascal PONCELET, Professeur, Ecole des Mines d'Alès,..... Président
Chantal REYNAUD, Professeur, Université de Paris 11,..... Examinateur
Danièle HERIN, Professeur, Université Montpellier II,.....Directrice de thèse
Pierre POMPIDOR, Maître de conférences, Université Montpellier II,..... Co-encadrant
Eric MINO, Coordinateur de l'UT SEMIDE,.....Membre invité
Anne LAURENT, Maître de conférences, Université Montpellier II,.....Membre invité

Numéro d'identification :

Académie de Montpellier

Université Montpellier II
— Sciences et Techniques du Languedoc —

Thèse

présentée à l'Université des Sciences et Techniques du Languedoc
pour obtenir le diplôme de DOCTORAT

Spécialité	: **Informatique**
Formation Doctorale	: **Informatique**
École Doctorale	: **Information, Structures, Systèmes**

Annotation de documents par le contexte de citation basée sur une ontologie

par

Lylia ABROUK

Soutenue le 27 novembre 2006 devant le Jury composé de :

Rose DIENG-KUNTZ, Directrice de recherche, INRIA Sophia antipolis,.....Rapporteur
Mohand-Said HACID, Professeur, LIRIS, Université Lyon 1,.....Rapporteur
Pascal PONCELET, Professeur, Ecole des Mines d'Alès,..... Président
Chantal REYNAUD, Professeur, Université de Paris 11,..... Examinateur
Danièle HERIN, Professeur, Université Montpellier II,.....Directrice de thèse
Pierre POMPIDOR, Maître de conférences, Université Montpellier II,..... Co-encadrant
Eric MINO, Coordinateur de l'UT SEMIDE,.....Membre invité
Anne LAURENT, Maître de conférences, Université Montpellier II,.....Membre invité

Je dédie cette thèse
à mes parents

Remerciements

Je remercie respectueusement mes directeurs de thèse, le Professeur Danièle Hérin, et le Dr Pierre Pompidor pour tous les conseils et encouragements dont j'ai bénéficié tout au long de ce travail.

Je tiens à remercier tout particulièrement Mr Eric Mino, coordinateur et gérant de l'UT SEMIDE pour sa disponibilité et son soutien constant et d'avoir permis que ma thèse se déroule dans les meilleures conditions.

Mes respects et ma gratitude vont également aux membres du jury qui m'ont fait l'honneur de juger ce travail et qui par leur disponibilité, leurs observations et leurs rapports m'ont permis d'enrichir mon travail.

Je tiens à remercier les membres du laboratoire d'informatique, robotique et microélectronique de Montpellier (LIRMM) de m'avoir accueillie pendant le déroulement de la thèse.

Mes remerciements vont également aux membres de l'équipe de l'UT SEMIDE : Lidy Thomas et Jauad EL-Kherraz pour leur soutien. Je tiens à remercier tout particulièrement Mr Joel Moncel qui a initié les contacts avec les membres du SEMIDE.

Merci à Anne Laurent pour toute l'aide qu'elle m'a apportée durant les derniers mois de rédaction.

Merci à Maguelonne Tesseire, Raissi cheddy et Marc plantvit pour toutes les longues répétitions.

Je tiens à remercier quelques personnes qui réalisent déjà ou qui connaîtront l'immense satisfaction qu'on éprouve le jour de la soutenance : John Tranier, Clément Jonquet, Luc Frabresse, Céline Fiot, Jérome Chappelle, Christopher Dartnell, Patita Suksomboon, Nunes Maria Augusta, Raissi cheddy, Marc plantvit, Gregory beurier, Didier Schwab, Xavier Baril, Nicolas Vidot, Adorjan Kiss. Je tiens à remercier particulièrement Mehdi Yousfi-Monod, Christophe Crespelle et Simon Jaillet pour leurs encouragements tout au long de ma thèse.

Je tiens à remercier toutes les personnes qui ont assisté à mes pré-soutenances sachant que la plupart avaient des thèses à préparer.

Merci à Fabien et Isabelle d'avoir fait le déplacement de Reims et d'avoir été là le jour J.

Je remercie Nicole Olivet, Nadine Tilloy, Elisabeth Petiot pour leur aide.

Je remercie Sabine pour son soutien et toutes les sorties qui m'ont permis de m'échapper un peu.

Je tiens à remercier la famille Tibaoui : Sid Ahmed, Malika et leurs enfants qui sont également des amis très proches : Nabila, Mehdi et Feriel.

Je remercie mes parents pour leur soutien constant.

Je remercie mon frère Fouad, ma sœur Sihem et ma belle sœur Véronique mais également mes cousines Ouarda et Fouzia pour leur confiance.

La rédaction de la thèse nécessite de la patience et beaucoup de courage, elle aurait été impossible à réaliser sans le soutien de quelqu'un : Merci à Kadda d'avoir toujours été là pour moi.

Sommaire

Introduction et Problématique 1

1 Introduction 3

1.1 Contexte de la thèse . . . 4
1.2 Problématique . . . 5
1.3 Contributions . . . 6
1.3.1 Annotation des documents . . . 7
Outil d'annotation RAS . . . 7
1.3.2 Enrichissement de l'ontologie . . . 7
1.4 Méthodologie . . . 8
1.5 Plan de la thèse . . . 9
1.5.1 Première partie . . . 10
1.5.2 Deuxième partie . . . 10
1.5.3 Troisième partie . . . 10

I Etat de l'art 11

2 Annotations des documents 13

2.1 Processus de recherche d'informations . . . 15
2.2 Annotation des documents par le contenu . . . 16
2.2.1 Indexation classique . . . 16
2.2.2 Indexation sémantique . . . 17
2.2.2.1 Indexation avec une terminologie orientée ontologie . . . 18
2.2.2.2 Indexation dans un référentiel métier . . . 19

2.2.2.3 Indexation dans OntoSeek 20
2.2.2.4 Indexation dans le modèle DocCore 20
2.2.2.5 Indexation avec une ontologie pour la désambiguïsation 22
2.3 Annotation des documents par le contexte 23
2.3.1 Annotation des documents en utilisant le contexte de citation 23
2.3.1.1 La méthode de couplage bibliographique . . 25
2.3.1.2 La méthode de co-citations 26
2.3.2 Annotation des documents en utilisant le contexte de référencement 27
2.3.2.1 La méthode de propagation de mots clés . . 28
2.3.2.2 La méthode de propagation de métadonnées sur le *World Wide Web* 29
2.3.2.3 Méthode de propagation de signatures lexicales 31
2.3.3 Propagation d'annotations sur des images 32
2.3.3.1 Méthode de propagation d'annotations d'images 32
2.3.3.2 Méthode de propagation d'annotations de photographies 33
2.4 Discussion et Conclusion 33

3 Les ontologies 37
3.1 Introduction 39
3.2 Définitions 39
3.3 Les ontologies 41
3.3.1 De la philosophie à l'informatique 41
3.3.2 Les niveaux des ontologies 43
3.4 Principes et méthodes pour la construction d'ontologies . . . 44
3.4.1 Principes méthodologiques 44
3.4.2 Cycle de vie des ontologies 44
3.4.3 Méthodologies générales 45
3.4.3.1 Méthodes basées sur les experts 45
3.4.3.2 Méthodes basées sur le texte 46
3.4.4 L'analyse de textes 47

3.5 Différentes approches pour la construction et l'enrichissement d'ontologies . . . 49
3.5.1 Méthodes de construction d'ontologies . . . 49
3.5.1.1 Méthodes basées sur les utilisateurs . . . 49
3.5.1.2 Méthodes basées sur une analyse linguistique 50
3.5.1.3 Méthodes basées sur l'analyse formelle des concepts (AFC) . . . 52
3.5.2 Méthodes d'enrichissement d'ontologies . . . 52
3.5.2.1 Méthodes basées sur une analyse linguistique 52
3.5.2.2 Méthodes basées sur une analyse linguistique et des ressources externes . . . 55
3.6 Discussion et conclusion . . . 56

II Approche retenue : Annotation de documents basée sur une ontologie et enrichissement semi-automatique de l'ontologie 59

4 Propagation d'annotations entre les documents 61
4.1 Introduction . . . 63
4.2 Les étapes de l'annotation . . . 64
4.3 Regroupement thématique des documents . . . 66
4.3.1 Construction du graphe de co-citations . . . 67
4.3.2 Calcul de la similarité thématique entre les documents 69
4.3.3 Choix de l'algorithme pour le regroupement . . . 72
4.3.3.1 Fouille de données . . . 72
4.3.3.2 Classification non supervisée . . . 73
4.3.4 Regroupement avec fuzzy C-means . . . 77
4.4 Importation et ordonnancement des annotations . . . 79
4.5 Conclusion . . . 82

5 Enrichissement semi-automatique de l'ontologie 85
5.1 Introduction . . . 87
5.2 Prérequis . . . 88
5.2.0.1 Notions sur les treillis de concepts . . . 88

Contexte 88
Treillis 88
Correspondance de Galois 88
Concept formel 88
Treillis de galois 89
5.3 Ontologie du SEMIDE 89
5.4 Enrichissement de l'ontologie par exploitation des requêtes . 90
5.4.1 Principe de la méthode 90
Exemple de session de recherche 90
5.4.2 Représentation des sessions de recherche par le treillis de concepts 92
5.4.2.1 Exemple de représentation de requête . . . 92
5.4.2.2 Construction du treillis de Galois 93
5.4.3 Construction de la hiérarchie 95
5.5 Analyse linguistique des termes composés 96
5.5.1 Règle sur le groupe prépositionnel 97
5.5.2 Règle sur le groupe adjectival 98
5.6 Etude de l'impact sur l'annotation d'un document 100
5.6.1 Résumé des étapes de notre approche 100
5.6.2 Approche pour la révision de la phase d'annotation . 101
5.7 Conclusion 102

III Expérimentation et évaluation 105

6 L'outil RAS (Reference Annotation System) 107
6.1 Introduction 109
6.2 Les étapes d'annotation 109
6.3 Les fonctionnalités de l'outil 110
6.3.1 Visualisation d'un document existant 110
6.3.2 Insertion d'un nouveau document 112
6.3.3 Ontologie pour l'annotation 112
6.4 Le résultat de l'annotation 112
6.4.1 Calcul des similarités thématiques entre les références 112
6.4.2 Regroupement thématique des références 113

6.4.3 Importation et propagation des annotations 113

7 Constitution du corpus de tests et évaluation 117

7.1 Introduction 119
7.2 Descriptif du protocole d'expérimentation 119
7.3 Constitution du corpus de tests 120
7.3.1 Collection Citeseer 120
7.3.2 Construction de la base 121
7.3.3 Utilisation d'une ontologie 122
7.3.3.1 L'ontologie choisie 122
7.3.3.2 Dmoz 122
7.3.3.3 Annotation avec Dmoz 122
7.3.4 Calcul des indices de co-citations 123
7.4 Evaluation 124
7.4.1 Evaluation basée sur la comparaison d'annotations . 124
7.4.1.1 Description de la méthode 124
7.4.1.2 Résultat 125
7.4.2 Evaluation basée sur l'avis d'experts 127
7.4.2.1 Description de la méthode 127
7.4.2.2 Calcul de l'échantillon minimal de test . . . 127
7.4.2.3 Résultat 129
7.4.2.4 Analyse 131
7.5 Conclusion et discussion 132

Conclusion et perspectives 135

8 Conclusion 137

8.1 Synthèse 137
8.2 Résumé des principales contributions 139
8.2.1 Annotation de documents 139
8.2.2 Ontologies 139
8.2.3 Communauté du SEMIDE 140
8.3 Perspectives 140
Utilisation de la structure pour l'annotation . 140

Similarité thématique des documents 141
Utilisations d'autres algorithmes de classification 141
Ontologie par profil d'utilisateur 141

Table des figures **143**

Annexes **145**

A Schéma de la base de données de documents test **147**
A.1 Modèlisation EA du corpus 147
A.2 Schéma relationnel 148
A.3 Structure de la table *article* 148
A.4 Structure de la table *reference* 148
A.5 Structure de la table *reference_par* 148
A.6 Structure de la table *indicecocitation* 149
A.7 Structure de la table *ontologie* 149
A.8 Structure de la table *ontologie_is_a* 149
A.9 Structure de la table *article_ontology_annotation* 149
A.10 Structure de la table *resultat* 150

Annexe **147**

B Extraits de scripts RAS (Python ™) **151**
B.1 Calcul des annotations d'un document 151
B.2 Regroupement thématique par l'algorithme Cmeans 155

C Système d'information du SEMIDE **161**
C.1 Introdction 161
C.2 Les métadonnées dans le SEMIDE 162
C.2.1 Métadonnées et Web 162
C.2.2 Métadonnées dans le SEMIDE 163
C.3 Processus de recherche d'information existant au SEMIDE . 164

D L'ontologie du SEMIDE **167**

Bibliographie 173

Introduction et Problématique

1

Introduction

La gestion de l'eau représente un enjeu majeur et une priorité pour les millions de personnes confrontées aux divers problèmes liés à l'approvisionnement, la pollution, et la gestion d'un développement durable.

Les informations sur l'eau sont nombreuses, d'ordre qualitatif ou quantitatif et sont dispersées chez de très nombreux acteurs. Que ce soit au niveau local, national ou européen, le citoyen, le gestionnaire, le chercheur ou le décideur a besoin d'accéder aux informations disponibles.

Afin de répondre à ce besoin, dans le contexte de l'accord de Barcelone (1995) sur le partenariat Euro méditerranéen, la conférence Euro Méditerranéenne sur la gestion Locale de l'Eau de Marseille (nov. 1996) a mis en évidence la nécessité de disposer dans tous les pays partenaires de connaissances larges et approfondies.

L'information disponible sur ces connaissances n'existant que de façon fragmentaire, dispersée et hétérogène, il est apparu nécessaire d'engager un effort de rationalisation et de lisibilité pour la rendre facilement accessible et utilisable. C'est pourquoi, il a été décidé d'étudier les modalités de mise en œuvre d'un système d'information qui, à travers l'utilisation des moyens modernes de communication, permettrait de mettre en réseau les sources existantes : le SEMIDE[1] (Système Euro Méditerranéen d'Information sur les savoir-faire dans le Domaine de l'Eau).

Dans tous les domaines de connaissances, le Web sémantique vise à fournir un médium d'échange d'information et de savoir en partageant les ressources. En particulier, nous avons cherché à appliquer les technologies du Web sémantique afin d'offrir une plateforme technique réalisant les objectifs du SEMIDE.

Les techniques employées dans le Web sémantique afin de partager les ressources nécessitent que celles-ci soient explicitement décrites. Ces données additionnelles de description seront exploitées par les utilisateurs ou par les agents logiciels afin de localiser et extraire les ressources pertinentes répondant à un besoin particulier. Par conséquent, la description sémantique des ressources est un concept fondamental du Web sémantique. Cependant, et comme nous allons

[1] www.semide.org

le voir dans le cas du SEMIDE, le contenu des ressources n'est pas toujours disponible pour diverses raisons. C'est pourquoi, nous allons présenter des techniques afin de pallier ce problème en décrivant sémantiquement les ressources dans le système d'informations du SEMIDE afin de les rendre exploitable et partageable.

1.1 Contexte de la thèse

Dans le SEMIDE, l'information est mise à disposition par un Point Focal National (PFN) pour chaque pays et par une Unité Technique (UT) centrale.

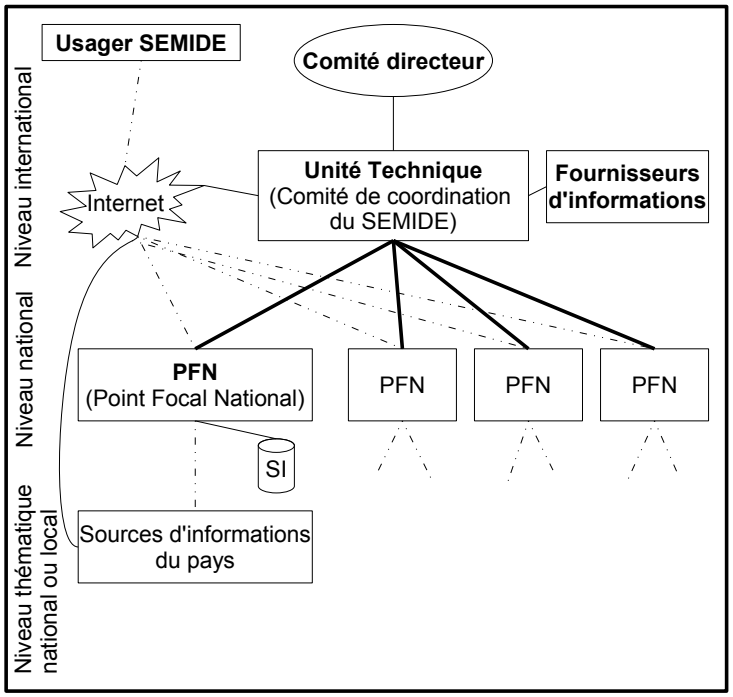

FIG. 1.1 – Organisation du semide

La figure 1.1 donne une vue d'ensemble de deux aspects, architecture organisationnelle et technique. L'aspect organisationnel regroupe :

- le comité directeur, pour le suivi et les décisions stratégiques ;
- les PFN sont composés d'équipes réduites travaillant au sein d'un organisme public ou parapublic chargé de la documentation et de l'information dans le domaine de l'eau. Leur rôle est de créer et développer un serveur national d'informations, organiser les procédures de communication et d'accès aux informations labellisées, et assurer les relations avec les usagers du pays du PFN ;
- l'UT pour soutenir les pays dans le développement de leur système national et agir en tant que point focal international ;

– le comité de coordination faisant participer tous les coordonnateurs des PFN et de l'UT.

L'architecture technique est entièrement distribuée, basée sur l'Internet pour l'accès par les utilisateurs et le partage d'information/services entre les divers nœuds (PFN) au niveau international avec les fournisseurs d'informations.

La démarche décentralisée du SEMIDE a été plébiscitée au niveau international (Banque Mondiale, Forum Mondial de l'Eau) comme un bon modèle de coopération et de développement. Aujourd'hui, le SEMIDE axe ses efforts sur l'aide aux pays partenaires méditerranéens pour la mise en œuvre de Systèmes Nationaux d'Information sur l'Eau (SNIE) interopérables. L'objectif est l'intégration transparente dans le système régional SEMIDE de services nationaux. A plus long terme, cette architecture pourrait être appliquée à une échelle géographique plus large en incluant d'autres fournisseurs de ressources.

1.2 Problématique

Malgré l'émergence de nouvelles technologies très prometteuses, à l'heure actuelle, le SEMIDE manque d'outils performants pour une mise en œuvre suffisamment flexible et à un coût modéré pour être acceptés par les différents acteurs du domaine. La diffusion de l'information au niveau du SEMIDE engendre un besoin d'accéder aux informations disponibles sur l'eau issues de sources distantes et hétérogènes. Ceci ne peut se faire qu'en arrivant à :

– identifier les connaissances disponibles dans les pays Euro Méditerranéens ;
– rendre ces informations accessibles et disponibles aux différents acteurs.

Plus précisément, les objectifs du SEMIDE sont de :

– développer le partage de l'information sur le savoir-faire dans le secteur de l'eau entre les pays partenaires ;
– faciliter l'accès à cette information en prenant en compte son hétérogénéité.

De plus, étant une initiative à long terme, le SEMIDE doit s'assurer d'intégrer les évolutions à venir, notamment vers une architecture orientée Web sémantique. Le Web sémantique étant défini comme : *"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."* [TBL01].

La recherche d'informations dans le Web actuel se base essentiellement sur la structure des documents, ce qui rend l'exploitation du contenu quasiment impossible par les machines. A la différence de cela, dans le Web sémantique, les machines accèdent aux ressources grâce à la représentation sémantique du contenu. Le Web sémantique vise à fournir un moyen d'échange et de partage des ressources et connaissances entre les différents acteurs d'un domaine. Ces ressources sont décrites par des informations structurées additionnelles :

les métadonnées. Une métadonnée est une donnée sur une donnée : c'est une information qui décrit le contenu d'une ressource et qui rend possible son exploitation dans le système d'information.

Enfin, le SEMIDE étant un système euro méditerranéen, les acteurs sont de pays différents et les ressources sont spécialisées et multilingues. Par conséquent, pour pouvoir répondre au problème du partage d'informations, nous devons pouvoir utiliser un vocabulaire commun du domaine (spécialisé au domaine de l'eau), indépendant de la langue afin de partager la connaissance entre les différents acteurs de la communauté. Les ontologies sont un des concepts de base du Web sémantique, elles représentent un vocabulaire défini par une communauté afin de représenter un domaine. Elles servent pour la structuration et l'exploitation des métadonnées.

Comme nous allons le voir dans le chapitre 2, l'annotation sémantique avec des ontologies représente une méthode pertinente pour pallier aux problèmes de recherche d'informations dans un contexte multilingue et hétérogène. Néanmoins, l'annotation sémantique soulève deux problèmes principaux :

1. Le grand volume de données et l'accès au contenu : il est peu réaliste de penser que les centaines de milliers de ressources mises à disposition des utilisateurs puissent être annotées manuellement par leurs auteurs ou les documentalistes. Cette tâche exige un effort intellectuel considérable. Il existe plusieurs types de métadonnées : sur le contenu, sur l'accès, administratives, etc. Nous nous intéressons particulièrement à l'information sur le contenu de la ressource que nous décrivons par une liste de mots clés décrivant la ressource. Par ailleurs, une ressource bien décrite doit l'être non pas par une liste plate de mots clés, mais par une liste de concepts reliés par des relations. En effet, cette liste de mots clés sera un sous ensemble des concepts d'un vocabulaire partagé par la communauté. Ces relations nous permettront de faire une meilleure recherche sur les ressources.
2. L'enrichissement d'ontologies : une ontologie correspond à un vocabulaire contrôlé et organisé, et à la formalisation explicite des relations créées entre les différents termes du vocabulaire. Ces ontologies peuvent évoluer en même temps que la masse d'informations. Il faut donc trouver des moyens d'enrichissement de cette ontologie, ainsi que l'impact que ceci peut avoir sur l'annotation des documents.

1.3 Contributions

Ce travail s'inscrit dans le cadre de la recherche d'information en utilisant les technologies du Web sémantique. Nous présentons dans ce qui suit les principales contributions de cette thèse.

1.3.1 Annotation des documents

Une ressource doit être bien décrite, sinon elle peut demeurer pratiquement inexploitable et impossible à retrouver. L'annotation des documents est une solution à ce problème, mais qui est difficile à effectuer manuellement. Notre contribution consiste à définir une nouvelle approche d'annotation de documents. Notre travail traite du problème d'annotation de ressources tels que les documents techniques et les publications.

Annoter des ressources sans avoir accès au contenu est une tâche difficile. Effectivement, dans un système tel que le SEMIDE où l'accès aux documents est restreint, généralement pour des raisons de confidentialité, les fournisseurs de ressources ne mettent à disposition qu'un ensemble de métadonnées.

Partis du constat qu'une ressource référence généralement une autre ressource, nous avons défini une approche d'annotation complètement indépendante du contenu. En utilisant le contexte de citation, nous propageons les annotations des références afin d'annoter le document citant.

Etant dans un système distribué et collaboratif, l'environnement linguistique est nécessairement hétérogène. La langue du pays ne doit pas être une barrière au partage de l'information. L'utilisation de simples mots clés marque vite la limite d'une telle solution. L'annotation dans notre approche est basée sur une ontologie qui définit le domaine. L'annotation de la ressource peut être faite en utilisant les annotations des références, l'idée étant de faire hériter une ressource des concepts d'autres ressources (références) en restant dans l'hypothèse que tous les fournisseurs partagent une ontologie.

Outil d'annotation RAS L'approche d'annotation n'est pas restée qu'au niveau théorique, nous avons développé un outil nommé RAS (Reference Annotation System) qui regroupe toutes étapes de notre approche. Cet outil est validé sur une grande base de documents techniques.

1.3.2 Enrichissement de l'ontologie

La masse d'informations étant toujours en évolution, une ontologie ne peut être figée et définitive. Elle doit pouvoir s'enrichir avec l'apparition de nouvelles ressources.

Notre contribution consiste à proposer une approche d'enrichissement de l'ontologie du SEMIDE. Elle est motivée par le manque de termes dans l'ontologie utilisée pour la recherche de documents par les utilisateurs. On exploite les termes saisis par les utilisateurs et qui sont absents de l'ontologie.

L'ontologie ayant pour but principal le partage d'un vocabulaire entre les différents acteurs du domaine, nous nous basons sur le besoin de ces derniers pour enrichir l'ontologie. Cette approche est complétée par une analyse linguistique.

1.4 Méthodologie

Les moteurs de recherche permettent de retrouver des ressources associées généralement à des mots clés résultant d'une phase d'annotation. L'attribution de mots clés à un document est une tâche difficile, voire impossible à effectuer manuellement, d'une part à cause du temps et de l'effort que cela représente, et d'autre part à cause du fait que le résultat d'une annotation peut varier d'un expert à un autre. L'utilisation des ontologies dans le processus d'annotation est motivée par plusieurs raisons. On peut citer par exemple la difficulté à laquelle se heurtent les outils de recherche lorsqu'on est dans un vocabulaire libre. Ceux-ci génèrent du bruit et du silence ne sachant pas mettre des relations entre les termes (cf. chapitre 2).

Notre approche d'annotation de documents est motivée par le constat qu'un document référence généralement d'autres documents que l'auteur estime pertinents pour leur contenu.

Prenons par exemple les documents de type événement. Un événement est un séminaire, une conférence, un cours et peut être organisé par le SEMIDE. Pour chaque événement on retrouve généralement un document Web qui le décrit et cite des références à d'autres documents pour fournir des informations complémentaires.

La figure 1.2 est un exemple de l'organisation des événements[2]. Chaque évènement cite d'autres documents, comme la page Web du point focal national organisateur de l'évènement, un document qui décrit le thème de l'évènement ou un ensemble de présentations et publications. La plupart des documents du SEMIDE ne sont pas annotés et cette tâche est quasiment impossible à réaliser manuellement, comme nous l'avons expliqué précédemment.

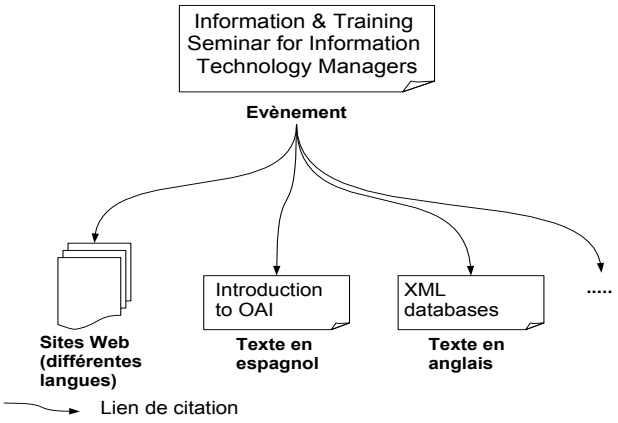

FIG. 1.2 – organisation des documents événements

Notre solution consiste à utiliser le contexte d'un document (liens de citation avec d'autres documents) afin de l'annoter, et ceci en prenant en compte

[2]le terme événement est utilisé en faisant référence au *document de l'événement*

les thèmes dans le contenu, ainsi que la similarité thématique des citations. Cette annotation suit une ontologie définie par la communauté. L'ontologie utilisée peut évoluer avec la masse de documents, elle doit être enrichie si le vocabulaire devient insuffisant pour décrire les documents, ou bien pour représenter tout simplement le domaine. Ce manque se traduit par l'insatisfaction des acteurs du système lors de la phase de recherche.

Un aperçu général de notre approche, illustrant les différentes phases, est présenté sur la figure 1.3. Les phases sont détaillées dans les chapitres suivants.

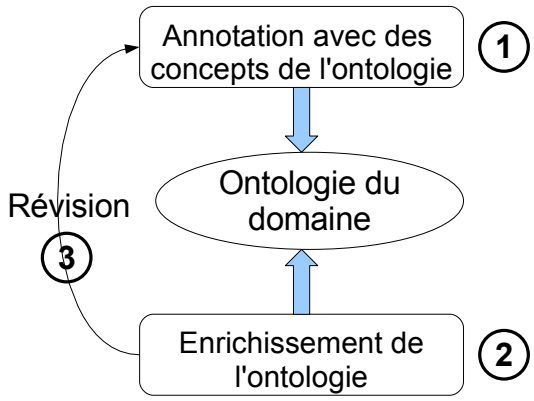

FIG. 1.3 – Méthodologie

1. Partant de l'hypothèse que tous les documents sont annotés en utilisant les concepts de l'ontologie, la première phase consistera à construire l'annotation du document D en utilisant les annotations des références de D. Cette étape est *la propagation d'annotations (phase 1)* ;
2. la deuxième phase est *l'enrichissement de l'ontologie globale*, elle consiste à mettre à jour l'ontologie du domaine par des concepts proposés par les acteurs du système *(phase 2)* ;
3. l'enrichissement de l'ontologie entraînera une *révision de l'annotation des documents (phase 3)*.

Dans ce travail de thèse, nous avons traité principalement le problème d'annotation, que nous avons complété par une proposition d'une approche pour l'enrichissement de l'ontologie et d'un algorithme pour la révision répondant aux besoins des acteurs du système.

1.5 Plan de la thèse

La suite de cette thèse est composée de trois parties.

1.5.1 Première partie

Dans la première partie, nous présentons l'état de l'art de notre travail portant sur l'annotation des documents et l'enrichissement des ontologies et nous positionnons les problèmes à résoudre et nos propositions.

Les différents travaux dans le domaine de l'annotation des documents sont présentés dans le chapitre 2. Deux aspects sont développés, celui de l'annotation en utilisant uniquement le contenu du document et celui de l'annotation en utilisant le contexte du document. Ceci nous amènera à donner des définitions sur le processus et les différents types d'indexation.

Concernant l'annotation du document en utilisant le contexte, nous présentons essentiellement l'utilisation des liens de citation et de référencement. Notre approche se situe dans ce deuxième type d'annotation.

Le chapitre 3 est consacré à une introduction générale au concept d'ontologie. Nous présentons ensuite les travaux sur la construction et l'enrichissement des ontologies.

1.5.2 Deuxième partie

Dans la deuxième partie, nous présentons notre proposition d'annotation de documents et d'enrichissement d'ontologie.

Dans le chapitre 4, l'approche d'annotations de documents en se basant sur les citations déjà annotées est détaillée. Ceci permet d'annoter une ressource sans connaître son contenu.

Dans le chapitre suivant, l'enrichissement de l'ontologie générale est traité en se basant sur la fonction syntaxique des termes ainsi que les sessions de recherche des utilisateurs. Ce processus, qui s'est avéré essentiel lorsqu'on annote des documents, nous a amenés à décrire l'impact de l'enrichissement sur le processus d'annotation.

1.5.3 Troisième partie

La dernière partie de cette thèse est consacrée à l'expérimentation de notre approche ainsi que l'évaluation des résultats.

Afin de tester notre approche, nous avons implémenté un outil nommé RAS (Reference Annotation System) qui regroupe les différentes phases de notre approche. Le chapitre 6 est consacré à la présentation de l'outil RAS.

Le dernier chapitre présente l'expérimentation effectuée pour évaluer l'approche proposée. Le nombre des documents du SEMIDE n'étant pas important, les expérimentations ont été réalisées sur la base Citeseer (articles scientifiques) compte tenu de sa taille et ses caractéristiques.

Première partie

Etat de l'art

2

Annotations des documents

L'annotation des documents est une phase essentielle dans la recherche d'information car elle permet de la rendre exploitable. Ce processus est développé dans ce chapitre.
Deux types d'annotation sont différenciés, l'un effectué par le contenu et l'autre par le contexte des documents (relations inter-documents). Tout au long de ce chapitre, sont décrites les différences et les limites des deux types d'annotation.

Sommaire

2.1 Processus de recherche d'informations 15
2.2 Annotation des documents par le contenu . . 16
2.3 Annotation des documents par le contexte . . 23
2.4 Discussion et Conclusion 33

2.1 Processus de recherche d'informations

Pour faciliter l'accès à l'information ainsi que son exploitation au mieux, on a besoin de décrire cette information. Dans notre contexte, ceci se traduit par des documents annotés par un ensemble de termes représentatifs suivants ou non un vocabulaire contrôlé, et ceci de manière automatique ou bien guidée par des experts à différents degrés. Nous illustrons dans la figure 2.1 le processus de recherche d'informations (RI) et par cela l'étape d'annotation. Le problème central en RI est de localiser les termes que d'autres ont utilisés dans le document recherché [Bla90].

Définition 2.1 *Nous définirons un document comme le support d'une information.*

Le processus de recherche d'informations se compose de deux principales phases :

- la phase d'annotation : qui a pour but de représenter au mieux le contenu du document. C'est l'étape traitée tout au long de notre travail.
- la phase de recherche : qui consiste à restituer à un utilisateur les réponses les plus pertinentes par rapport à sa requête, en utilisant l'annotation des documents.

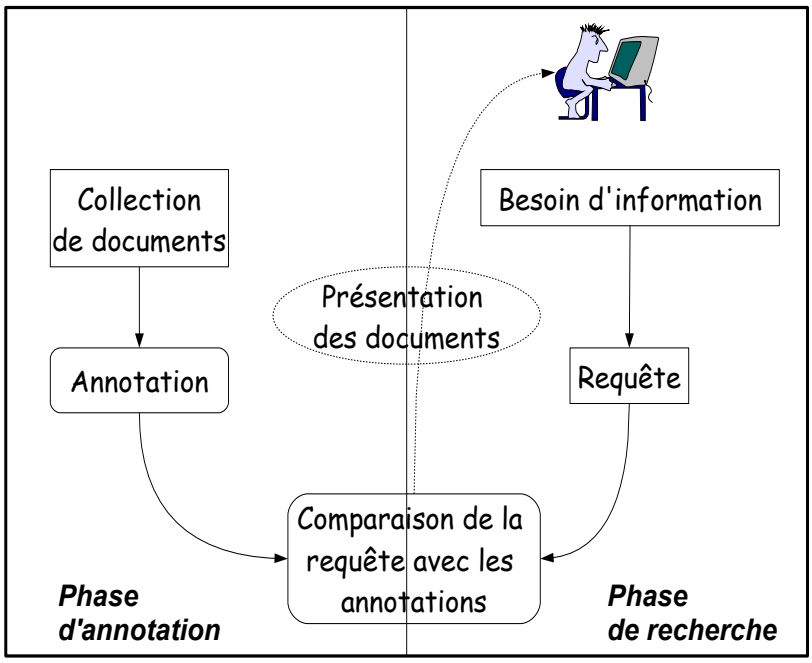

FIG. 2.1 – Les étapes dans le processus de recherche d'informations

Dans ce chapitre, nous présentons succinctement les deux types d'annotations que nous avons retenus : l'*annotation par le contenu*, en utilisant uniquement le contenu du document et l'*annotation par le contexte* qui consiste

à utiliser les relations entre les documents. Nous nous attarderons sur les travaux d'annotation utilisant le contexte de citation en introduisant la notion de bibliométrie ainsi que quelques méthodes connues.

Ces travaux sont proches de notre méthode de propagation d'annotations qui constitue le cœur de notre travail.

2.2 Annotation des documents par le contenu

Le premier type d'annotation, qu'on appelle aussi *annotation statique* ou indexation, utilise uniquement le contenu du document pour le décrire. Le terme statique est ici utilisé dans le sens où les relations que le document peut avoir avec d'autres documents n'influent pas sur son annotation. On présente ici deux types d'indexation : *classique* et *sémantique*, le deuxième type utilisant le sens des mots et les relations entre eux.

2.2.1 Indexation classique

L'indexation d'un texte [Sal71], [Rij79] consiste à repérer dans son contenu certains mots ou expressions particulièrement significatifs (appelés termes d'indexation) dans un contexte donné, et à créer un lien entre ces termes et le texte d'origine. Il existe trois types d'indexation :

1. manuelle : lorsque le document est analysé par un spécialiste du domaine ou un documentaliste ;
2. automatique : lorsque le processus d'indexation est complètement informatisé ;
3. semi-automatique : lorsqu'une première sélection de termes est réalisée automatiquement mais le choix final reste au spécialiste. Les systèmes les plus simples et les plus répandus sont basés sur la sélection de mots-clés dans les textes [EMT92].

L'indexation manuelle assure une bonne correspondance entre les documents et les termes. Cependant, cette méthode demande un travail manuel qui est non seulement très difficile mais très long à réaliser par les indexeurs.

Salton [Sal86] a démontré les inconvénients de ce type d'indexation, par exemple : deux indexeurs peuvent indexer deux documents identiques avec des termes différents et des différences d'indexation peuvent également exister chez la même personne qui indexe à des moments différents.

L'indexation automatique est complètement informatisée et est réalisée en plusieurs étapes : *(i)* l'extraction automatique des mots clés qui correspondent au mieux au contenu informationnel du document, *(ii)* l'élimination des mots vides ou mots fonctionnels (ex : conjonctions de coordination), *(iii)* la lemmatisation pour retrouver la racine des mots, *(iv)* la pondération des mots, pour affecter un poids élevé aux mots les plus importants.

Le résultat du processus d'indexation classique est un index plat, qui est une liste de mots alphabétiques et/ou thématiques sans aucune relation sémantique entre ses termes, les mots sont indépendants et les relations qui peuvent exister entre eux ne sont pas spécifiées.

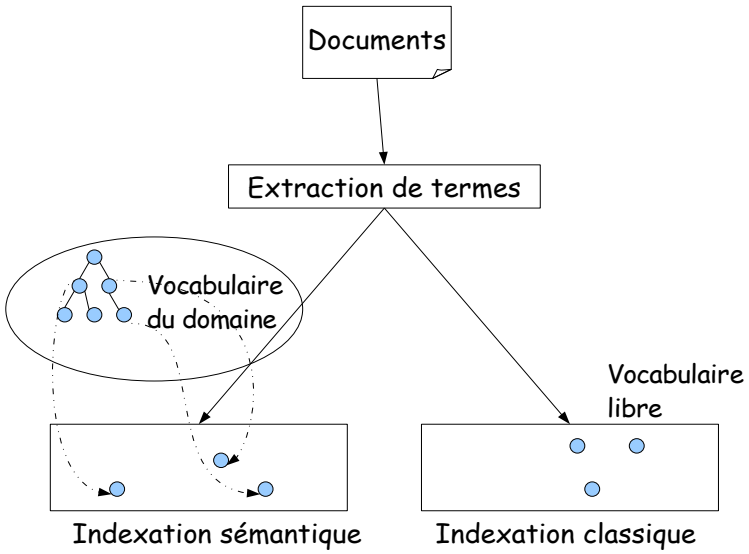

FIG. 2.2 – Comparaison entre l'indexation classique et l'indexation sémantique

2.2.2 Indexation sémantique

L'indexation sémantique (figure 2.2) est une spécialisation de l'indexation classique, elle prend en compte la sémantique des mots au travers des relations entre les termes indexés.

Selon Baziz [Baz05], l'indexation sémantique consiste à utiliser les sens des mots afin d'indexer les documents en désambiguïsant les mots dans le document. Elle se distingue de l'indexation conceptuelle qui est, selon le même auteur, une généralisation de l'indexation sémantique et consiste à identifier des concepts dans les documents, ces concepts véhiculent du sens.

Définition 2.2 *Nous définirons l'indexation sémantique comme l'utilisation d'ontologies de domaine ou de thésaurus afin d'effectuer le processus d'indexation, une ontologie étant un ensemble de concepts reliés par une relation de "spécialisation/généralisation" définissant un domaine donné. Chaque concept est dénoté par un ou plusieurs termes.*

Il existe plusieurs travaux traitant de l'utilisation de l'indexation sémantique. Nous en avons retenus cinq que nous avons estimé pertinents afin de montrer l'utilité d'une telle approche :

- indexation avec une terminologie orientée ontologie ;
- indexation dans un référentiel métier ;
- indexation dans OntoSeek ;
- indexation dans le modèle DocCore ;
- indexation avec une ontologie pour la désambuïguisation.

2.2.2.1 Indexation avec une terminologie orientée ontologie

Parce que l'information sur le Web n'est pas facile à retrouver, compte tenu de l'hétérogénéïté et la mise à jour de cette information, une approche basée sur l'indexation de sites Web avec des mots clés rattachés à des concepts les représentant a été proposée par Desmontils [DJ02] et [DJM02].

Le but est de construire un index associé à une ontologie. Les différentes étapes de ce processus, illustrées dans la figure 2.3, sont les suivantes :

- pour chaque page, un index de termes est construit, un poids est associé à chaque terme en utilisant les marqueurs HTML qui lui sont associés (par exemple un titre est plus important qu'un paragraphe) ;
- un thésaurus est utilisé afin de construire une liste de concepts candidats à partir de l'index de la première étape, ici le thésaurus *Wordnet* [MBF+90] ;
- la représentativité de chaque concept candidat dans le document est déterminée. Ce calcul est fait à partir du poids des termes et de leurs relations avec les autres concepts. Cela permet de choisir le meilleur sens du concept dans la page. Plus un concept est proche des autres concepts et plus il est significatif dans la page ;
- l'ontologie et la représentativité des concepts sont utilisées comme un filtre. Les pages sont associées aux concepts de l'ontologie.

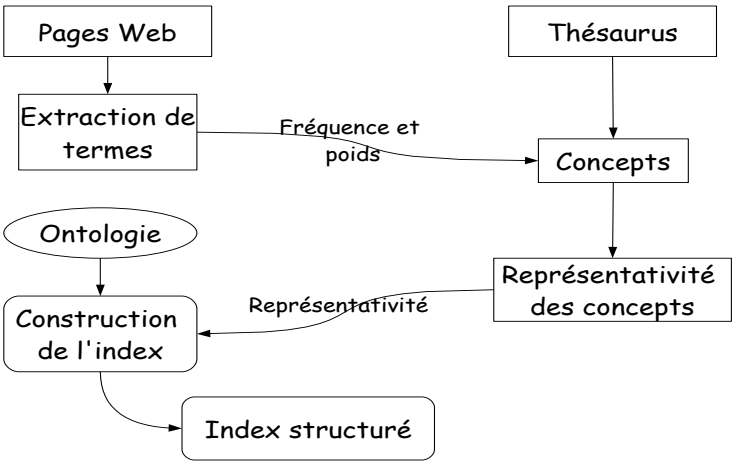

FIG. 2.3 – Processus d'indexation dans [DJ02]

Ce processus est semi-automatique, l'utilisateur à la possiblité d'intervenir sur les différentes étapes). Son but est d'avoir une vue globale du site Web et ainsi faciliter la recherche d'informations.

2.2.2.2 Indexation dans un référentiel métier

L'indexation dans un référentiel métier est motivée par la contrainte que l'indexation d'un document dépend des activités de l'entreprise et non pas des mots clés du document. Ici l'approche d'indexation combine une analyse linguistique du document et une analyse statistique ainsi qu'un traitement sémantique [NFF04b], [NFF04a] :

1. L'*analyse linguistique* consiste à extraire les termes composant les documents. Ceci se fait en utilisant un outil d'extraction terminologique existant.

   Le traitement linguistique représente le document à indexer par un ensemble de termes simples et importants en se basant sur le référentiel métier (l'ontologie du domaine). Le résultat est un fichier balisé qui comprend les lemmes non ambiguës et un fichier lemme des termes lemmatisés avec leurs occurrences. Lemmatiser consiste à réduire les formes fléchies des mots à leur forme canonique.

2. L'*analyse statistique* qui succède l'analyse linguistique vise à determiner l'importance d'un terme dans un document. Ceci est réalisé en utilisant : *(i)* la pondération des termes se basant sur les composantes locales[3], globales[4] et de normalisation[5], *(ii)* la sélection des termes représentatifs et *(iii)* la pondération à base de co-occurrence.

3. La dernière phase consiste à affecter le document au référentiel métier. Afin de répondre au problème de synonymie et de polysémie, l'idée est de représenter le document comme une structure dont les mots sont liés. L'approche consiste à adopter la méthode LSI [DDL+90] (*Latent Semantic Indexing*). Le système part d'une matrice (termes × thèmes), le thème étant le chemin de l'arborescence qui est le référentiel métier. Les colonnes de la matrice représentent la distribution du sens de chaque thème pour le document. Le but est de représenter un document par son contenu et par les thèmes du domaine. Les thèmes sont les métiers prédéfinis dans le référentiel, et un document est représenté par rapport à ces métiers.

La limite de cette méthode observée par les auteurs est le "silence" lorsque : *(i)* il n'y a pas de termes du domaine ou du thème dans le document,*(ii)* il n'y

[3]un terme qui apparaît fréquemment dans un document semble plus important pour le contenu d'un document.

[4]permet d'accorder un poids plus important aux termes qui apparaissent moins fréquemment dans la collection des documents

[5]impose certaines contraintes lors de la pondération des termes pour les documents courts/longs lors de l'indexation.

a pas beaucoup de texte dans le document ou lorsque *(iii)* l'information dans le document n'est pas explicite.

2.2.2.3 Indexation dans OntoSeek

OntoSeek [GMV99] est un système de recherche documentaire pour des documents de type pages "jaunes" en ligne ou "catalogues de produits". Il utilise l'ontologie Sensus : 50000 nœuds concepts issus de la fusion de l'ontologie linguistique (psycholinguistique[6]) *Wordnet* et de l'ontologie Penman [BKMW89]. Le contenu des documents et des requêtes est guidé par un formalisme de graphes conceptuels. Les objectifs de ce système sont :

- l'utilisation des termes "arbitraires" du langage naturel (ontologie linguistique généraliste) qui permettent de décrire le contenu des documents ;
- une flexibilité terminologique pour formuler les requêtes, grâce à un mécanisme d'intersection sémantique entre les requêtes et la description des produits ;
- une assistance à la formulation, la généralisation ou la spécialisation des requêtes ;
- des résultats précis et justes, et une efficacité raisonnable avec des volumes de données importants ;
- une grande portabilité et extensibilité.

Différentes techniques de recherche d'informations ont été testées dans OntoSeek, avec : *(i)* une liste de mots, *(ii)* une liste structurée de mots *(ii)* une liste de sens de mots avec une ontologie linguistique (*Wordnet*) et *(iv)* une liste structurée de sens de mots avec une ontologie linguistique (*Wordnet*). Par ces techniques, OntoSeek a montré pour les pages jaunes l'intérêt d'utiliser une ontologie linguistique couplée avec une description du contenu structurée pour l'amélioration de la recherche d'informations.

2.2.2.4 Indexation dans le modèle DocCore

Ce modèle a pour but de représenter le contenu sémantique des documents en les projetant sur une ontologie linguistique générale. Contrairement aux réseaux sémantiques qui sont une structure de graphe avec des nœuds étiquetés par des constantes de relations, ce modèle a les caractéristiques suivantes [Baz05] :

- il utilise une ontologie pour identifier les concepts du document et calcule les liens de proximité entre eux ;
- les arcs entre les nœuds (concepts) ne sont pas étiquetés mais pondérés par rapport à la similarité sémantique entre les documents (nœuds).

Le modèle construit un *noyau sémantique* pour chaque document avec les concepts et leur proximité. L'idée de cette approche est de représenter l'importance d'un terme dans un document avec : *(i)* sa fréquence d'apparition et

[6] La psycholinguistique est l'étude des facteurs psychologique et neurologique permettant aux êtres humains d'apprendre, d'utiliser et de comprendre les langues.

(ii) la proximité avec les autres termes dans le document. Les étapes de cette approche de manière plus détaillée sont les suivantes :

1. L'*extraction de concepts candidats*. Le but de cette étape est d'extraire des termes du document pouvant représenter les concepts de l'ontologie de deux manières : *(i)* projeter l'ontologie sur le document avec le risque d'omettre des termes dans le document qui n'ont pas la même forme dans l'ontologie ; *(ii)* projeter le document sur l'ontologie, en utilisant les formes de base des concepts représentatifs dans le document s'ils ne correspondent pas aux concepts dans l'ontologie, par exemple si on ne retrouve pas le terme "systèmes" dans l'ontologie, on le met à sa forme canonique "système". Dans cette étape, les termes composés sont aussi extraits afin de réduire l'ambiguïté, ensuite une phase de pondération consiste à affecter un poids aux termes afin de déterminer leur importance. Il faut noter que les termes clés du document peuvent avoir plusieurs sens et de ce fait peuvent être associés à plusieurs concepts ;
2. Le *calcul de similarité entre les concepts* en utilisant un nombre important de relations (synonymie, hyperonymie,...). Le résultat de cette étape est un ensemble de concepts candidats reliés par des valeurs de proximité sémantique ;
3. La *construction du noyau sémantique*, les nœuds d'un réseau sémantique $SN(j)$ sont définis comme suit : $SN(j) = \{C^1_{j_1}, C^2_{j_2}, .., C^m_{j_m}\}$, où $SN(j)$ représente la *jime* configuration des sens des termes extraits du document D. $j_1, j_2, ..., j_m$ sont des index pris entre 1 et le nombre de sens possibles pour les terme $t_1, t_2, ..., t_m$. Cette configuration qui représente le mieux le document est choisie en supposant que le concept candidat est celui qui a le plus de liens avec les autres concepts. Il correspond au score maximum calculé pour ce concept, le score du sens d'un terme T_i étant la somme des valeurs de similarités obtenues avec les autres termes candidats.

Les résultats de ce modèle ont montré que l'indexation conceptuelle combinée à l'indexation classique améliore les résultats du processus de recherche d'information. Baziz [Baz05] met comme perspectives, l'utilisation des noyaux sémantiques afin d'attribuer des thèmes aux documents, en utilisant la mesure de score des termes.

L'indexation des documents en utilisant des ontologies du domaine est aussi pratiquée dans plusieurs spécialités :

- le domaine des biopuces[7] [KDK04] : ce travail consiste à annoter des articles dans le domaine de biopuces en se basant sur une ontologie, améliorant par cela la pertinence des résultats d'une recherche d'informations ;
- le domaine médical [PDB02], consiste à indexer des textes médicaux en se basant sur un thésaurus médical.

[7] permettent de dépister, de détecter, d'identifier des séquences d'ADN grâce aux propriétés des acides nucléiques.

2.2.2.5 Indexation avec une ontologie pour la désambiguïsation

Khan [Kha00] utilise l'ontologie dans le processus d'indexation afin de désambiguïser les termes extraits d'un document par la co-occurrence des mots clés, et la proximité sémantique des concepts dans l'ontologie. Il utilise un algorithme de désambiguïsation des concepts représentant un document et ceci par une liste de mots clés en utilisant les définitions suivantes :

- le score d'un élément : chaque concept C_i est composé d'une liste complémentaire de synonymes $C_i = \{l_1, l_2, ..., l_n\}$. Les mots-clés dans le texte sont appariés avec chaque élément l_j du concept. Le score pour l'élément l_j du concept C_i est le nombre de mots clés l_j s'accordant avec l'annotation du document.

$$Score_element_{ij} = \frac{nombre\ de\ mots\ clés\ de\ l_j\ appariés}{nombre\ de\ mots\ clés\ de\ l_j} \tag{2.1}$$

- le score d'un concept correspondant au plus grand $Score_element_{ij}$;

$$Score_concept_i = max\ Score_element_{ij}\ \ où\ 1 \leq j \leq n \tag{2.2}$$

- le score d'une région dans l'ontologie : est égale à la somme des $Score_concept_i$ séléctionnés ;
- la distance sémantique $SD(C_i, C_j)$ entre les concepts C_i et C_j correspond au plus court chemin entre les deux concepts dans l'ontologie et est égale à 1 s'ils sont directement reliés ;
- le score de propagation : s'il existe une corrélation entre un concept C_i avec un ensemble de concepts $\{C_j, C_{j+1}, C_{j+2}, ...\}$, le score de propagation S_i est le suivant :

$$S_i = Score_concept_i + \sum_{k=j}^{k=n} \frac{Score_concept_k}{SD(C_i, C_k)} \tag{2.3}$$

Quand deux concepts sont corrélés, ils ont un S_i plus grand que les concepts non corrélés.

Toutes ces définitions sont utilisées afin d'affecter un terme dans un texte à un concept dans l'ontologie. Afin d'éliminer des concepts non pertinents, un seuil basé sur le score de propagation est fixé. L'hypothèse de cette approche est qu'un terme dans un document pris isolément de son contexte peut être ambigu.

On constate ici une couche supplémentaire de l'utilisation de l'ontologie, qui ne consiste pas juste à utiliser une ontologie pour seulement se restreindre à un domaine mais utiliser les relations inter-concepts afin d'exploiter les différents sens d'un terme dans un texte et supprimer les ambiguïtés.

Dans la section suivante, nous présentons le deuxième type d'annotation retenu, qui utilise les relations inter-documents.

2.3 Annotation des documents par le contexte

Contrairement au premier type d'annotation, cette annotation ne dépend pas seulement du contenu du document mais aussi du contexte. Nous décrivons dans cette section deux types d'annotation de documents : le premier utilise les liens de citations entre les documents et le deuxième utilise les liens de référencement. Nous définissons dans ces deux parties les concepts de citation et de référencement, ainsi que les travaux effectués dans ce domaine.

2.3.1 Annotation des documents en utilisant le contexte de citation

Définition 2.3 *La scientométrie [Gau98] est la mesure de l'activité de recherche scientifique et technique. La bibliométrie est la composante de la scientométrie qui a pour objet principal l'étude quantitative des publications scientifiques à des fins statistiques.*

Définition 2.4 *La bibliométrie [Lau97] est l'application de méthodes statistiques ou mathématiques sur des ensembles de références bibliographiques. Il s'agit donc d'une mesure utilisée pour aider à la comparaison et à la compréhension d'un ensemble de références bibliographiques.*

La bibliométrie est à l'origine de l'exploitation statistique des publications et ne se limite pas à la stricte élaboration et détermination d'indicateurs synthétiques spécifiques à un sujet traité. Il existe deux approches dans la bibliométrie [PC04] :

- l'approche lexicale où l'analyse se fait sur les mots, cette approche dépend de la langue du texte ;
- l'approche citationniste où l'analyse se fait sur les citations dans le texte (références bibliographiques). Cette approche est détaillée dans la suite, en nous intéressant aux principales utilisations des liens dans les documents. L'**analyse des citations** examine les relations entre les auteurs et les publications, par exemple les relations de co-citations et de co-référence.

Avant de développer les travaux utilisant le contexte de citation, il est utile d'introduire le graphe de citations ainsi que ses propriétés.

Les documents scientifiques peuvent être représentés par un graphe orienté $G = (N, A)$ où un nœud est un article et un arc est un lien de citation. Un exemple de graphe de citation est représenté dans la figure 2.4.

Voici deux spécificités du graphe de citation :

- *dynamique* : le graphe de citation évolue à l'apparition de nouveaux documents. Les nœuds et les liens existants ne changent pas mais un nœud correspondant au nouveau document et des arcs de citations reliant ce document aux différents documents existants qu'il référence, sont ajoutés ;

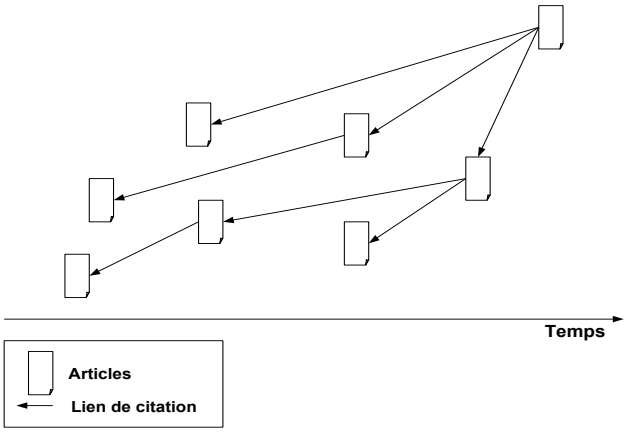

FIG. 2.4 – Le graphe de citation

– *orienté*, et *unidirectionnel* : la relation de citation dépend de la date de publication d'un document, un document cité est apparu forcément après un document citant, et ce dernier ne peut pas être cité par le document cité.

La figure 2.5 illustre les différentes relations de citations entre les documents :

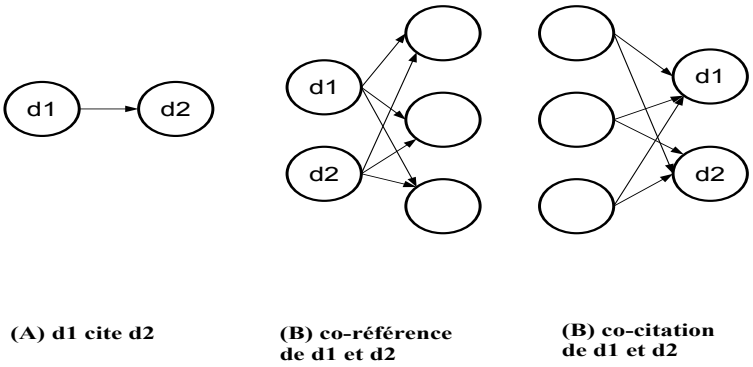

FIG. 2.5 – Les relations entre les documents

– la relation de citation : un document d_1 référence un document d_2. Généralement l'analyse de citation détermine l'impact d'un auteur dans un domaine particulier, en déterminant le nombre de fois où cet auteur a été cité ;
– le couplage bibliographique : les documents sont considérés couplés bibliographiquement quand ils partagent une ou plusieurs références bibliographiques ;

– la relation de co-citations : représente des documents qui sont cités par les mêmes documents.

Les deux méthodes présentées ci-dessous, utilisent la relation de citation non pas pour une annotation, mais pour essayer de trouver une similarité thématique entre les documents.

2.3.1.1 La méthode de couplage bibliographique

Kessler [Kes65] est le premier qui a eu l'idée d'utiliser les citations comme relation entre les documents scientifiques. Kessler [Kes65] a utilisé l'analyse des citations de manière évoluée en élaborant une méthode d'analyse bibliométrique par association bibliographique. Le principe est basé sur l'hypothèse que deux articles qui citent un ou plusieurs documents communs ont une relation significative avec une force d'association traduite par le nombre d'articles en commun. Les articles sont couplés s'ils partagent au moins une référence bibliographique.

Kessler a conçu un tableau qui est construit de la manière suivante :

1. chaque ligne correspond à un document i étudié ;
2. chaque colonne correspond au document j cité dans le document i ;
3. l'ensemble des éléments x_{ij} du tableau sont à 1 si le document i cite le document j et 0 sinon.

Il a comparé les résultats de cette méthode avec ceux d'une analyse de thèmes indexés, et a conclu qu'il existait une forte corrélation entre les groupes formés par les deux méthodes [Kes65]. La figure 2.6 illustre la construction du graphe de couplage à partir du graphe de citation, les nœuds représentent les documents et les arêtes les degrés de couplage. Weinberg [Wei74] pense que le couplage bibliographique fonctionne mieux sur de la littérature où il y a une certaine redondance comme les articles d'état-de-l'art, car les auteurs citent souvent des travaux anciens.

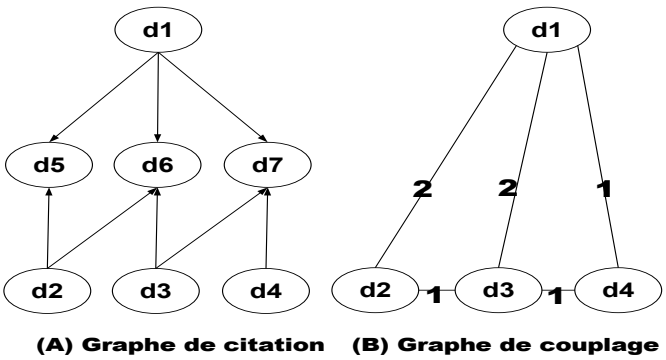

FIG. 2.6 – Graphe de couplage

Cette méthode n'a pas été utilisée par la suite, probablement à cause du volume important des données [Ros96] pour les systèmes informatiques de l'époque. Mais cette méthode a été aussi critiquée par Martyn [Mar64], qui spécifie qu'un article scientifique peut avoir plusieurs idées, et peut être cité pour plusieurs thèmes différents. Prenons l'exemple de la figure 2.6, le document d_7 est cité par d_3 et d_4, on ne peut pas conclure à une proximité en les deux documents car d_7 peut traiter de deux thèmes différents et chaque document le cite pour un thème. Cette critique peut s'avérer vraie quand la force de couplage est faible, mais prenons deux documents qui citent plusieurs mêmes documents, la probabilité pour que tous les documents soient cités pour un thème différent est assez faible.

2.3.1.2 La méthode de co-citations

L'idée de couplage bibliographique a été reprise par Garfield et Small [Sma73], [Gar93] (fondateurs de l'école de pensée) en utilisant la méthode de co-citations. L'utilisation de telles techniques a dû attendre l'arrivée de la micro-informatique. D'autres analyses sont apparues plus tard comme les *co-auteurs* et *co-occurences de mots*.

La méthode de co-citation [Gar93], utilisée en bibliométrie depuis 1973, a pour but de créer à partir d'articles scientifiques d'un même domaine de recherche et en utilisant leurs références bibliographiques, des relations entre ces articles. Cette méthode repose sur l'hypothèse que deux références bibliographiques de dates quelconques, fréquemment citées ensemble, ont une parité thématique. De la même façon que pour le tableau des couplages, la matrice de co-citations est construite de la manière suivante :

1. chaque ligne est l'ensemble des citations i étudiées ;
2. chaque colonne l'ensemble des citations j ;
3. l'ensemble des éléments x_{ij} de la matrice correspond au nombre de documents qui ont cité le document i et le document j en même temps.

La figure 2.7 illustre la construction du graphe de co-citations à partir du graphe de citation.

En 1985, la mesure de co-citations est calculée à partir d'une pondération. Le nombre de citations d'un document est divisée par le nombre de citations présentes dans le document citant [SSG85].

Limites de la méthode de co-citations

De nombreuses limites ont été soulignées sur la méthode de co-citations et plus généralement sur les citations [Ros96] :

- les citations erronées, quand un auteur cite un travail mais en référençant une autre source que l'auteur principal ;
- la nature des citations qui peut être critique. Un auteur peut citer un document afin de critiquer le travail développé dans ce document, par

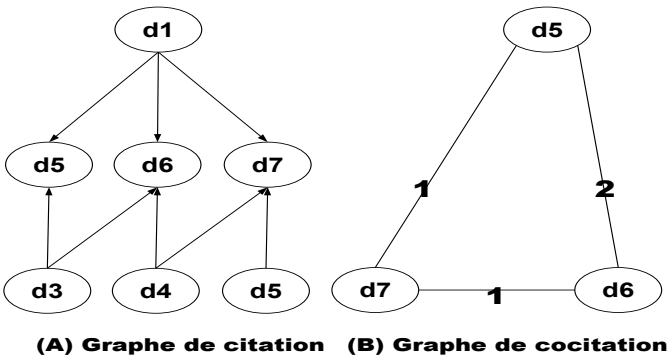

FIG. 2.7 – Graphe de co-citations

exemple pour une erreur méthodologique.
- l'auto-citation, les références faites par un auteur à ses travaux antérieurs ;
- la période entre la citation et la publication. En effet un document qui vient de paraître ne peut pas être cité immédiatement, il faut attendre une longue période avant qu'il ne soit cité ;
- le nombre de citations dépend de la nature de la publication ;
- le choix de l'auteur des citations, il cite plus facilement des personnes de son pays ;
- on cite plus facilement des auteurs de référence (qui sont beaucoup cités ailleurs).

L'analyse des citations montre l'importance que peut avoir une citation dans un document. Lorsque l'auteur cite une référence c'est qu'il estime qu'elle ajoute de l'information. L'utilisation du contexte d'un document (relations avec les autres documents) est très importante pour la description et une compréhension complète d'un document. Cependant, les limites citées plus haut nous poussent à faire un choix dans la sélection des citations utilisées, ainsi qu'à l'approche d'utilisation.

La section suivante traite de l'annotation de documents par les liens de référencement dans le Web.

2.3.2 Annotation des documents en utilisant le contexte de référencement

Définition 2.5 *Un lien de référencement est un lien de citation sur le Web sous la forme d'un hypertexte.*

En parlant de lien de référencement, on se réfère au graphe du Web. Le graphe du Web se présente comme un système hypertexte sous la forme d'un graphe orienté où les nœuds correspondent aux pages, et les arcs aux liens hypertextes.

Avant de décrire les travaux sur l'utilisation des liens de référencement pour l'annotation, nous présentons quelques utilisations des liens dans le Web :

- la classification des pages Web est un exemple de l'utilisation de l'analyse des liens afin de retrouver les pages les plus importantes, les deux algorithmes les plus connus de classement de pages sont l'algorithme *Page Rank* [BP98], [ANTT01], [MTF04] et l'algorithme *Hits*[Kle99]. L'algorithme *Page Rank* est utilisé par le moteur de recherche Google[8] afin de classer les pages du Web. Le principe de cette approche est d'évaluer l'importance d'une page en fonction des pages pointant vers elle. La mesure est basée sur trois hypothèses [Agu02] :
  1. la popularité d'une page est fonction de la popularité des pages qui la citent ;
  2. toutes les citations d'une page n'ont pas la même importance ;
  3. la popularité d'une page ne dépend pas des requêtes.

  L'algorithme *Hits* utilise un moteur de recherche pour identifier les meilleures pages autorité (*authorities*) et les pages index (*hubs*). Les *hubs* sont les pages contenant peu d'informations pertinentes, mais beaucoup d'hyperliens et les *authorities* sont les pages contenant peu de liens, mais beaucoup d'informations pertinentes. L'algorithme calcule les relations de renforcement mutuel (*mutually reinforcing relationship*) entre les *hubs* et *authorities* : les bonnes pages pivots (ou *hubs*) sont celles qui pointent vers de bonnes pages références et les bonnes pages références sont citées par de bonnes pages pivots. Finalement, les pages sont classées par ordre décroissant selon leurs scores d'autorité et d'index.

On peut aussi citer d'autres travaux utilisant les liens comme :

- la portée géographique d'un document [TD04], [BCGM+99],
- la découverte de pages [PK02] où les auteurs utilisent les liens afin de calculer la similarité entre deux pages hypertextes,
- la découverte de communautés dans le Web [GKR98], [KRRT99], [VFD04].

Aprés une brève introduction sur l'utilisation des liens de citations dans le Web, nous présentons l'utilisation des liens pour l'annotation des documents. Les travaux présentés plus bas traitent de la propagation de métadonnées de façon générale sur les documents. Cette partie nous intérèsse plus spécialement car notre approche est basée sur la propagation d'annotations de documents cités.

2.3.2.1 La méthode de propagation de mots clés

Marchiori [Mar98] a été le premier à s'intéresser à la propagation des métadonnées. Son approche permet de propager des mots clés sur des pages Web. Ces mots clés sont pondérés par un cœfficient compris entre 0 et 1.

[8] www.google.com

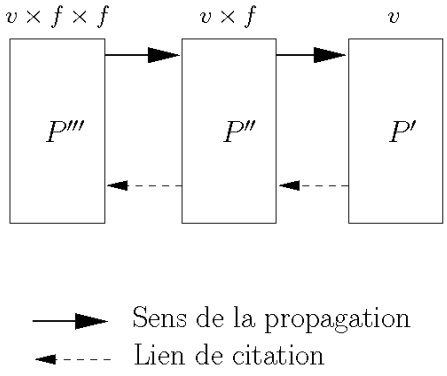

FIG. 2.8 – Propagation de métadonnées (Marchiori)

L'hypothèse de Marchiori est la suivante : si une ressource P' du Web a des métadonnées (mots clés) associées $A.v$, indiquant que le mot clé A a un poids v et s'il existe une ressource P'' dans le Web avec un hyperlien vers P', alors les métadonnées de P' sont propagées à P''. L'idée est que l'information contenue dans P' est accessible par P'', étant donné qu'il existe un lien.

La pertinence de P'' pour le mot clé A n'est pas identique à P'. En effet, l'information pour P' est seulement potentiellement accessible à P'', mais n'est pas directement contenue dans P''. Pour résoudre ce problème, il suffit d'atténuer la valeur v de l'attribut en multipliant par *un facteur d'affaiblissement* f. Ainsi, dans l'exemple ci-dessus, P'' peut avoir sa liste de mots clés avec $A : v \times f$. S'il existe une autre ressource P''' avec un lien vers P'', nous pouvons propager les métadonnées $A : v \times f$ exactement de la même manière ; l'ensemble de métadonnées associé sera alors $A : v \times f \times f$ (figure 2.8).

Appliquer un facteur d'affaiblissement important peut supprimer des mots clés dans les références qui peuvent être importants.

De la même façon, lorsque le facteur d'affaiblissement n'est pas important, la portée de la propagation est rapidement ingérable.

2.3.2.2 La méthode de propagation de métadonnées sur le *World Wide Web*

L'affectation des métadonnées aux pages est une tâche difficile, c'est pour cela que Prime [PC04] s'est intéressée à l'attribution de ces métadonnées en les propageant dans le graphe du Web. Ces métadonnées ne représentent pas des mots clés mais le type d'autorité, le type d'information et le type du site. Tout comme Marchiori, Prime se base sur l'hypothèse que si une page P contient un lien vers une autre page P', alors ces pages partagent des métadonnées communes. Contrairement à Marchiori, l'approche ne s'applique pas sur tout le graphe du Web mais sur un graphe construit avec la méthode de *co-**s**itation* définie par Prime selon le principe de la méthode de co-citations appliquée sur des documents Web. La relation de *co-**s**itation* définie ici est établie entre deux

pages Web P_1 et P_2 si elles sont citées par une autre page sur un site différent de P_1 et P_2. Pour cela, les auteurs considèrent deux étapes principales :

- la construction d'un corpus (un corpus est un ensemble de documents, regroupés dans un but précis) avec la méthode de *co-sitation* afin d'obtenir des classes partageant des propriétés. Ensuite, un travail manuel d'attribution de métadonnées est effectué sur un ensemble de documents ;
- la propagation des métadonnées dans les sous corpus

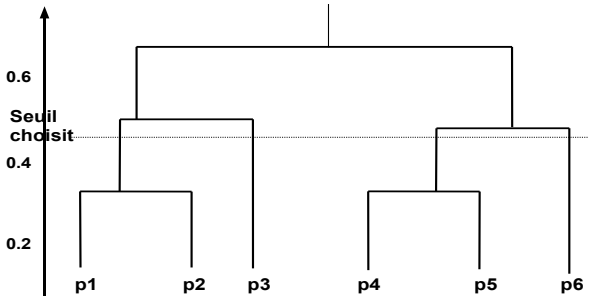

FIG. 2.9 – Dendogramme de la méthode de classification

La construction du corpus se fait en trois étapes : *(i)* le calcul de la matrice de co-sitations, *(ii)* la calcul de la similarité entre les pages, *(iii)* le regroupement en classes. Pour cela une méthode de classification hiérarchique ascendante est utilisée.

Pour la propagation des annotations, deux méthodes sont utilisées. Elles diffèrent dans le choix des pages à indexer manuellement :

- la première méthode consiste à choisir un niveau dans un dendogramme (figure 2.9) construit à l'étape de la classification. Le dendogramme est le résultat d'une classification hiérarchique ascendante. Dans un premier temps, les classes sont formées d'un seul document, ensuite les plus proches sont identifiées et regroupées afin de recalculer la similarité avec les autres classes.
  Le principe de cette méthode est que si les deux pages les plus éloignées (par exemple P_1 et P_6) partagent les mêmes métadonnées, alors les autres pages ont une forte probabilité de partager les mêmes métadonnées, elles sont alors propagées sur les pages P_2 à P_5 ;
- la deuxième méthode repose sur l'hypothèse que l'élément le plus proche de tous les autres (élément central) est le plus représentatif de la classe (par exemple P_4). Cette page est indexée et ses métadonnées sont propagées sur toutes les autres.

Cette approche s'applique sur les trois types de métadonnées "type d'autorité, type d'information et type du site". Une telle approche de propagation ne peut pas s'appliquer sur la métadonnée *mots clés* car cela générerait des

similarités entre les annotations des différents documents car dans les deux méthodes, la propagation se fait d'un ou de deux documents sur un autre ensemble de documents estimés proches. Même avec l'évolution des documents dans le temps, ce type d'approche donnerait comme résultat des documents qui seraient tous annotés avec les mêmes termes.

2.3.2.3 Méthode de propagation de signatures lexicales

Cette approche de Bouklit et Lafourcade [BL06] a pour but de propager des signatures lexicales dans le graphe du Web. Quelques définitions données par les auteurs de ce travail sont nécessaires à sa compréhension :

- une *signature lexicale* $S(p)$ est un ensemble de termes pondérés décrivant une page ;
- la *signature interne* $I(p)$ d'une page P est la signature lexicale que souhaite donner l'auteur à la page ;
- la *signature externe* $E(p)$ d'une page P est la signature lexicale perçue par les auteurs des pages qui pointent la page P ;
- le *contenu d'une page* $C(p)$ est la signature lexicale, en se basant uniquement sur le contenu de la page. Il est calculé en utilisant les techniques d'indexation classique (Section 2.2.1) ;
- l'*équation de propagation avant* consiste à calculer la signature externe d'une page P à partir des précédentes signatures internes des pages qui la pointent ;
- l'*équation de propagation arrière* consiste à calculer la signature interne d'une page P à partir de son contenu et des précédentes signatures externes des pages qu'elle pointe.

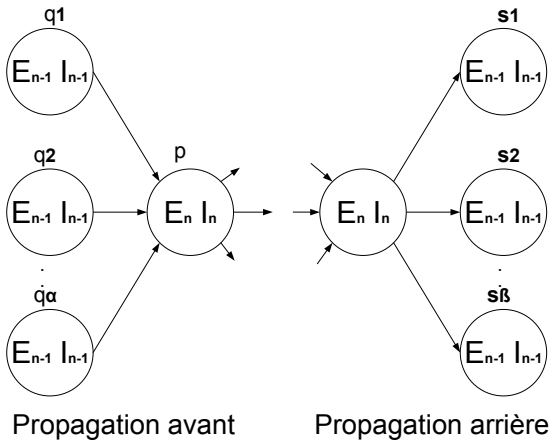

FIG. 2.10 – Propagation avant et arrière

Comme représenté dans la figure 2.10 , $q_1, q_2, .., q_\alpha$ désignent les prédécesseurs de la page P dans le graphe et $s_1, s_2, .., s_\beta$ désignent les successeurs de la page P. Les équations de propagation avant et arrière sont les suivantes :

$$\begin{cases} E_0(p) \leftarrow \emptyset \\ E_n(p) \leftarrow f(I_{n-1}(q_1), ..., I_{n-1}(q_\alpha)) \\ et \\ I_0(p) \leftarrow C(p) \\ I_n(p) \leftarrow g(E_{n-1}(s_1), ..., E_{n-1}(s_\beta), C(p)) \end{cases}$$

L'algorithme de propagation décrit dans [BL06], consiste à appliquer itérativement les deux équations de propagation des signatures lexicales. Ensuite, on combine respectivement les signatures internes et les signatures externes avec le contenu afin de calculer les signatures externes et internes.

Ici, on combine contexte et contenu en supposant qu'on possède le contenu des documents. Ceci est intéressant, mais le problème reste identique car toutes les références sont utilisées, ce qui génère forcément du bruit avec des références qui ne sont pas pertinentes pour le document.

Les auteurs ont posé une limite qui est la combinaison avec le contenu. Ceci ne va pas suffire à éliminer ce qui n'est pas pertinent ou bien cela peut avoir complètement l'effet inverse et éliminer trop de termes provenant des références si on se base trop sur le contenu.

Un autre point important est le domaine du travail et le type de documents. Une telle approche ne peut être appliquée sur le Web, les informations provenant de sources complètement hétérogènes et les références pouvant être des liens sur des publicités. La propagation ne donnerait pas une annotation adéquate. Une solution améliorant le résultat serait de créer une restriction à ce domaine en utilisant une ontologie.

2.3.3 Propagation d'annotations sur des images

L'utilisation des liens afin d'annoter une ressource s'applique également sur les images. L'annotation de ce type de ressource reste indispensable car sans elle, la ressource reste inexploitable. Dans cette section, nous présentons des approches de propagation d'annotations d'images :

2.3.3.1 Méthode de propagation d'annotations d'images

La recherche d'images est une tâche impossible si ces dernières ne sont pas annotées. Ce travail [HL05] consiste à propager les annotations d'images en se basant sur le degré de similarité entre elles, comme suit :

1. utiliser des techniques existantes afin de représenter les images par du texte ;
2. déterminer la similarité entre les textes ;
3. créer un corpus d'images pré-annotées ;
4. propager sur les images qui ne sont pas annotées.

Ce travail est basé sur l'hypothèse que la similarité des images revient à la similarité sémantique d'un contenu.

2.3.3.2 Méthode de propagation d'annotations de photographies

Un autre travail dans le même domaine s'intéresse à l'annotation de photographies [ZHL+04]. Il consiste à propager les annotations de photos d'un album de famille. Partant d'un système qui permet à un utilisateur de labelliser une photo en se basant sur une liste de labels proposée par le système, ce travail est une continuité de ce système. Les auteurs présentent un système qui, contrairement au premier, nécessite de parcourir photo par photo. Il peut sélectionner plusieurs photos et labelliser un individu.

2.4 Discussion et Conclusion

Les méthodes et outils présentés dans ce chapitre montrent l'apport de l'utilisation d'une ontologie ou autre ressource utilisant un vocabulaire contrôlé, ainsi que les relations entre les termes afin d'indexer les documents. Le résultat de l'indexation n'est pas une liste plate de mots qui sont complètement indépendants mais sont reliés entre eux. L'approche générale s'effectue en passant par une analyse linguistique du texte qui peut être complétée par d'autres méthodes.

L'apport de l'utilisation d'une ontologie est à deux niveaux :

- d'un côté, l'indexation des documents en utilisant un vocabulaire contrôlé et spécialisé à un domaine bien défini permet d'éviter d'avoir du bruit (trop de réponses dans le processus de recherche) ;
- d'un autre côté, la recherche d'informations en exploitant les relations entre les termes pour retrouver les documents les plus pertinents.

L'indexation sémantique à partir des ontologies est la méthode la plus pertinente et la plus prometteuse pour pallier aux problèmes de volatilité et d'hétérogénéité des documents, mais n'utiliser que le contenu du document génère un manque qui peut être dû aux raisons suivantes :

- un auteur peut vouloir faire passer une idée dans un document et ne pas utiliser les termes de l'ontologie ;
- un document peut contenir beaucoup d'idées et ne pas avoir beaucoup de contenu. Par exemple un document décrivant les présentations d'une conférence va avoir peu de texte, mais de nombreuses références vers les présentations ;
- enfin, la raison qui nous pousse essentiellement à nous intéresser au contexte des documents est le cas du SEMIDE où le contenu des documents n'est pas toujours disponible mais seulement un ensemble de métadonnées associées.

Par ailleurs, plusieurs points critiques nous ont amenés à nous poser un certain nombre de questions, les plus importantes étant : **est-ce que tout le contexte est pertinent ?** et **faut-il tout propager ?**

Afin d'étudier ces questions, nous avons utilisé l'algorithme de Marchiori sur un petit corpus de notre documentation, afin d'analyser le résultat. Le

choix de cet algorithme est dû à la similarité avec notre approche dans le sens où on utilise les liens de citations afin d'annoter les documents. Ceci revient à une propagation d'annotations.

L'application sur notre corpus de test nous a amenés dans un premier temps à prendre aléatoirement 10 documents de notre base, mettre les poids des mots clés à 1 avec un facteur d'affaiblissement à 0.8. En ne gardant que les mots clés ayant un poids supérieur à 0.5, la propagation se fait sur trois niveaux $1 \times 0.8 \times 0.8 \times 0.8 = 0.512 > 0.5$ comme on peut le voir sur la figure 2.11.

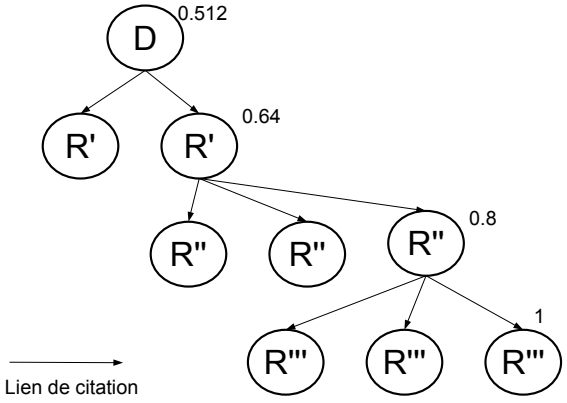

FIG. 2.11 – Propagation selon Marchiori

Le nombre d'annotations propagées a un comportement exponentiel.

De la même façon, les travaux présentés utilisent toutes les références pour créer une annotation. L'idée de prendre toutes les références et toutes les annotations et de propager génère beaucoup de bruit et, avec le temps, un nombre exponentiel d'annotations. L'utilisation du contexte de citation pour l'annotation rejoint l'idée de base de l'*analyse des citations* où, si un auteur cite un document, c'est qu'il estime que ce document est important. Cette hypothèse est valide mais uniquement dans l'absolu. La réalité des bases documentaires est que toutes les références ne sont pas toujours pertinentes, et dans le sous ensemble d'annotations pertinentes, on ne peut pas mettre toutes les annotations au même niveau de pertinence.

Pour conclure, nous avons vu dans ce chapitre le processus de recherche d'informations, en nous intéressant à l'étape d'annotation. Celle-ci consiste à définir un ensemble de mots clés pertinents décrivant le document afin de tirer profit du contenu. Deux types d'annotations ont été retenus : *l'annotation par le contenu* et *l'annotation par le contexte*.

Le premier type consiste à annoter des documents à partir de leurs contenus, cette annotation peut se faire en utilisant ou non un vocabulaire contrôlé. Nous avons vu dans cette partie l'utilité de suivre une ontologie ou un langage

commun afin d'annoter les documents. Les ontologies servent dans le processus d'annotation à plusieurs niveaux :

- l'utilisation d'un vocabulaire contrôlé et partagé par tous les acteurs, ce qui facilite le partage des informations ;
- la désambiguïsation des termes lors de l'indexation, en se basant sur les co-occurrences des termes ;
- l'annotation des documents indépendamment de la langue, une ontologie ayant un ensemble de concepts dénotés par des termes dans plusieurs langues.

Le deuxième type concerne l'annotation par le contexte. Pour cela on a introduit quelques méthodes utilisées dans le domaine de bibliométrie traitant des différents liens de citations entre les documents et les travaux correspondants. Nous nous sommes particulièrement intéressés aux travaux traitant de l'annotation des documents en utilisant des liens de citation/référencement. Effectivement, notre travail rejoint celui de Marchiori [Mar98] dans le sens ou on utilise aussi les liens afin de propager des annotations. L'évaluation de l'approche de Marchiori nous a permis de nous poser les bonnes questions pour une telle annotation.

Les différentes étapes de notre approche avant la phase de propagation décrites dans le chapitre prennent en compte :

- les références pertinentes ;
- les différents thèmes traités dans un document (un document peut être traité par plusieurs thèmes différents, comme décrit par Martyn [Ros96]) ;

Étant dans un contexte technique, spécialisé et multilingue, nous utilisons une ontologie afin d'effectuer l'annotation des documents.

Le chapitre suivant est consacré aux techniques et travaux sur la construction et l'enrichissement d'ontologies.

3

Les ontologies

L'annotation de documents, dans notre approche, nécessite l'utilisation d'un vocabulaire contrôlé partagé par les acteurs du domaine. Ce vocabulaire se traduit par une ontologie qui doit être évolutive en fonction du besoin de la communauté et de l'information existante.
Dans ce chapitre, nous présentons quelques approches de construction et d'enrichissement des ontologies. Le nombre important de travaux dans ce domaine nous a amenés à décrire des méthodologies et outils classiques, pour ensuite retenir quelques approches présentant des similitudes avec notre problématique.

Sommaire

3.1 Introduction 39
3.2 Définitions 39
3.3 Les ontologies 41
3.4 Principes et méthodes pour la construction d'ontologies 44
3.5 Différentes approches pour la construction et l'enrichissement d'ontologies 49
3.6 Discussion et conclusion 56

3.1 Introduction

La rapidité de l'évolution de la masse d'informations dans tous les domaines a généré un besoin d'organisation et de structuration des contenus. Les ontologies servent à la représentation des données échangées dans un domaine particulier afin de faciliter la communication interne au système informatique et externe entre les différents acteurs du domaine. Leur utilisation peut varier de la représentation des données à la recherche d'informations. Notre travail a pour objectif d'enrichir une ontologie existante (celle du SEMIDE) afin de l'utiliser pour l'annotation (représentation des données) et la recherche de documents (recherche d'informations).

Plusieurs travaux se sont intéressés à la construction de ces ontologies dans différents plans : *(i)* extraction de termes représentatifs dans un domaine spécialisé, *(ii)* identification de relations lexicales entre les termes, *(iii)* placement des nouveaux termes dans une ontologie existante. On verra que dans ces travaux, le terme ontologie prend plusieurs sens, que l'on parle de thésaurus, taxonomies ou plus généralement de vocabulaire contrôlé.

Ce chapitre aborde essentiellement les différentes approches que nous avons étudiées, et qui traitent complètement ou même partiellement du problème de construction ou d'enrichissement d'ontologies. La suite de ce chapitre est organisée comme suit :

- dans la section 3.2, sont présentées des définitions générales sur l'utilisation de vocabulaires contrôlés ;
- la section 3.3 traite de l'introduction et de l'usage du terme ontologie dans le monde de l'informatique ;
- les principes et méthodologies générales, ainsi que quelques outils classiques pour la construction d'ontologies sont présentés dans la section 3.4 ;
- dans la section 3.5, les approches de construction et d'enrichissement sont étudiées, puis discutées dans la section 3.6.

3.2 Définitions

Vocabulaire contrôlé

Un vocabulaire contrôlé [the01] est une liste de termes définis généralement par une communauté afin de pouvoir décrire du contenu, et de rechercher l'information. Il est souvent utilisé pour les documents techniques ou plus généralement dans un domaine spécialisé. L'utilisation d'un vocabulaire contrôlé pour la description des documents facilite l'accès à l'information pour les utilisateurs par la rapidité et la délimitation d'un domaine de connaissances. Les résultats d'une recherche sont plus précis et pertinents par rapport à une description libre car un sujet sera décrit avec les mêmes termes.

Un vocabulaire contrôlé est utilisé dans les thésaurus, les ontologies, les

réseaux sémantiques, ainsi que toute autre ressource décrivant un domaine avec des termes préférentiels. L'organisation du vocabulaire est traitée à différents niveaux par exemple (relations entre les termes ou le multilinguisme).

Taxonomie

La structure la plus simple d'un vocabulaire contrôlé est la taxonomie. Il s'agit d'une hiérarchie de termes, organisée généralement avec la relation de spécialisation / généralisation. D'autres relations sont utilisées comme la composition mais dans une taxonomie, un seul type de relation est représenté [Tex05].

Exemples de taxonomies

Relation de divsion
- Monde
  - Afrique
  - Europe
    - France
    - Italie

Cet exemple montre une décomposition hiérarchique ou division du monde en continents et chaque continent en pays.

Thésaurus

Les termes d'un thésaurus servent à représenter ou à annoter des documents. Le thésaurus utilise un vocabulaire contrôlé, structuré et souvent restreint à un domaine particulier. En plus des relations de spécialisation (relation verticale) présentes dans une taxonomie, un thésaurus élargit le contexte d'un terme en ajoutant d'autres relations : terme interdit, terme préféré (relations horizontales). Cet élargissement sert au processus de recherche de documents afin d'augmenter la pertinence avec des résultats correspondant à tout un contexte et non pas à un seul terme.

Un thésaurus est une sorte de dictionnaire[9] hiérarchisé. Les termes normalisés sont reliés par des relations sémantiques. Il est organisé alphabétiquement, et propose généralement les définitions pour les termes.

Exemples de thésaurus

GEMET, thésaurus multilingue de l'environnement[Age99]
- Terme faune
- TG(Terme général) organisme
- TS(Terme spécifique) faune marine

[9]Recueil de mots rangés par ordre alphabétique et suivis de leur définition ou leur traduction dans une autre langue (Larousse 2000).

– TT(Thème) biologie

Réseau sémantique

Le réseau sémantique est un outil qui simule la représentation humaine de la mémoire. C'est un modèle qui montre comment l'information pourrait être représentée en mémoire et comment on pourrait accéder à ces informations [Qui68],[Sch05].

Un réseau sémantique est une représentation de l'ensemble des connaissances qu'un individu se construit pour un domaine spécifique. Les éléments de connaissance sont représentés par des nœuds et sont reliés par des liens associatifs ou sémantiques [HCGS99]. Contrairement au thésaurus qui peut associer un terme à plusieurs contextes, dans un réseau sémantique un concept est défini de façon unique. Chaque concept est représenté par des attributs propres aux nœuds et par des relations qui l'associent aux autres nœuds [JFBR92].

Rastier [Ras04] rapproche les ontologies des réseaux sémantiques, et note que la principale différence réside dans la nouveauté de l'utilisation des ontologies.

> "...La nouveauté réside dans leur échelle sans précédent (par dizaine de milliers de concepts) et dans leur utilisation pour servir de base de connaissances interlangues."

Toutes ces définitions montrent la difficulté de distinguer tous ces concepts, ainsi que la subtile différence qui existe entre eux.

Définition 3.1 *Notre ontologie se rapproche d'un thésaurus par les relations qui la forment, ainsi que des réseaux sémantiques par le souhait de ne pas se limiter à des relations prédéfinies. Le principe est ici l'utilisation d'un vocabulaire contrôlé pour décrire un domaine traitant de ressources hétérogènes.*

Avant de présenter les différents travaux sur la construction et l'enrichissement de l'ontologie, nous donnons quelques définitions du terme ontologie ainsi que son évolution dans le monde informatique.

3.3 Les ontologies

3.3.1 De la philosophie à l'informatique

Il est difficile de donner une définition unique du terme ontologie car il apparaît dans plusieurs domaines. Le terme ontologie est apparu en premier en philosophie, il signifie "être" (du grec *ôn, onton*, participe présent du verbe einai) et discours, étude, sciences (*logos*). L'ontologie est la science ou la théorie de l'être [PMB03].

Allier donne la définition suivante [Lam02] extrait de [All87] :

> "Science de l'être en tant qu'être, c'est-à-dire de l'être en général, de ses diverses espèces, de ses propriétés et de ses relations. Les modes de l'être sur lesquels elle discute sont : le possible, le réel et l'impossible, le potentiel et l'actuel, le contingent et le nécessaire, le déterminé et l'indéterminé, le fini et l'infini, le parfait et l'imparfait, la substance et le mode, l'essence et l'accident."

Le monde de l'informatique s'est approprié plus tard ce terme.

Intelligence artificielle

McCarthy a été le premier dans le milieu de l'intelligence artificielle (IA) à s'intéresser aux ontologies de la philosophie, afin de construire des théories logiques de systèmes d'intelligence artificielle, en affirmant que pour pouvoir construire des systèmes intelligents fondés sur la logique on devait construire une ontologie du monde afin d'énumérer tout ce qui existe [PMB03].

Le terme ontologie a commencé à se répandre au début des années 90, une première définition est donnée par Neches et ses collègues [NFF+91] :

> "An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary."[10]

Gruber a ensuite donné sa définition, c'est celle qui est la plus citée en informatique [Gru93] :

> "An ontology is an explicit specification of a conceptualization."[11].

Web sémantique

Le Web sémantique est une extension du Web actuel où l'information est mieux définie dans le but de permettre aux machines et aux humains de coopérer [LHL01] . Dans le Web sémantique, une ontologie est vue comme un ensemble de connaissances, y compris le vocabulaire et les relations sémantiques, avec quelques règles simples d'inférence et de logiques relatives à des sujets particuliers [GBJJ05].

Une ontologie correspond à un vocabulaire contrôlé et organisé et à la formalisation explicite des relations créées entre les différents termes du vocabulaire. Cette formalisation peut être faite par des langages définis par la communauté du Web sémantique. Une ontologie formelle [Gua99] peut être vue comme la distinction entre les entités du monde réel et les catégories utilisées pour décrire des entités (concepts, propriétés, relations). Une ontologie

[10] Une ontologie définit les termes et les relations comportant le vocabulaire d'un thème aussi bien que les règles afin de combiner les termes et les relations pour définir les extensions pour le vocabulaire.

[11] Une ontologie est une spécification formelle et explicite d'une conceptualisation .

formelle [Dam03] est un développement systématique, formel et axiomatique de la logique de toutes les formes et modes d'existence.

Les ontologies ont pour rôle principal la représentation d'un domaine.

> "An ontology is a shared and common understanding of some domain that can be communicated between people and application systems".[12][Fen00]

Elles sont utilisées dans divers domaines dont les suivants :

1. la description de l'information dans différentes spécialités (biologie, médecine...). Par exemple *The Gene Ontology (GO)*[ABB00]
2. la recherche d'informations ;
3. le commerce électronique afin de permettre, par exemple, la communication entre les fournisseurs et les acheteurs ;
4. les environnements de formation à distance [PMB03].

3.3.2 Les niveaux des ontologies

On distingue différents niveaux d'ontologies selon le but pour lequel elles sont conçues. Van Heijst distingue quatre types d'ontologies [Ass98] :

- les *ontologies du domaine* rassemblent les connaissances dans un domaine particulier ;
- les *ontologies applicatives* sont spécifiques et non réutilisables. Elles ont un domaine de validité restreint et correspondent à l'exécution d'une tâche ;
- les *ontologies génériques* expriment des conceptualisations valables dans différents domaines, elles ne sont pas propres à un seul domaine ;
- les *ontologies de représentations* conceptualisent des primitives des langages de représentations des connaissances.

Guarino donne une autre classification illustrée dans la figure 3.1.

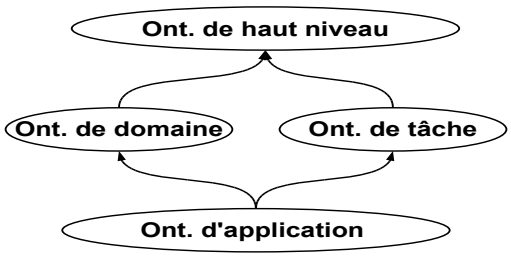

FIG. 3.1 – Les types d'ontologies, extrait de [Gua98]

[12] une compréhension partagée et commune qui permet la communication entre les humains et les systèmes d'application.

Il définit les *ontologies de haut niveau* qui décrivent des concepts généraux comme le temps, une action, et indépendants d'un domaine particulier. Ce type ressemble aux *ontologies génériques* de Van Heijst. Les *ontologies de domaine* et les *ontologies de tâche* spécialisent les termes de l'ontologie de haut niveau. Enfin, *les ontologies d'application* décrivent des concepts qui dépendent de l'ontologie de domaine et de tâche.

3.4 Principes et méthodes pour la construction d'ontologies

3.4.1 Principes méthodologiques

Plusieurs travaux se sont intéressés à l'élaboration de principes de construction d'ontologies. Gomez [PB99] a énuméré un certain nombre de principes à suivre pour l'élaboration d'une ontologie, inspirés par les différents travaux existants :

- *clarté et objectivité* : des définitions objectives des termes doivent être fournies afin de clarifier le sens des termes ;
- *exhaustivité* : une définition exprimée par une condition nécessaire et suffisante est préférable à une définition exprimée seulement par une condition nécessaire ou seulement par une condition suffisante ;
- *cohérence* : une ontologie doit être cohérente afin de formuler des inférences cohérentes avec les définitions ;
- *extensibilité* : l'enrichissement de l'ontologie ne doit pas influer sur les définitions existantes ;
- *interventions ontologiques minimales* : une ontologie doit faire un minimum d'hypothèses sur le monde en phase de modélisation ;
- *distinction ontologique* : les classes de l'ontologie doivent être séparées ;
- *minimisation des distances sémantiques entre les concepts frères* : les concepts frères doivent être proches sémantiquement.

3.4.2 Cycle de vie des ontologies

Le cycle de vie d'une ontologie est composé des étapes suivantes [Für02], [KCG+01] :

1. évaluation des besoins ;
2. construction ;
3. diffusion ;
4. utilisation.

Avant de construire une ontologie, il faut définir son domaine, son utilité et son maintien.

Les trois étapes de la construction sont la conceptualisation, l'"ontologisation" et l'opérationnalisation. La conceptualisation consiste

à identifier dans un corpus les connaissances du domaine en se basant sur les documents ou bien des interviews avec des experts. L'ontologisation, qui peut être l'importation d'autres ontologies, doit suivre les principes décrits dans la section 3.4.1. L'opérationnalisation consiste à rendre l'ontologie compréhensible par la machine.

3.4.3 Méthodologies générales

Des approches issues de l'ingénierie des connaissances pour la construction d'ontologies sont présentées dans cette section. Ces méthodes sont basées essentiellement sur :

- des experts, généralement en utilisant des entretiens comme sources d'informations ;
- des corpus de textes, en utilisant des approches linguistiques.

3.4.3.1 Méthodes basées sur les experts

La méthode KOD : L'anthropologue[13] Claude Vogel a développé en 1988 une approche utilisant des techniques de l'ethnologie et de la linguistique. La méthode *KOD* (Knowledge Oriented Design) consiste à élaborer trois modèles [Ahm05] :

- le *modèle pratique* : consiste à se baser sur les entretiens avec l'expert et en ayant comme résultat : des taxèmes, des actèmes et des schémèmes. Les taxèmes sont tout objet du monde physique, les actèmes sont tout ce qui peut changer d'état et les schémèmes sont les shémas d'interprétation ;
- le *modèle cognitif*, pour structurer et valider les connaissances acquises ;
- le *modèle informatique*, pour mettre en forme les connaissances dans un programme informatique.

KOD utilise des principes linguistiques et terminologiques, les groupes nominaux pour extraire les concepts et les verbes pour extraire les activités. Ce travail peut produire un grand volume de connaissances et est coûteux [AGBS00].

La méthode CommonKADS : CommonKADS [SWdH+94] est une méthode pour la modélisation des systèmes de base de connaissances. Il y a une phase de spécification et une phase de conception. Les différentes étapes de la modélisation conceptuelle des connaissances de CommonKADS sont les suivantes :

- *le modèle organisationnel* : permet d'évaluer les impacts du système sur l'organisation et d'établir la faisabilité du SBC (Système de Base de Connaissances) ;
- *le modèle des tâches* : permet de décrire les tâches et leur répartition ;

[13] L'anthropologie s'intéresse aux pratiques comme aux représentations. Comparative, elle vise à l'inter-compréhension des sociétés et des cultures.

- *le modèle des agents* : décrit les caractéristiques des agents, les agents peuvent être des humains, des systèmes d'informations ou bien d'autres entités capables d'effectuer une tâche ;
- *le modèle de communication* : permet de modéliser, de façon conceptuelle et indépendante de la plate-forme, les communications entres plusieurs agents impliqués dans une tâche ;
- *le modèle d'expertise* : permet la représentation de connaissances de résolution de problèmes dans une application et les connaissances du domaine ;
- *le modèle de formalisation* : donne les spécifications techniques du système en termes d'architecture, de plate-forme d'exécution.

3.4.3.2 Méthodes basées sur le texte

La méthode BCT : Les Bases de Connaissances Terminologiques (BCT) sont des représentations conçues pour structurer le résultat d'analyses terminologiques afin de normer, puis diffuser les usages consensuels de ces termes dans un domaine technique [Seg01]. Cette approche est entièrement linguistique [AGBS00]. La construction d'une BCT consiste à constituer un corpus en fonction du but visé. Les termes sont ensuite extraits à partir de ce corpus avec un extracteur automatique de termes, en s'appuyant sur les critères linguistiques ainsi que sur le nombre d'occurrences. Un réseau de concepts est construit et défini par les termes et leurs relations.

La figure 3.2 illustre les différentes composantes de la BCT. Elles comprennent :

- le terme qui comporte les données linguistiques ;
- le concept qui comporte des données sur le concept associé au terme ;
- le lien terme-concept ;
- le texte pour extraire les liens entre les différents termes dans le corpus.

La méthode Terminae : Terminae vise à faciliter la modélisation d'un domaine à partir de textes, c'est à la fois une méthode et un outil. Terminae [BS99] est un outil d'aide à la construction d'ontologies à partir de textes. L'approche de Terminae se base sur l'analyse linguistique des textes pour élaborer une ontologie. La méthodologie adoptée est la construction de concepts en identifiant des relations lexicales entre les termes dans le texte. Terminae se base sur les principes suivants [BAG03] :

- prendre les textes du domaine comme source de connaissance ;
- enrichir le modèle conceptuel d'une composante linguistique ;
- utiliser des outils de traitement linguistique ;
- construire des ontologies pour un domaine bien défini.

Terminae vise à construire des terminologies, réseaux conceptuels et ontologies. Les différentes étapes pour la construction dans Terminae sont [BS99], [Lam02] :

1. la description des besoins ;

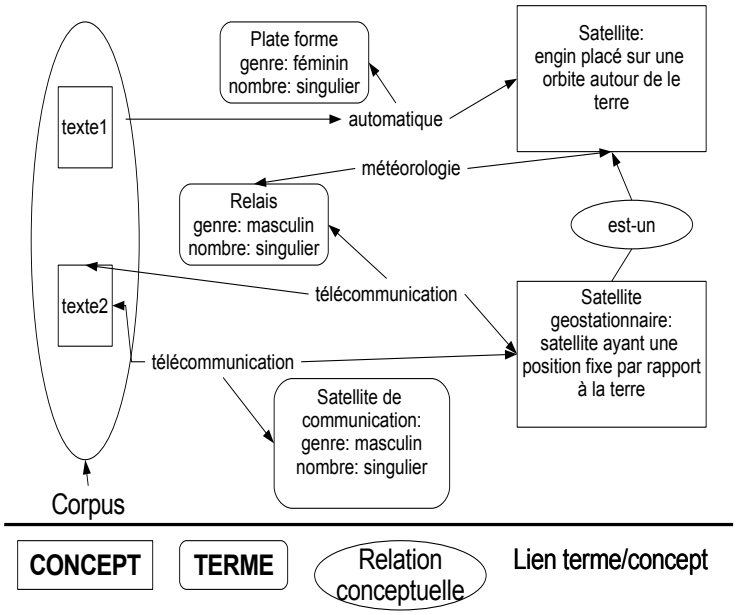

FIG. 3.2 – Les différentes composantes de la BCT, extrait de [CR97]

2. la constitution d'un corpus de documents représentant le domaine. Ici le choix s'est porté sur des documents techniques ;
3. l'identification des termes, et de leurs relations lexicales. Ce travail linguistique se fait en utilisant des outils linguistiques. Le résultat est proposé à un expert pour qu'il choisisse la liste finale de termes ;
4. la définition des concepts et des relations dans un langage formel. La structuration en réseau se fait en tenant compte de l'objectif de l'ontologie tout en se basant sur le texte, et nécessite l'ajout de nouveaux concepts et relations ;
5. la validation du modèle par un expert.

Cette méthode est principalement basée sur l'étude des besoins, la documentation technique..., et pour cela reste générale [AGBS00].

3.4.4 L'analyse de textes

Sont présentés ci-dessous quelques outils d'extraction de termes candidats et de relations.

Le système Lexter : Le système *Lexter* est un outil d'extraction de terminologies. Il effectue une analyse morpho-syntaxique à partir d'un corpus de textes techniques, en français, dans un domaine de spécialité. Le résultat de

cette analyse est un réseau de termes candidats. Les deux analyses de textes morphologique et syntaxique se définissent comme suit [Sch05] :

- la *morphologie* est l'étude de la façon dont sont formés les mots. L'analyse morphologique consiste à identifier les items lexicaux possibles d'une forme fléchie[14] donnée ;
- la *syntaxe* étudie la manière dont les mots se combinent pour former des phrases. L'analyse syntaxique a pour but d'identifier les relations qui existent entre les mots d'une phrase (sujets, verbes, compléments...).

Lexter a été développé par Bourigault [Bou94], il peut être découpé en trois étapes :

- l'extraction des groupes nominaux maximaux ;
- la décomposition des groupes nominaux maximaux ;
- la présentation des résultats sous forme d'un réseau sémantique.

La première étape, dite aussi *découpage*, a pour principe de base de découper le texte en repérant des frontières qui peuvent être de la forme de patrons morpho-syntaxiques, par exemple : verbe, conjonction, etc. Le résultat de ce découpage est un ensemble de groupes nominaux, dits maximaux parce qu'ils constituent des groupes complexes constituant des sous-groupes de termes candidats.

La deuxième étape, *la décomposition*, consiste à décomposer les groupes nominaux résultant de la première étape. L'hypothèse est qu'un terme complexe (groupe nominal) est composé d'une tête et d'une expansion. La décomposition suit deux règles :

- si le groupe nominal a la forme "nom1 adjectif" alors la tête sera "nom1" et l'expansion "adjectif" (par exemple : Réseau régional) ;
- si le groupe nominal a la forme "nom1 de nom2" alors la tête sera "nom1" et l'expansion "nom2". (par exemple : Centre de tourisme).

Lexter gère aussi la suppression d'ambiguïté au cas où le système renvoie plusieurs possibilités de découpage.

La dernière étape, *la structuration*, consiste à représenter les résultats de la deuxième phase par un réseau sémantique constitué d'un ensemble de termes candidats reliés entre eux. Chaque terme est relié à sa tête, à son expansion et à tous les termes dont il est tête ou expansion.

Le système SEEK : Le système SEEK (Système Expert d'Exploration (K) contextuelle) [Jou93] a pour objectif l'aide à l'acquisition et la modélisation d'un domaine de connaissances. Il attribue des valeurs sémantiques aux relations entre des termes extraits manuellement de textes.

Le système détecte des relations "statiques" (identification, incompatibilités, mesures, comparaison, inclusion, appartenance, localisation, partie/tout, possession et attribution) et localise l'expression. Le système utilise une liste de schémas et des règles morphologiques qui permettent la déduction des relations

[14] les mots fléchis comportent un radical, et des désinences qui sont par exemple des indications du nombre et du genre pour les noms.

sémantiques.

Le traitement informatique de SEEK s'appuie sur une exploration contextuelle qui vise à rechercher dans les textes des indices pertinents ou éléments *déclencheurs*. Une fois ces indices trouvés, on recherche dans leur contexte des indices *complémentaires*. A partir de ces deux indices, des relations sémantiques sont construites entre deux termes associés à deux concepts.

3.5 Différentes approches pour la construction et l'enrichissement d'ontologies

Une ontologie sert à décrire un domaine de connaissances. Ce domaine, quel qu'il soit, ne peut pas être figé et représenté d'une manière définitive qui n'évolue pas. Dans notre travail, on s'intéresse à une ontologie existante servant à annoter les documents du domaine et qui évolue en même temps que la masse d'informations.

Dans les travaux que nous allons énumérer, nous nous intéressons à deux façons d'enrichir une ontologie :

- l'enrichissement pendant la phase de construction, nous considérons ici que la phase d'enrichissement est incluse dans ce processus ;
- l'enrichissement d'une ontologie déjà existante.

Dans les deux cas, les approches utilisées dépendent de l'existant et du besoin. Nous avons néanmoins dégagé les plus importantes et qui consistent à la participation d'experts et à l'analyse du corpus du domaine. Ces deux critères, étant les plus importants, peuvent être complétés par d'autres analyses.

3.5.1 Méthodes de construction d'ontologies

3.5.1.1 Méthodes basées sur les utilisateurs

Le système *Adaptiva* [BCW02] est l'implémentation d'une méthode centrée sur les utilisateurs. Il n'est pas nécessaire que les utilisateurs aient une connaissance particulière du domaine. Par contre ces derniers doivent être capables de : *(i)* créer des ontologies, *(ii)* valider des phrases qui expriment une relation entre deux termes et *(iii)* nommer la relation.

Le système doit pouvoir analyser un volume important de textes, identifier les occurrences des termes et les classer, et enfin extraire une relation entre deux termes si elle existe. La méthodologie de la construction de l'ontologie consiste en l'apprentissage de la relation d'hyponymie (*is-a*) par le système et la validation de l'ontologie par l'utilisateur. Elle se fait suivant les étapes suivantes :

1. l'utilisateur choisit le corpus de textes et une ontologie associée à des termes d'un thésaurus ;
2. le thésaurus est utilisé par le système afin de retrouver des exemples qui serviront à la détermination des relations entre les concepts ;

3. l'utilisateur valide ou supprime ces exemples, ainsi des patrons sont déterminés. L'utilisateur peut affiner ces patrons ;
4. l'utilisateur peut enfin fusionner deux concepts s'il estime qu'il s'agit du même concept (exemple synonymie).

La méthode est ici centrée sur l'utilisateur. L'automatisme de l'approche dépend de la première phase et du choix de l'utilisateur de l'ontologie.

3.5.1.2 Méthodes basées sur une analyse linguistique

Sanderson et Croft [SC99] construisent de manière automatique une hiérarchie de termes à partir d'un ensemble de documents en se basant sur les principes suivants :

- les termes sont extraits des documents et reflètent les thèmes couverts par les documents ;
- un terme parent est associé à un concept plus général que le terme fils, le terme parent peut subsumer le terme fils ;
- un terme fils couvre un thème spécifique au terme parent ;
- un terme peut avoir plus d'un parent ;
- les termes ambigus ont des entrées différentes dans la hiérarchie et ceci par rapport au sens du terme dans le document. Il existe une entrée par sens de terme.

La propriété de transitivité où un terme subsume tous ses descendants n'est pas obligatoire, les auteurs donnent un exemple dans [Woo97]. Sanderson et Croft se basent sur la relation de subsomption, avec l'idée qu'un parent subsume les thèmes de ses fils. La relation de subsomption (qui désigne une relation hiérarchique entre deux concepts) se définit ici comme suit :

$$P(x|y) \geq 0.8,\ P(x|y) < 1 \tag{3.1}$$

La règle 3.1 signifie que si x et y apparaissent dans plus de 80% des cas ensemble et que y est plus fréquent que x, alors y est parent de x. L'approche de Sanderson et Croft se base sur les co-occurrences des termes ; x subsume y si les documents où y apparaît est un sous ensemble des documents où x apparaît.

Les termes pour la construction sont sélectionnés en se basant sur les requêtes. Les auteurs font une comparaison entre la fréquence d'un terme dans les documents répondant à la requête x_r et l'occurrence du terme dans toute la collection x_c. Les termes sélectionnés répondent à la condition suivante :

$$x_r/x_c \geq 0.1 \tag{3.2}$$

Cette valeur a été choisie empiriquement afin de sélectionner les termes candidats. Seuls les termes apparaissant dans des parties importantes des documents (par exemple les parties similaires à la requête) sont pris en considération.

Cette approche a été étendue par Hermine [NFG04] pour les thèmes représentés par des mots clés. Ici, les relations entre thèmes sont extraites en se basant sur la hiérarchie de concepts. Hermine [Fot04] présente des approches qui permettent de calculer les probabilités conditionnelles d'un thème sachant un autre sans passer par une décomposition en mots. Cette approche tente de pallier la faiblesse de celle de Sanderson et Croft qui, d'après les expérimentations de l'auteur, ne s'appliquerait que sur un corpus homogène et avec une forte répétition des termes.

L'auteur de ce travail [NFGL03] décrit l'algorithme de génération de hiérarchies de concepts et de documents. La hiérarchie entre documents comme celle entre les concepts, se traduit par un document qui traite des thèmes spécifiques d'un autre document.

Les différentes étapes de cet algorithme sont les suivantes :

- l'extraction des thèmes, ici la méthode de Salton est utilisée afin de décomposer un document en thèmes pour faire un regroupement entre les thèmes et ne garder qu'un petit ensemble de concepts ;
- la construction des hiérarchies de concepts en se basant sur l'approche de Sanderson et Croft pour créer ensuite une hiérarchie de concepts et ainsi rattacher les documents aux différents concepts ;
- la génération des relations de généralisation/spécialisation entre les documents en se basant sur la hiérarchie des concepts.

Ferret [FFHS02] aborde la construction de hiérarchies à partir d'un corpus en se basant sur le contexte des termes. Le système trouve des relations sémantiques entre les termes et plus particulièrement les relations d'hyperonymie afin de construire une hiérarchie.

L'approche se base sur le sens d'un terme qui se traduit par un ensemble de contextes associés au terme dans le corpus, le paragraphe étant le délimiteur du contexte ou bien la délimitation se fait en utilisant un outil.

Chaque terme est représenté par un *contexte sémantique* où un terme t_1 est parent d'un terme t_2 si le contexte sémantique de t_2 est inclus dans celui de t_1.

Les différentes étapes de la construction de la hiérarchie de termes sont les suivantes :

1. extraction et sélection de termes : cette étape consiste en l'extraction des termes du corpus et la sélection des plus représentatifs en se basant sur leurs rôles dans les phrases et leurs occurrences ;
2. construction du contexte sémantique en sélectionnant des ensembles de termes qui co-occurent dans les paragraphes et en associant quelques paramètres comme le nombre d'occurrences dans le paragraphe ;
3. construction de la hiérarchie de manière itérative.

Cette approche construit une hiérarchie en se basant sur un contexte sémantique d'un terme qui se traduit par les termes associés dans un paragraphe qui lui sont associés. D'un côté, cette approche ne construit pas une relation directe entre deux termes et d'un autre côté, le contexte sémantique peut être

différent d'un document à un autre et il est rare de retrouver la même inclusion de contextes.

3.5.1.3 Méthodes basées sur l'analyse formelle des concepts (AFC)

L'Analyse Formelle de Concepts est une formalisation mathématique de l'analyse des données qui utilise la structure des treillis pour la représentation d'une relation binaire entre deux ensembles. L'utilisation de cette structure pour la construction de hiérarchies a débuté dans les années 80 [GW97].

La hiérarchisation de concepts par Cimiano [CST03] basée sur l'analyse formelle des concepts (AFC) vise à créer une taxonomie de manière automatique qui peut être appliquée sur des domaines différents. L'idée est d'extraire automatiquement un ordre partiel entre les concepts. Les connaissances sont représentées par une matrice objets-verbes (par exemple un appartement peut être acheté). Ensuite ces informations sont représentées par un treillis, où l'on retrouve un ordre partiel entre les nœuds verbes-concepts. La hiérarchie entre les concepts est extraite directement du treillis.

L'acquisition de ces relations est faite en utilisant le parseur LoPar[15]. LoPar extrait les concepts et les verbes ainsi que les relations qui les relient. La sélection des termes est basée sur un calcul de probabilité. Cette méthode se rapproche de celle de Ferret [FFHS02], dans la mesure où la construction de la hiérarchie se base sur le contexte du concept et les termes peuvent être regroupés selon les verbes qu'ils partagent.

Latiri [LMB05] présente une approche utilisant aussi l'AFC afin de construire une ontologie. Ceci se fait en passant des relations du treillis aux relations ontolologiques. La première phase de cette méthode consiste en un pré-traitement linguistique du corpus textuel pour une représentation ordonnée des concepts par un treillis de Galois. Des relations ontologiques sont dérivées à partir des relations du treillis (relation père/fils, relation d'héritage) afin de construire l'ontologie de termes.

3.5.2 Méthodes d'enrichissement d'ontologies

3.5.2.1 Méthodes basées sur une analyse linguistique

L'approche de Faatz et Steinmetz [FS02] s'applique sur une ontologie déjà existante afin de l'enrichir en se focalisant uniquement sur la relation généralisation/spécialisation, et en extrayant de nouveaux concepts automatiquement. Pour cela les étapes sont les suivantes :

- utilisation d'un corpus spécialisé à partir du Web ;
- extraction des concepts candidats ;
- calcul de la similarité de ces concepts avec les concepts existant dans l'ontologie.

[15] Analyseur récursif gauche pour des grammaires hors-contexte probabilistes. http ://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LoPar-en.html

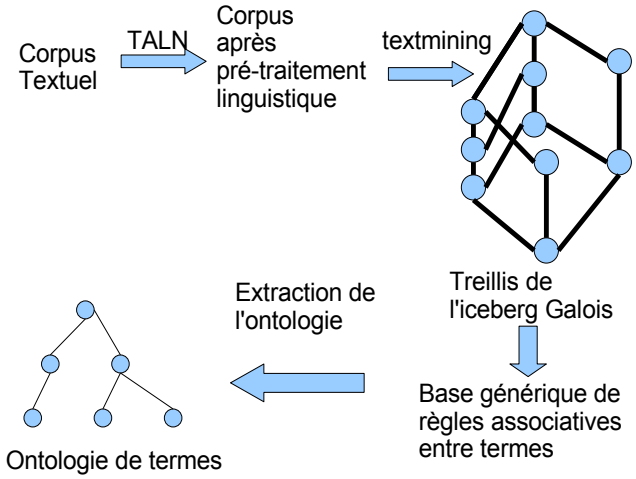

FIG. 3.3 – Méthode de construction de Latiri [LMB05]

Les auteurs posent le problème de l'enrichissement d'ontologie comme un problème d'optimisation de paramètres. Ces paramètres indiquent combien un concept co-occurre avec d'autres concepts candidats par rapport à des règles prédéfinies dans une collection de textes. L'idée est de comparer les mots qui co-occurent dans le corpus avec l'hypothèse qu'il existe une mesure calculant la distance sémantique entre deux concepts dans l'ontologie.

En plus de définir le contexte d'un terme afin de le relier aux concepts de l'ontologie, l'analyse linguistique consiste à extraire des règles pour la construction de relations. Amardeilh [ALM05b] enrichit une ontologie à partir de documents textuels en s'appuyant sur l'analyseur linguistique Insight Discoverer Extractor (IDE) [ALM05a]. L'analyseur produit en sortie un arbre conceptuel étiqueté où chaque nœud porte une étiquette sémantique attribuée à l'unité textuelle extraite en fonction du domaine traité. L'idée de cette approche est d'intégrer les sorties de IDE avec les concepts ontologiques de l'outil Intelligent topic Manager ITM[16]

Ceci se fait en plusieurs étapes :

- parcours de l'arbre résultant de l'extraction linguistique ;
- définition de règles d'acquisition entre les étiquettes linguistiques et les concepts de l'ontologie ;
- déclenchement des règles sur les textes.

L'application de cette approche a été faite sur un petit corpus de compte rendus de décisions de cours de cassation. Un dixième de ce corpus a servi à la définition des règles. L'enrichissement d'ontologie ici peut utiliser n'importe

[16]Outil de la société Mondeca, c'est une plateforme logicielle pour la gestion des connaissances et l'exploitation d'ontologies.

quel outil linguistique, mais la définition des règles suppose que tous les documents analysés aient la même forme afin d'utiliser le minimum de documents pour les définir. Or, à part des domaines très spécifiques et réduits, nous ne pouvons pas nous baser sur la structure des documents afin d'enrichir une ontologie du domaine.

L'enrichissement se fait aussi par des instances, en utilisant l'ontologie existante dans le processus d'annotation. Ainsi Valarakos [VSkGP03] s'intéresse à l'enrichissement d'ontologie avec de nouvelles instances dans la relation de synonymie. Son approche consiste en 3 étapes :

- annotation du corpus du domaine en utilisant son ontologie ;
- utilisation du corpus annoté afin de former des chaînes de Markov cachées [AMS04] pour extraire de nouvelles instances [VSkP03] ;
- extraction de nouvelles instances du corpus ;
- validation des nouvelles instances par les experts du domaine qui les ajoutent manuellement dans l'ontologie.

L'idée de cette approche est de se baser sur la mise en correspondance des mots. Cela revient à détecter les instances identiques qui ont une orthographe différente mais représentent les mêmes concepts en comparant les caractères de ces instances. Les instances qui existent déjà dans l'ontologie globale sont classées par les concepts auxquels elles sont associées. Le codage de Huffman[17] est utilisé pour coder les classes d'instances. Des classes ayant une seule instance sont créées lorsque l'instance n'est pas associée à un concept.

Une autre approche d'enrichissement consiste à ne s'intéresser qu'au contexte d'un concept. Harabagiu [HMM99] a noté que les réseaux de type Wordnet présentaient un manque, dans la mesure où il existe une absence de relation sémantique (appartenance de mots à un même thème). Afin de combler ce manque, Harabagiu [HM02] propose une approche afin d'extraire des relations thématiques de Wordnet en se basant sur les définitions associées aux concepts.
Dans [AAMH01], une méthode d'enrichissement d'ontologie est proposée, en construisant des signatures thématiques pour chaque concept dans Wordnet. Le travail consiste à construire, pour chaque concept, la liste de mots qui lui sont relatifs. Ceci est réalisé par Agirre [AAMH00] selon les étapes suivantes :

- extraction du Web de tous les documents relatifs aux concepts de l'ontologie, en utilisant le moteur de recherche altavista[18]. Les auteurs se basent sur les synonymes et les définitions ;
- extraction des mots en fonction de leur fréquence ;
- filtrage des signatures en appliquant quelques règles comme par exemple prendre un seul document par site Web.

[17] Le codage de Huffman est un algorithme de compression, il permet de coder les octets revenant le plus fréquemment avec une séquence de bits beaucoup plus courte que d'ordinaire.
[18] http ://fr.altavista.com/

Cette approche permet à un utilisateur de manipuler tout le contexte d'un concept, ce qui peut être utile à la recherche d'informations. Par contre cette méthode ne donne pas une relation explicite entre deux concepts (par exemple hiérarchie). Contrairement à cette méthode, notre approche essaie d'enrichir l'ontologie en utilisant les relations existantes (généralisation/spécialisation).

3.5.2.2 Méthodes basées sur une analyse linguistique et des ressources externes

Le manque de connaissances dans une ontologie se ressent généralement lors de la phase d'indexation, lorsqu'on souhaite décrire un document avec par exemple un terme qui n'existe pas dans l'ontologie.

Simon et al [SDJ03] proposent une approche se basant sur un thésaurus et sur des techniques du *Text Mining* afin d'enrichir l'ontologie de manière semi-automatique en laissant l'utilisateur sélectionner tout ou partie des concepts proposés. Ce travail vise à améliorer le processus d'indexation, car les documents sont indexés avec des concepts d'une ontologie afin d'améliorer le processus de recherche.

Se baser uniquement sur les concepts d'une ontologie afin de représenter les documents, peut de ne pas être suffisant car on risquerait d'omettre des concepts importants mais absents de l'ontologie, d'où la nécessité de faire évoluer cette ontologie en l'enrichissant avec de nouveaux concepts.

Cette approche retient l'enrichissement par spécialisation. Elle est réalisée en 3 étapes :

- l'utilisation d'un thésaurus pour l'enrichissement de l'ontologie lors de l'indexation ;
- la gestion du domaine de l'ontologie ;
- la supervision et le contrôle par l'expert.

Notons que l'étape d'indexation en se basant sur les documents génère des concepts candidats qui ne sont pas présents dans l'ontologie. Le travail consiste à ajouter ces concepts en utilisant un thésaurus en appliquant des règles afin de limiter l'ajout des concepts et pouvoir réorganiser l'ontologie si besoin. Le thésaurus utilisé est *Wordnet*.

La deuxième étape consiste ici à contrôler l'enrichissement de l'ontologie en ne conservant que les concepts qui appartiennent au domaine de spécialité et en utilisant la représentativité des concepts dans les pages.

Les expérimentations de ce travail ont montré que la deuxième étape d'élagage de concepts est essentielle, la première phase qui vise juste à enrichir l'ontologie, donnant un nombre important de nouveaux concepts. Ce travail est basé sur l'utilisation d'une hiérarchie de concepts afin d'enrichir une autre hiérarchie. Le point important dans cette approche est le contrôle de l'ajout des concepts en veillant à représenter toujours un domaine de spécialité.

D'autres ressources utilisant un vocabulaire contrôlé peuvent être utilisées pour l'enrichissement d'une ontologie. Parekh [PGF04] décrit comment enrichir

une ontologie en se basant sur des textes, glossaires et dictionnaires du domaine afin de trouver des groupes de concepts/termes reliés entre eux.

Une première ontologie est construite par un expert manuellement, incluant les termes représentatifs du domaine en collectant dictionnaires/glossaires et textes spécifiques au domaine (figure 3.4).

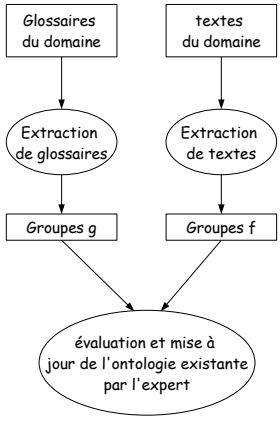

FIG. 3.4 – Enrichissement d'ontologie d'après [PGF04]

Cette approche est une aide à l'expert, car malgré l'extraction automatique depuis des dictionnaires, l'expert doit enrichir l'ontologie manuellement. Aussi trouver des textes représentant le domaine reste possible, mais il est moins réaliste de trouver des dictionnaires ou glossaires spécifiques à un domaine particulier.

Pompidor [PSH03] utilise des dictionnaires afin d'enrichir une ontologie dans le but de construire un cours. Il décrit une approche pour la génération de cours d'un enseignant, de la manière suivante :

- analyser des définitions de plusieurs dictionnaires pour l'extraction de patrons syntaxiques ;
- construire manuellement une ontologie en commençant par un ensemble minimal de termes, cela est fait pour créer son cours ;
- interroger les moteurs de recherche à partir des termes de l'ontologie et analyser les pages réponses pour trouver de nouveaux concepts ;
- compléter l'ontologie avec les nouveaux concepts.

Le brouillon de cours est généré lorsqu'il n'y a plus d'insertion de nouveaux concepts.

3.6 Discussion et conclusion

Plusieurs chercheurs se sont intéressés au problème de représentation d'un domaine par une ontologie, et cela en développant des approches et des outils

pour sa construction et son enrichissement. Compte tenu du nombre important de ces travaux, nous n'avons pas pu tous les énumérer, tentant de décrire dans ce chapitre les travaux qui se rapprochent le plus de nos besoins. Les travaux de construction et d'enrichissement se rejoignent dans le sens où on ne peut pas construire sans enrichir, d'où notre choix de présenter les deux approches. L'élaboration d'une ontologie dans les approches présentées débute par une étude de besoins, en se basant essentiellement sur l'analyse de corpus et/ou analyses d'experts du domaine. Cette construction se fait en identifiant les termes du domaine ainsi que les relations entre eux.

La construction d'ontologie est généralement faite en se basant sur les textes. Effectivement, la première étape pour définir un vocabulaire est d'identifier les concepts candidats. Une analyse syntaxique est utilisée ensuite pour extraire les relations entre les termes, supprimer les ambiguïtés s'il y en a. Cette analyse est généralement complétée par d'autres étapes, comme par exemple la validation par un expert.

L'analyse des textes pour la construction d'une ontologie passe par une phase déterminante qui est la constitution d'un corpus. Or cette tâche n'est pas facile dans le sens où on ne retrouve pas forcément des textes très représentatifs pour n'extraire que les termes importants du domaine. Cette analyse génère généralement trop de résultats (bruit) ou pas assez (silence), ce qui nécessite un travail fastidieux et complémentaire par l'expert.

Évidemment, il n'est pas réaliste de penser qu'on peut se passer du travail humain, car il est impossible de créer une ontologie de manière complètement automatique ou du moins jusqu'à présent. Bien que la méthodologie dépende principalement du besoin, on a déjà pu tirer les constats suivants :

1. le traitement linguistique est essentiel, même s'il ne représente qu'une petite étape dans le processus ;
2. le choix d'un corpus représentatif est une tâche difficile qui ne dépendra que d'un petit nombre d'experts, si ce n'est d'un seul ;
3. enfin, l'ontologie concerne un domaine et par ce fait les acteurs du domaine. Une collaboration de ces derniers pourrait être une approche intéressante.

L'utilisation de techniques linguistiques est intéressante mais le travail de constitution de corpus représentatif n'est pas évident comme nous l'avons déjà expliqué. Une approche intéressante serait d'utiliser de tels outils sur des termes proposés par des experts ou sur des utilisateurs confirmés. Se baser sur des ressources telles que des ontologies existantes revient en fait à reprendre une partie de ces dernières pour construire une ontologie. Cela est possible dans le cas ou nous nous trouvons dans un domaine assez général et où des taxonomies plus spécialisées peuvent être utilisées pour enrichir l'ontologie construite. Dans notre problématique, nous nous situons dans le cas d'une ontologie existante mais insuffisante et où les manques

peuvent être mis en évidence par des experts ou par des utilisateurs insatisfaits.

En conclusion, l'étude des techniques de gestion des ontologies est motivée par le temps que prendrait un enrichissement manuel. Dans ce chapitre, nous nous sommes intéressés aux méthodes de construction et d'enrichissement des ontologies. Nous avons :

- situé notre définition de l'ontologie par rapport aux autres vocabulaires contrôlés ;
- présenté les méthodologies et les approches classiques dans le domaine de la construction d'ontologies ;
- présenté des approches de construction et d'enrichissement d'ontologies. Ces deux approches sont positionnées au même niveau, le travail d'enrichissement étant inclus dans celui de la construction, la distinction entre les deux n'ayant pas lieu d'être.

L'enrichissement est souvent fait sur une ontologie construite en se basant sur un corpus du domaine et en s'aidant parfois d'autres taxonomies ou bien l'ajout se fait par un documentaliste. La construction d'ontologies est une tâche difficile et demande la validation d'un expert. Cette validation se fait à deux niveaux, soit au début du processus, soit pour valider le résultat du traitement automatique.

La troisième partie de ce document décrit le processus d'annotation de documents complété par l'enrichissement de l'ontologie utilisée.

Deuxième partie

Approche retenue : Annotation de documents basée sur une ontologie et enrichissement semi-automatique de l'ontologie

4

Propagation d'annotations entre les documents

Dans ce chapitre, nous proposons une approche pour l'annotation semi-automatique de ressources selon les liens de référencement. Cette approche permet d'annoter une ressource sans connaissance préalable de son contenu selon un regroupement thématique des références construit à partir d'un classifieur flou non-supervisé.

Sommaire

4.1 Introduction 63
4.2 Les étapes de l'annotation 64
4.3 Regroupement thématique des documents . . 66
4.3.1 Construction du graphe de co-citations 67
4.3.2 Calcul de la similarité thématique entre les documents 69
4.3.3 Choix de l'algorithme pour le regroupement . . 72
4.3.4 Regroupement avec fuzzy C-means 77
4.4 Importation et ordonnancement des annotations 79
4.5 Conclusion 82

4.1 Introduction

L'annotation d'un document consiste en un ensemble de termes clés issus ou non d'un vocabulaire contrôlé. Dans notre contexte ces termes sont reliés par des relations sémantiques appartenant à l'ontologie du domaine. L'utilisation d'annotations de documents permet de décrire et d'utiliser au mieux les ressources. Une ressource non annotée peut demeurer inexploitable et impossible à retrouver, cette annotation est une phase importante dans la recherche d'informations (RI), car la phase d'interrogation se base essentiellement sur la description des documents pour les retrouver. En effet, dans le processus de RI, dans la phase d'interrogation, la requête de l'utilisateur est comparée à l'annotation du document afin de retrouver les résultats correspondants.

L'annotation des documents est une tâche quasiment impossible à faire manuellement. D'un côté, ce travail demande une quantité considérable de ressources "humaines" et de temps. D'un autre coté nous avons illustré dans le chapitre 2 les inconvénients d'un traitement manuel, qui peut être différent d'un humain à un autre, une seule personne pouvant proposer deux annotations différentes pour un même document. Ces inconvénients ont été soulignés pour l'indexation. Notons que l'annotation diffère de l'indexation dans le sens ou les termes issus de l'indexation (index) apparaissent dans le document, contrairement à l'annotation ou les termes peuvent être différents.

A partir de ces constatations et du besoin des utilisateurs, nous avons développé une méthode d'annotation semi-automatique, l'expert du domaine restant décideur de la fiabilité de l'annotation. Parti du constat qu'un document technique référence d'autres documents, nous présentons dans ce chapitre une méthode d'annotation qui consiste à propager les annotations en utilisant les liens de citation entre les documents.

Le premier à avoir utilisé les liens de citation dans le Web afin d'annoter des pages Web est Marchiori [Mar98]. Son approche consiste à propager les mots clés des documents référencés dans les pages citantes. Prime [PC04] quant à elle s'est intéressée à propager un autre type de métadonnées (type d'autorité, type d'information et type de site).

Comme *Marchiori* [Mar98], nous utilisons les liens de citation afin d'annoter des documents. Cette annotation est un ensemble de termes appartenant à l'ontologie du domaine. Cependant, comme nous l'avons vu dans l'état de l'art, les citations ne peuvent pas être utilisées de la même manière, et considérées comme ayant le même degré de pertinence dans le document. En effet, les citations dans le contenu d'un document n'ont pas toutes la même importance pour l'auteur. Prenons par exemple un document scientifique, un auteur cite plusieurs documents dans l'état de l'art mais donnera plus d'importances aux travaux qui sont proches des siens. Dans ce cas ces références devront être considérées comme plus pertinentes que les autres.

Notre approche est basée sur le contexte de citation, le contenu n'étant pas utilisé pour annoter un document. Dans un système tel que le SEMIDE, les documents ne sont pas mis à disposition, et cela pour des raisons de confi-

dentialité ou tout simplement parce que les fournisseurs souhaitent que les utilisateurs récupèrent le contenu auprès d'eux après avoir repéré l'information souhaitée. Néanmoins, afin de retrouver cette information, une phase de description (annotation) est obligatoire. Les documents sont fournis sous forme de métadonnées contenant les références, le titre et éventuellement quelques mots clés.

Nous proposons dans cette thèse une approche d'annotation basée uniquement sur le contexte du document. L'approche permet d'annoter un document sans connaissance préalable du contenu, en se basant sur les références. On suppose pour cela qu'on dispose d'une base de départ annotée. Ces annotations pourront être propagées sur les nouveaux documents.

4.2 Les étapes de l'annotation

Afin d'annoter les documents sans se baser sur le contenu, nous nous appuyons sur le contexte du document qui se traduit par les liens de citations. Avant d'utiliser les annotations des documents références, nous devons répondre à trois questions :

1. Quelles citations considérer pour effectuer la propagation ? Evidemment, toutes les citations ne sont pas significatives pour le document source et pour déterminer son thème.
2. Comment annoter le document ?
3. Comment fusionner toutes les annotations issues des références sélectionnées ?

Les différentes étapes de l'approche sont décrites dans l'algorithme 1, elles répondent aux trois questions posées.

Ces étapes sont schématisées dans la figure 4.1.

Pour ajouter un nouveau document, noté d, dans la base initiale, nous procédons de la manière suivante :

1. Récupérer l'ensemble des documents cités par d dans un ensemble noté Ref_d.
2. Sélectionner les annotations les plus proches thématiquement. Pour cela nous devons trouver les références les plus proches par un regroupement thématique. En effet, regrouper thématiquement les documents de l'ensemble Ref_d consiste à déterminer les groupements thématiques les plus pertinents et éviter ainsi les références non pertinentes mais présentes dans Ref_d. Elles correspondent aux références qui sont citées mais qui ne sont pas significatives pour le document.
3. Importer les annotations des documents cités par d.
4. Sélectionner parmi les annotations importées les plus pertinentes pour les proposer comme annotations du document d.

Algorithme 1 : Etapes de l'annotation d'un document d

Données :
$D \longleftarrow \{d_1, d_2, ..., d_i\}$ ensemble des documents à annoter
A_{d_i} annotation de d_i
$Ref_{d_i} \longleftarrow \{Ref_{1_{d_i}}, Ref_{2_{d_i}}, ..., Ref_{j_{d_i}}\}$ ensemble des références de d_i
A_{Ref_j} annotation de Ref_j
début
  répéter
    sélectionner d_i de D
    supprimer d_i
    // Annotation de d_i
    classer Ref_{d_i} par proximité thématique
    $A_{d_i} \subset A_{Ref_j}$ annotations des références les plus pertinentes
    Retourner A_{d_i}
  jusqu'à $(D = \emptyset)$;
fin

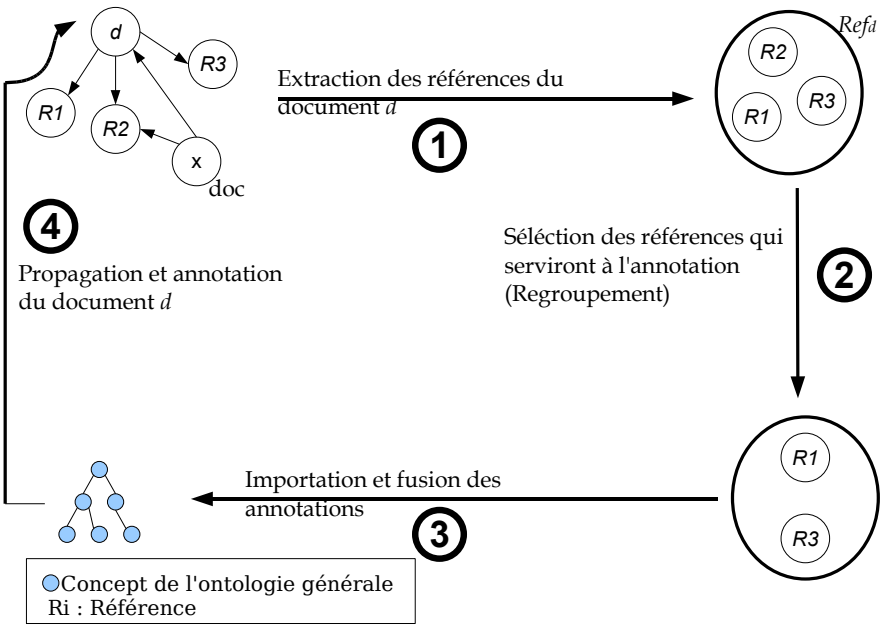

FIG. 4.1 – Les différentes étapes de la propagation

Afin de positionner notre approche, nous reprenons dans le tableau 4.1 les différents types d'annotations retenus dans la chapitre 2, et positionnons les caractéristiques de l'approche avec l'existant.

– L'utilisation d'une ontologie dans le processus d'annotation répond aux problèmes de l'indexation classique où l'absence de relations entre les termes se ressent lors de l'étape de recherche ;

		Informations utilisées pour annoter				
Méthodes d'annotation		Contenu	Contexte	Ontologie	Référence	Selection de référence
	Indexation classique	☒				
	Indexation sémantique	☒		☒		
	Propagation de mots clés		☒		☒	
	Propagation de métadonnées		☒		☒	
	Propagation de signatures lexicales	☒	☒		☒	

TAB. 4.1 – Comparaison des méthodes d'annotation retenues

– L'utilisation du contenu des documents n'est pas suffisant pour décrire tout le contexte d'un document. Par ailleurs, le contenu des document n'est pas forcément disponible ;
– L'utilisation de toutes les références d'un document pour l'annoter entraîne du bruit avec un nombre important de concepts et une faible pertinence.

N'ayant généralement pas accès aux documents pour des raisons que nous avons expliquées, et afin d'utiliser l'apport du contexte d'un document, l'approche présentée dans cette thèse se base sur les liens de citations pour l'annotation d'un document en se basant sur une ontologie du domaine.

L'approche développée répond aux critères du tableau 4.1, elle :
– se base sur le contexte de citation ;
– utilise une ontologie ;
– sélectionne les références pertinentes dans le document.

Les différentes étapes ainsi que nos choix sont détaillés tout au long de ce chapitre. Nous supposons disposer d'une base qui contient des articles ainsi que leurs relations de citation. Nous supposons également qu'une partie seulement de ces articles sont déjà annotés. Le problème consiste alors à annoter un nouveau document ajouté à cette base.

4.3 Regroupement thématique des documents

Un document, notamment lorsqu'il est technique ou scientifique, peut faire référence à plusieurs autres documents. En utilisant ce contexte de citation, cela nous offre la possibilité de situer thématiquement le document [Gar93]. Par la suite, nous utilisons cette caractéristique pour déterminer le thème d'un

document sans avoir accès à son contenu mais en utilisant simplement les références de celui-ci.

Lorsqu'un auteur cite un autre document, il estime que ce dernier donne de la valeur ajoutée à son document, et qu'il apporte de l'information. Parmi les limites de l'analyse des citations que nous avons énumérées dans le chapitre 2, les citations peuvent être erronées ou bien de nature critique. Un document peut aborder certains aspects mineurs, les références qui sont utilisées dans ces aspects mineurs ne devront pas être prises en compte pour l'importation des annotations. Toutes les références d'un document ne sont alors pas pertinentes pour la détermination du thème du document citant. Dans ce contexte, il est alors important de retrouver les sujets les plus importants abordés par le document, et d'ignorer les sujets les moins importants.

La première étape de notre démarche consiste à choisir les références à utiliser afin d'annoter le document cible. Le but de cette première étape consiste à ne garder que les références qui traitent du même sujet que le document source. Afin de déterminer les thèmes les plus importants dans l'ensemble des références d'un document, nous utilisons l'hypothèse de la co-citations. Cette mesure importante dans le domaine de la bibliométrie [Ros96] tient compte de l'hypothèse que, **si deux documents sont souvent cités ensemble alors ils sont thématiquement proches**.

Dans un premier temps, la matrice de co-citations représentant le graphe de co-citations des documents références Ref_d est calculée. Ensuite, les valeurs de similarités entre les couples de documents sont déduites, ceci représente la distance thématique entre les documents. Enfin, des classes de documents qui correspondent aux ensembles de documents partageant le même thème sont construites.

4.3.1 Construction du graphe de co-citations

L'utilisation de la méthode de co-citations dans l'annotation se traduit par le fait d'utiliser les annotations des références proches thématiquement afin d'annoter le document citant.

La figure 4.2 est un extrait du graphe de citation. Ce graphe indique que le document D cite les documents $d_1, d_2, d_3, d_4, d_5, d_6$. Dans ce cas, ces 6 documents sont co-cités au moins fois (par le document D).

La méthode de co-citations [Gar93] sert à calculer la ressemblance entre les documents cités par un document et non pas la ressemblance entre les documents citant. Elle utilise les références des articles scientifiques afin de calculer leur ressemblance thématique.

Le graphe de co-citations est crée à partir du graphe de citation où les nœuds sont les documents et les arcs valués par le nombre de fois où les noeuds (documents) sont cités ensemble. Par exemple dans la figure 4.3, la valuation

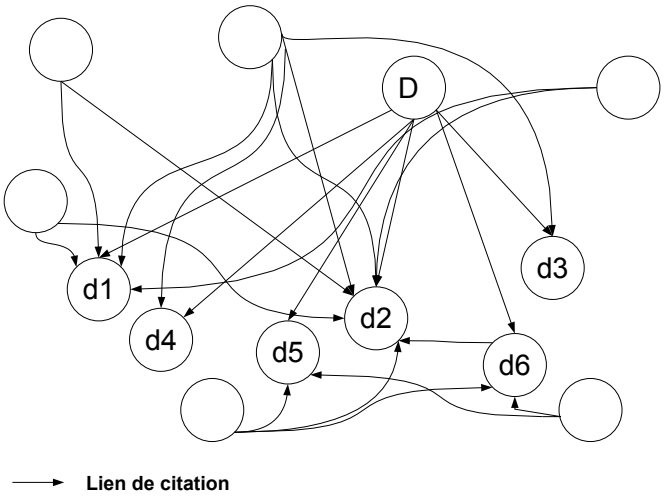

FIG. 4.2 – Extrait du graphe de citation des documents

2 entre les documents d_1 et d_3 indique que ces deux documents sont cités ensemble par deux documents.

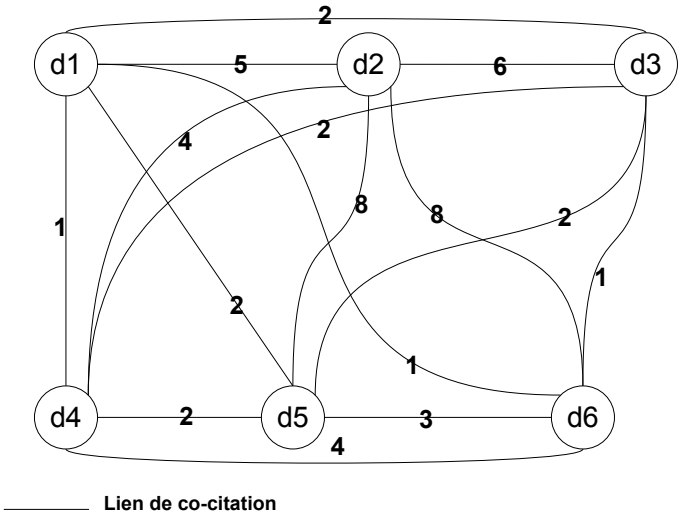

FIG. 4.3 – Le graphe de co-citations

La matrice de co-citations est une représentation du graphe de co-citations, elle correspond à une matrice carrée, cette matrice ne prenant pas en compte les documents citant.

La matrice de co-citations de l'exemple présenté dans la figure 4.3 est :

$$\begin{pmatrix} 0 & 5 & 2 & 1 & 2 & 1 \\ 5 & 0 & 6 & 4 & 8 & 8 \\ 2 & 6 & 0 & 2 & 2 & 1 \\ 1 & 4 & 2 & 0 & 2 & 4 \\ 2 & 8 & 2 & 2 & 0 & 3 \\ 1 & 8 & 1 & 4 & 3 & 0 \end{pmatrix}$$

- C_{34} (3ème ligne, 4ème colonne) correspond à la fréquence de co-citations de d_3 et d_4 qui est égale à 2 car ils sont cités ensemble par deux documents ;
- C_{14} est égale à 1 car d_1 et d_2 ne sont cités ensemble que par le document d.

4.3.2 Calcul de la similarité thématique entre les documents

Afin d'utiliser les références les plus proches thématiquement pour la propagation, nous calculons la similarité thématique entre ces documents en se basant sur l'hypothèse de la méthode de co-citations.

Un exemple est donné par Prime [PC04], qui expose une fonction de distance représentant la similarité thématique entre les documents :

$$S_{i,j} = 1 - \frac{{C_{(i,j)}}^2}{C_i \times C_j} \tag{4.1}$$

Dans l'équation 4.1 :

- $C_{(i,j)}$ représente l'indice de co-citations qui est défini comme le nombre de fois où les documents i et j sont cités ensemble ;
- C_i représente le nombre de fois où le document i est cité ;
- C_j représente le nombre de fois où le document j est cité ;

Cependant, dans cette fonction de distance, le dénominateur était conçu à l'origine pour normaliser la fraction, le résultat étant dans l'intervalle $[0,1]$.

En effet, les documents i et j sont indépendants et peuvent par exemple apparaître à des périodes différentes (si on suppose que le document i est plus ancien que j). Dans ce cas, le document i peut être cité plusieurs fois et on aura par conséquent un grand C_i. Cependant, si j est récent, alors dans ce cas C_j sera petit et comme $C_{i,j} \leq C_j$, nous aurons également $C_{i,j}$ petit.

Dans ce cas de figure, si on suppose qu'à chaque fois que le document j est cité, le document i est cité dans le même document, on s'attend à une proximité thématique. Cependant, on ne retrouve pas ce résultat si on applique la fonction de distance de Prime. En effet, comme $C_{i,j} \leq C_j << C_i$ alors $S_{i,j} \approx 1$. Ceci laisse entendre une disparité entre le document i et j. Or, même si le document j est cité à chaque fois avec le document i, cette disparité ne peut se résorber car le

paramètre C_i est complètement indépendant du document j. Ceci est illustré par l'exemple suivant :

- un document d_1 a été publié récemment et il est cité 10 fois, ce qui donne $C_1 = 10$;
- un document d_2 a été publié depuis longtemps et est cité 500 fois, ce qui donne $C_2 = 500$;
- les documents d_1 et d_2 sont cités 10 fois, ce qui implique qu'à chaque fois que d_1 a été cité alors d_2 a été cité aussi ;

À partir de ces données, on s'attend à avoir un résultat qui démontre que les deux documents sont proches, or en appliquant le fonction 4.1 le résultat est $S_{i,j} = 0.98 \approx 1$.

Afin de résoudre ce problème nous proposons une autre fonction qui utilise uniquement l'indice de co-citations.

Nous définissons la fonction de distance suivante :

$$S_{i,j} = \frac{1}{{C_{(i,j)}}^2} \tag{4.2}$$

L'équation 4.2 prend en compte simplement l'indice de co-citations entre deux documents afin de déterminer leur proximité thématique. Ainsi, plus deux documents sont cités ensemble, plus la distance $S_{(i,j)}$ sera proche de zéro. Dès que les références d'un document d sont récupérées, nous construisons le graphe de distances GC_d.

$$GC_d =< Ref_d, Ref_d \times Ref_d \times [0,1] > \tag{4.3}$$

Tel que décrit dans l'équation 4.4, le graphe de distances est un graphe complet où les nœuds représentent les documents cités dans d. Un lien entre deux documents i et j, est un lien valué avec la fonction de distance $S_{(i,j)}$ présentée dans l'équation 4.2. La représentation de ce graphe peut également être vue comme une matrice, appelée matrice de distance, MC, définie comme suit :

$$\begin{aligned} &MC_d : |Ref_d| \times |Ref_d| \\ &\forall i,j \in Ref_d,\ MC_d(i,j) = \begin{cases} S_{(i,j)} & \text{si} \quad i \neq j \\ 0 & \text{sinon} \end{cases} \end{aligned} \tag{4.4}$$

La matrice de distance de l'exemple présenté dans la figure 4.3 est :

$$\begin{pmatrix} 0 & 0.04 & 0.25 & 1 & 0.25 & 1 \\ 0.04 & 0 & 0.027 & 0.0625 & 0.015 & 0.015 \\ 0.25 & 0.027 & 0 & 0.25 & 0.25 & 1 \\ 1 & 0.0625 & 0.25 & 0 & 0.25 & 0.0625 \\ 0.25 & 0.015 & 0.25 & 0.25 & 0 & 0.11 \\ 1 & 0.015 & 1 & 0.0625 & 0.11 & 0 \end{pmatrix}$$

Après avoir calculé la similarité thématique entre les documents en nous basant sur la méthode de co-citations, il s'agit de regrouper les documents partageant le même thème et de séparer les thèmes traitant de thèmes différents. Pour cela, nous utilisons les méthodes de classification. Les documents regroupés dans la même classe ont une proximité thématique et par cela des annotations proches.

La classification a été utilisée pour améliorer le processus de recherche d'informations. D'aprés Rijsbergen [Rij79] : *"Closely associated documents tend to be relevant to the same requests"*, la requête était comparée à un document représentant de chaque classe de documents, ces documents traitant des mêmes thèmes. Les systèmes d'aujourd'hui permettent de faire une comparaison avec chaque document.

Il existe deux approches possibles pour classer les documents automatiquement.

- la classification supervisée,
- la classification non supervisée.

Classification supervisée

Dans la *classification supervisée*, on connaît les classes possibles et on a déjà un ensemble d'objets classés. Ces documents constituent un ensemble d'apprentissage. Le problème ici est d'arriver à associer un nouveau document à une classe en se servant de ceux déjà classés.

De nombreux algorithmes de classification supervisée existent. Par exemple, k-NN (k-nearest neighbor) est une méthode très connue dans le domaine de la classification supervisée automatique. Pour prédire la classe d'un nouveau cas (classer un nouveau document), l'algorithme cherche les k plus proches voisins de ce nouveau cas et sélectionne la réponse la plus fréquente de ces k plus proches voisins. La méthode utilise donc deux paramètres : le nombre k et la fonction de similarité pour comparer le nouveau cas à ceux déjà classés.

Dans le *classification supervisée*, un jeu d'entraînement est obligatoire. C'est un ensemble de documents dont la classe est connue afin qu'ils puissent être comparés avec les nouveaux documents. Ceci dans notre cas implique la classification d'un ensemble de documents de notre corpus avant la phase d'annotation, et demande un travail manuel considérable. Ceci nous a amenés à nous intéresser au deuxième type de classification qui est la *classification non supervisée*.

Classification non supervisée

La *classification non supervisée* ne nécessite pas de disposer de documents déjà classés, le seul paramètre à définir est le nombre de classes k. Le but ici est de mettre dans la même classe les documents similaires, afin de minimiser la distance entre les documents d'une même classe tout en augmentant la distance

entre les classes. Dans notre contexte, le choix du type de classification s'est donc porté vers la classification non supervisée

4.3.3 Choix de l'algorithme pour le regroupement

Dans cette section, nous présentons les différents algorithmes de regroupements de documents et présentons les raisons de notre choix pour un classifieur flou non supervisé.

Afin de regrouper les documents références servant à l'annotation, plusieurs algorithmes issus de différentes approches ont été étudiés :
- classification hiérarchique,
- algorithme des k-moyennes,
- extraction d'itemsets fréquents.

4.3.3.1 Fouille de données

Dans un premier temps, face au volume de données important, nous avons envisagé des méthodes issues de la fouille de données. Pour ce faire, nous recherchons les documents fréquemment cités ensemble. Ces "fortes" co-citations sont extraites sous la forme d'itemsets fréquents. Cette méthode a été très étudiée dans la littérature, notamment pour extraire des règles d'association dans la base de données.

Dans [RA93] le problème de la recherche de motifs d'association dans de grandes bases de données est défini de la manière suivante.

Définition 4.1 *Soit $I = \{i_1, i_2, ..., i_m\}$ un ensemble de m items. Soit $D = \{t_1, t_2, ...t_n\}$ un ensemble de n transactions ; chacune possède un unique identificateur appelé TID et porte sur un ensemble d'items (itemset) I. I est appelé un k-itemset où k représente le nombre d'éléments de I. Une transaction $t \in D$ contient un itemset I si et seulement si $I \subseteq t$. Le support d'un itemset I est le pourcentage de transactions dans D contenant I : $supp(I) = \frac{|\{t \in D | I \subseteq t\}|}{|D|}$. Une règle d'association est une application conditionnelle entre les itemsets, $I_1 \Rightarrow I_2$ où les itemsets I_1, $I_2 \subset I$ et $I_1 \cap I_2 = \emptyset$. La confiance d'une règle d'association r : $I_1 \Rightarrow I_2$ est la probabilité conditionnelle qu'une transaction contienne I_2 étant donné qu'elle contient I_1. Le support d'une règle d'association est défini par $supp(r) = supp(I_1 \cup I_2)$ et sa confiance par $conf(r) = supp(I_1 \cup I_2)/supp(I_1)$*

Exemple

Afin d'illustrer cette méthode, nous avons appliqué cette approche sur le graphe de citations extrait de notre base de documents. Le résultat est un ensemble de règles de la forme suivante (les numéros représentent les identifiants des documents) :

```
Frequent 2-itemsets:
itemset (occurrence)
38811 344990 (3)
13846 344990 (2)
11127 62675 (2)
11127 23953 (2)
Frequent 3-itemsets:
itemset (occurrence)
38811 23953 344990 (2)
11127 62675 23953 (2)
Frequent 4-itemsets:
itemset (occurrence)
11127 62675 23953 126094 (2)
```

Dans cet exemple, nous avons utilisé un ensemble de 18 documents. Les documents 344990 et 38811 sont cités ensemble par 3 documents, et par 2 documents avec le document 23953.

Par cette approche, les itemsets sont extraits en grande quantité. La définition d'un "bon" support est très difficile. De plus, les documents qui ne sont pas souvent cités sont éliminés alors qu'ils sont intéressants pour l'annotation. Même si des méthodes de clustering basées sur les itemsets fréquents ont été proposées [WXL99],[FWE03], il reste difficile de construire des classes quand de nombreux items sont non fréquents.

4.3.3.2 Classification non supervisée

Nous avons utilisé deux approches connues pour effectuer la classification non supervisée : l'algorithme des K-moyennes et la classification hiérarchique ascendante. Ces deux méthodes fonctionnent à partir de la notion de "distance" entre les objets à classer.

Afin d'utiliser la matrice de distances comme entrée à ces algorithmes, il faut donc veiller à ce que les valuations des liens définissent bien une distance au sens mathématique. Rappelons que les propriétés d'une distance sont les suivantes :

1. $d(x,y) \geq 0$ *(positivité)*
2. $d(x,y) = 0$ si et seulement si $x = y$ *(identité)*
3. $d(x,y) = d(y,x)$ *(symétrie)*
4. $d(x,z) \leq d(x,y) + d(y,z)$ *(inégalité triangulaire)*

Dans notre cas, comme les documents sont indépendants et que le calcul de $S_{(i,j)}$ ne prend en compte que l'indice de co-citations de ces documents, la spécification d'une distance au sens mathématique peut ne pas être satisfaite. En effet, le graphe de citation et la matrice de citation peuvent mener à des cas où :

$$MC_d(i,j) > MC_d(i,k) + MC_d(k,j), \qquad k \notin \{i,j\}$$

Plus généralement, on peut avoir une distance cumulée sur un chemin reliant deux documents qui est inférieure à la distance directe entre deux documents. Sur l'exemple de la figure 4.4 représentant le graphe de distances, nous avons par exemple :

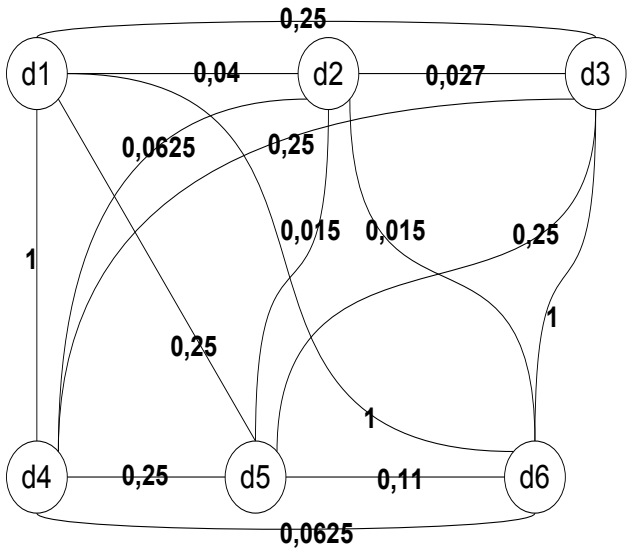

FIG. 4.4 – graphe de distance

Dans cet exemple, nous avons :

$$MC_d(1,3) > MC_d(1,2) + MC_d(2,3),$$

effectivement

$$0.25 > 0.04 + 0.027,$$

Dans ce cas, le graphe de distance ne respecte pas les propriétés d'une distance et l'utilisation de tels algorithmes ne sera pas appropriée.

Pour résoudre ce problème nous transformons la matrice de citation pour que la distance entre deux documents i et j soit minimale afin de répondre à l'inégalité triangulaire. On utilise pour cela l'algorithme Dijkstra [Dij59] afin de déterminer la distance minimale entre deux documents i et j :

$$\begin{aligned} &MC'_d : |Ref_d| \times |Ref_d| \\ &\forall i,j \in Ref_d,\ MC'_d(i,j) = \begin{cases} \text{Dijkstra}(i,j,MC_d) & \text{si} \quad i \neq j \\ 0 & \text{sinon} \end{cases} \end{aligned} \tag{4.5}$$

Après l'application de l'algorithme de Dijkstra, MC'_d définit bien un espace métrique car :

- la propriété de symétrie est satisfaite, S étant une fonction symétrique.
- S ne peut pas valoir zéro et par définition $MC'_d(i,j)$ vaut zéro quand $i = j$.

– l'inégalité triangulaire est satisfaite en utilisant Dijkstra.

Une fois cet algorithme appliqué, nous obtenons sur notre exemple les distances illustrées sur la figure 4.5.

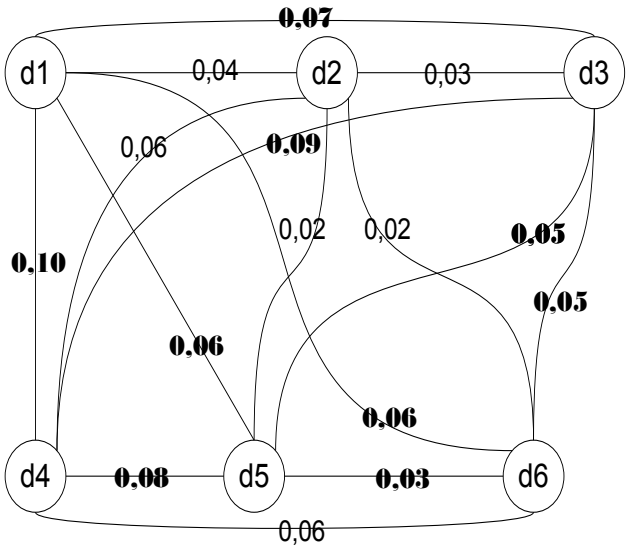

FIG. 4.5 – graphe de distance après application de l'algorithme de dijkstra

Nous avons par exemple $MC_d(1,3) > MC_d(1,2) + MC_d(2,3) = 0.07$. Nous pouvons appliquer les algorithmes de classification non supervisée, comme illustré ci-dessous.

Classification hiérarchique ascendante

L'approche hiérarchique ascendante a été testée car nous disposons de données ayant une distance définie. Ce type de classification se déroule en trois étapes :

1. trouver les similarités entre les paires d'objets ;
2. grouper les objets sous la forme d'un arbre, ceci est réalisé en considérant chaque élément dans une classe et regrouper les deux classes les plus proches au sens d'une distance ;
3. déterminer la coupe d'arbre en définissant le nombre de classes souhaité.

On appelle dendogramme l'arbre binaire reflétant la structure des données et permetant de regrouper les données similaires.

Cette méthode a été testée avec l'outil *XLSTAT* qui nous a fourni le résultat illustré dans la figure 4.6 pour l'exemple de la figure 4.3. La coupe définie nous donne comme résultat 4 classes, où la classe majoritaire regroupe les documents d_2,d_5 et d_6.

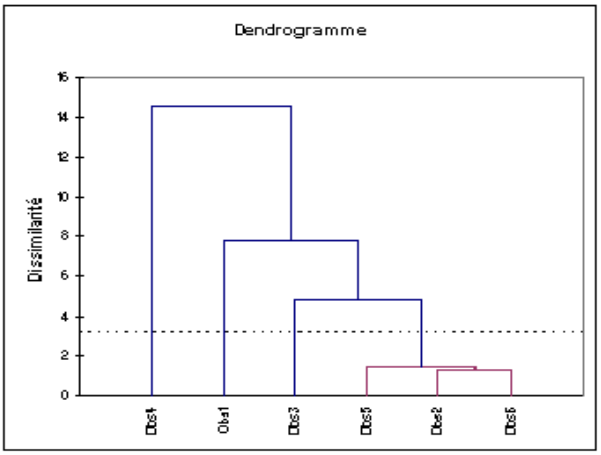

FIG. 4.6 – Résultat de la classification ascendante

Cependant, l'inconvénient avec une telle approche est la détermination du seuil de la coupe, ainsi que l'absence de chevauchement des classes. Un document peut être proche thématiquement de deux documents appartenant à deux classes différentes, s'il traite par exemple de deux thèmes différents. Dans cet exemple, ce cas est impossible car un document est affecté à une seule classe.

Algorithme des K-moyennes

L'algorithme des k-moyennes est une méthode de classification non supervisée. L'algorithme des K-moyennes est itératif, chaque itération est composée de deux étapes [GD02], [Dub04] :

- Recherche, pour chaque point d'observation, de son meilleur représentant parmi p référents, où chaque référent représente une classe ;
- Optimisation de chacun de ces référents pour qu'ils représentent au mieux les points d'observations en p classes.

Afin de classer les documents avec l'algorithme des k-moyennes à partir de la matrice de distances, nous avons utilisé le logiciel libre de data mining *WEKA* [WF05], qui nous a fourni le résultat suivant (figure 4.7).

Cet algorithme est simple et compréhensible et les éléments sont affectés automatiquement aux classes. Notons que les résultats des deux algorithmes dépendent des paramètres de départ. Pour les k-moyennes, nous avons fixé le nombre de classes à 3, en posant l'hypothèse qu'un document traite au maximum de 3 thèmes principaux. Le résultat des deux algorithmes aurait été le même en déterminant une coupe plus importante du dendogramme afin d'avoir 3 classes. Le résultat aurait été une classe majoritaire contenant les documents d_2, d_3, d_5, d_6.

Comme pour la classification hiérarchique ascendante, notons qu'avec "l'algorithme des k-moyennes" un document appartient à une seule classe, ce qui n'est pas pertinent dans notre cas. Effectivement, un document peut appar-

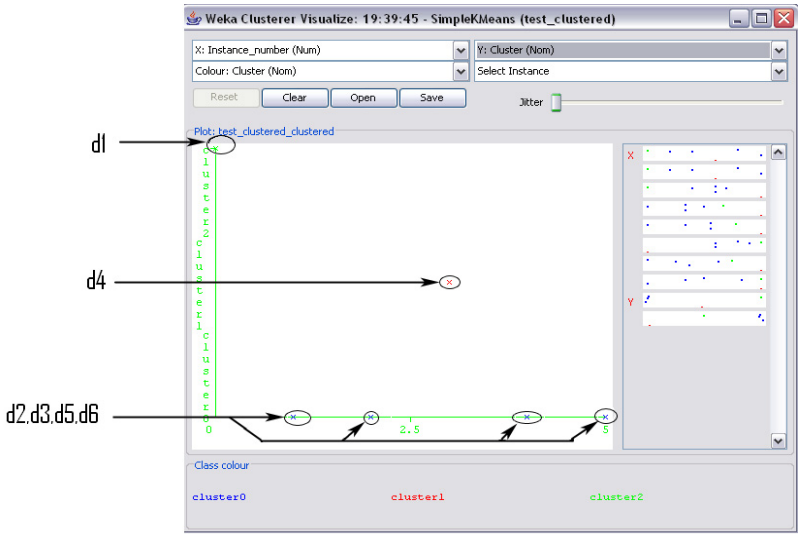

FIG. 4.7 – Résultat de l'algorithme k-means

tenir à un ou plusieurs thèmes ; par exemple : "l'informatique appliquée dans le domaine de l'eau". Ceci nous amène à choisir une solution permettant une multi classification, un document pouvant appartenir à plusieurs classes. Pour cette raison, nous avons décidé d'utiliser la forme étendu qui est l'algorithme des k-moyennes flou 'fuzzy c-means' [Dun73], [Bez81] qui utilise la théorie des ensembles flous.

4.3.4 Regroupement avec fuzzy C-means

Afin d'autoriser le chevauchement des groupes, nous utilisons l'extension des k-moyennes : les fuzzy C-means. Dans ce qui suit nous décrivons l'algorithme choisi pour notre classification *Fuzzy c-means* :

L'algorithme des c-moyennes (fuzzy c-means) fournit une partition de l'ensemble des données en sous-ensembles flous en optimisant la fonction 4.6 :

$$J = \sum_{i=1}^{n} \sum_{r=1}^{c} u_{ri}^{m} \|x_i - w_r\|^2 \tag{4.6}$$

avec la contrainte

$$\sum_{r=1}^{c} u_{ri}^{m} = 1 \tag{4.7}$$

- $X = \{x_i, i = 1....n\}$ est l'ensemble des données ;
- c le nombre de clusters cherchés. c doit être choisi ;
- w_r est le centre du cluster r ;
- u_{ri} est le degré d'appartenance de la donnée x_i au cluster r ;

– m est un hyper paramètre fixé généralement à 2 ;

Les différentes étapes de l'algorithme fuzzy c-means sont les suivantes :

1. choisir le nombre de classes c ;
2. initialiser la matrice $U = \left[u_{ij}\right]$, $U^{(0)}$;
3. à la kième étape : calculer les centres $C^{(k)} = \left[c_j\right]$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}.x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{4.8}$$

4. mettre à jour les degrés d'appartenance ;

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{4.9}$$

5. si $\left\| U^{(}k+1) - U^{(}k) \right\| < \varepsilon$ alors arrêt, sinon retour à l'étape 3.

Construction des classes

En spécifiant le nombre de groupes pour l'algorithme 'fuzzy c-means', noté N_{clusters}, le résultat du regroupement est alors une matrice MG_d de dimension $|Ref_d| \times N_{\text{clusters}}$ où chaque élément $MG_d(i,j)$ représente le degré d'appartenance du document i au groupe j. On notera que la somme des degrés d'appartenance d'un document aux différents groupes vaut 1.

Le résultat de notre exemple donne la matrice suivante en fixant le nombre de clusters à 3 :

$$\begin{pmatrix} 0.01 & 0.01 & 0.97 \\ 0.13 & 0.79 & 0.07 \\ 0.85 & 0.09 & 0.05 \\ 0.38 & 0.33 & 0.28 \\ 0.15 & 0.75 & 0.08 \\ 0.07 & 0.87 & 0.04 \end{pmatrix}$$

Ceci donne un regroupement des références illustré dans la figure 4.8. La première colonne définit le degré d'appartenance au cluster 1 et la deuxième au cluster 2,etc. Les documents d_2 et d_5 avec des degrés d'appartenance à C_2 respectivement 0.79 et 0.75 sont dans le deuxième cluster.

Fuzzy c-means est une classication à recouvrement qui indique le degré d'appartenance à la classe. Le document d_4 est entre le cluster 1 et le cluster 2 avec des degrés d'appartenance respectivement aux deux clusters 0.38 et 0.33.

Nous avons regroupé dans cette étape les documents par proximité thématique. Pour cela, nous avons testé les principales méthodes de classification, et le résultat s'est avéré très prometteur dans la mesure où les algorithmes fournissent des classes de références similaires afin d'annoter le document citant.

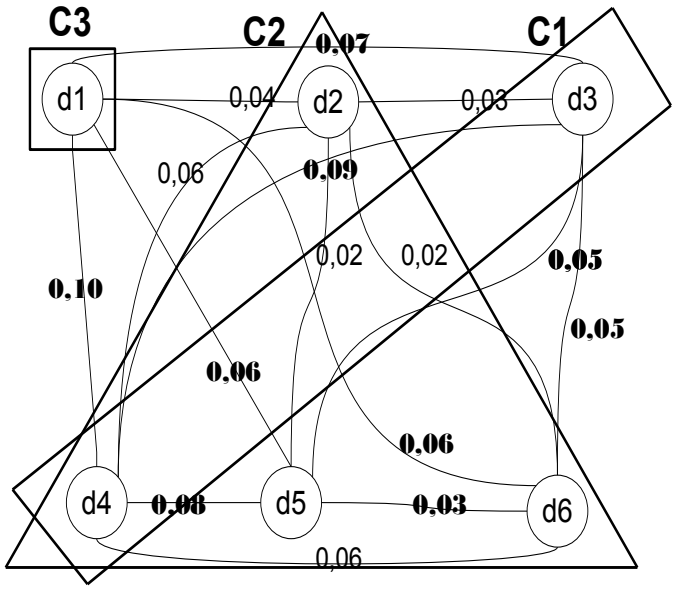

FIG. 4.8 – Le résultat du regroupement des références après application de l'algorithme fuzzy c-means

Nous avons opté pour une solution permettant le chevauchement des classes afin d'éviter d'éliminer des documents pertinents. Nous pensons tester d'autres algorithmes et les comparer avec les résultats actuels.

Les regroupements des documents par thème effectués, on annote le document d. La section suivante traite de l'importation des annotations des références.

4.4 Importation et ordonnancement des annotations

La présentation des annotations importées est faite en définissant un choix multi-critères pour sélectionner des annotations à utiliser dans la phase suivante. Le but dans cette section est d'importer et d'ordonner les annotations des documents cités par un document d. Nous définissons la liste des annotations importées pour le document d comme suit :

$$annotation_list_d = \bigcup_{i \in Ref_d} Annotation(i) \tag{4.10}$$

La fonction $Annotation(i)$ regroupe l'ensemble des annotations du document i, ces annotations correspondent aux annotations des références choisis dans l'étape précédente. Il faut rappeler que les éléments de l'annotation sont des concepts définis dans l'ontologie globale d'annotation. Il faut considérer

$annotation_list_d$ comme une liste ou bien comme un multi-ensemble, c'est à dire un ensemble avec une redondance des éléments, car deux références peuvent avoir des concepts communs.

L'ensemble des annotations du document d est alors défini comme suit :

$$annotation_set_d = \bigcup_{x \in annotation_list_d} \{x\} \tag{4.11}$$

Il s'agit maintenant de doter cet ensemble d'une fonction d'ordre total en se basant sur plusieurs critères. Le processus d'annotation est décrit dans l'algorithme 2.

Algorithme 2 : Importantion des annotations

Données :
A_d annotation de d
$Ref_{d_i} \longleftarrow \{Ref_{1_{d_i}}, Ref_{2_{d_i}}, ..., Ref_{j_{d_i}}\}$ ensemble des références de d_i
$C \longleftarrow \{C_1, C_2, ..., C_i\}$ ensemble des clusters des références de Ref_j
A_{Ref_j} annotation de Ref_j
début
  //Ordonner les clusters par ordre d'importance
  Ordonner(C)
  //Ordonner les références par degrés d'appartenance
  Ordonner (A_{Ref_j})
  //Doter les concepts de leur nombre d'occurrence
  Calcul (A_{Ref_j})
  Retourner A_d
fin

Nous retenons les critères suivants pour ordonner les annotations dans l'ensemble $Annotations_d$:

1. L'importance du cluster contenant le document d'où l'annotation a été importée. Si l'annotation apparaît dans plusieurs documents, on considérera l'importance du cluster le plus grand (maximal). Le thème du document à annoter étant le critère de sélection des références, nous considérons comme importants les références qui traitent du thème principal du document (la classe la plus importante).
2. Le degré d'appartenance du document au cluster d'où l'annotation a été importée. Si le concept apparaît dans plusieurs documents, on ne considérera que le document qui a un degré d'appartenance du cluster maximale. La similarité des documents se traduit par le degré d'appartenance du document à la classe.
3. Le nombre de fois où le concept apparaît dans la liste $annotation_list_d$. Un concept qui apparaît dans plusieurs références est plus pertinent qu'un concept qui n'apparaît qu'une seule fois.

Concernant le premier critère d'ordre, nous partons de l'hypothèse que les groupements importants définissent la thématique du document d. Nous utilisons la matrice d'appartenance aux clusters GR_d. En effet, on détermine l'appartenance d'un document à un groupe en utilisant le maximum des degrés d'appartenances. Le cardinal de chaque cluster est le nombre des documents contenus dans celui-ci. Le cardinal des groupes nous indique un premier critère afin de déterminer quels sont les groupes les plus importants. Dans la figure 4.8, le groupe le plus important est celui du cluster $C2$, avec un cardinal égal à 4.

Dans notre exemple on sélectionne les documents d_2, d_4, d_5 et d_6, avec les annotations suivantes (figure 4.9).

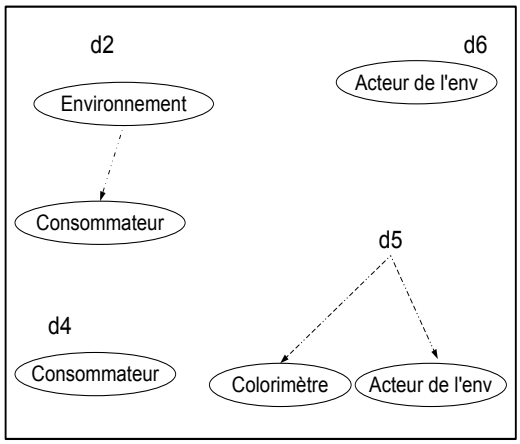

FIG. 4.9 – Annotation des documents cibles

En ce qui concerne le deuxième critère d'ordre, le degré d'appartenance d'un document aux différents groupes est déjà calculé dans la matrice GR_d. En ce qui concerne les documents de la figure 4.8, on se réfère à la matrice calculée lors de l'étape de regroupement, le degré d'appartenance des documents d_2, d_4, d_5 et d_6 est le suivant :

$$\begin{pmatrix} 0.79 \\ 0.33 \\ 0.75 \\ 0.87 \end{pmatrix}$$

Pour le troisième critère d'ordre, il s'agit simplement de calculer la répétition de l'annotation dans la liste $annotation_list_d$. On remarque que dans notre exemple *consommateur* est dans d_2 et d_4, alors que *environnement* n'est que dans d_2.

La fonction d'ordre est un ordre total et tous les éléments de l'ensemble $annotation_set_d$ peuvent être ordonnés. On trouve ainsi les annotations importantes en fonction de l'ordre suivant :

1. Les annotations qui proviennent des documents situés dans des clusters importants ;
2. Au sein des annotations qui proviennent d'un même cluster ou bien de clusters qui ont la même importance, les annotations importantes sont celles qui proviennent des documents qui ont un degré important d'appartenance au cluster. Ainsi, l'importance de l'annotation d'un document dépend de l'importance du document dans le cluster. Ici, les documents sont classés par ordre décroissant comme suit : d_4, d_6, d_5 et enfin d_2.
3. Si les annotations proviennent d'un même document ou de documents qui ont le même degré d'appartenance au même cluster, alors nous considérons comme importantes les annotations redondantes.

Dans la deuxième étape les annotations sont ordonnées de cette façon :
- consommateur,
- acteur de l'environnement,
- colorimètre,
- environnement.

avec *consommateur* et *acteur de l'environnement* dans deux documents différents.

4.5 Conclusion

Ce chapitre décrit notre système d'annotation automatique de documents. L'annotation manuelle de document est une tâche difficile, voire impossible à réaliser, compte tenu du temps que cela nécessite. Nous avons utilisé la relation de citation d'un document avec les autres documents de la base afin de l'annoter en se basant sur une ontologie du domaine sans avoir besoin de son contenu. Ce type d'annotation, contrairement à l'indexation classique avec une liste plate de mots clés, améliorera le processus de recherche de documents et résoudra le problème de multilinguisme puisqu'un terme exprimé dans différentes langues est associé à un seul concept.

Afin d'utiliser les références d'un document, nous nous sommes posés trois questions : *(i)* quelles références prendre ? *(ii)* comment annoter le document ? et *(iii)* comment importer et fusionner ces annotations ?

Pour répondre à ces questions nous nous sommes basés sur un regroupement thématique des citations en utilisant l'algorithme *fuzzy c-means*, le rapprochement thématique est construit à partir de la méthode de co-citations. Des critères ont été définis afin de sélectionner les annotations servant à la propagation, parmi ces critères il y a le degré d'appartenance d'une référence à un groupe et le nombre d'occurrences d'une annotation.

Afin de faciliter la lecture de ce chapitre, nous avons illustré les étapes des petits exemples. Notre approche n'est pas restée au niveau théorique, elle a été implémentée et le résultat de l'annotation a été testé sur une grande

base documentaire (plus de 500000), et les évaluations sont très encourageants. L'expérimentation et l'évaluation de l'approche sont présentées dans le chapitre 7.

Dans le chapitre suivant, nous présentons l'approche pour l'enrichissement de l'ontologie utilisée dans le processus d'annotation.

5

Enrichissement semi-automatique de l'ontologie

Dans ce chapitre, nous présentons une approche semi-automatique permettant d'enrichir l'ontologie du domaine, ainsi que la mise à jour des annotations des documents à partir des nouveaux concepts de l'ontologie. Ce travail a été développé en se basant sur deux constats :

- le manque de termes lors de la recherche guidée par les concepts du domaine ;
- le nombre important de termes composés.

L'approche est basée sur une représentation des sessions de recherche pour la découverte de nouveaux concepts. Elle est complétée par une analyse linguistique.

Sommaire

5.1 Introduction 87
5.2 Prérequis 88
5.3 Ontologie du SEMIDE 89
5.4 Enrichissement de l'ontologie par exploitation des requêtes 90
5.5 Analyse linguistique des termes composés . . 96
5.6 Etude de l'impact sur l'annotation d'un document 100
5.7 Conclusion 102

5.1 Introduction

L'ingénierie ontologique consiste en la recherche de concepts généraux, réutilisables, partageables et durables pour construire un modèle de connaissances capable d'aider des personnes à résoudre des problèmes [Miz04]. L'utilisation des ontologies dans différents secteurs (biomédical, automobile...) a connu une expansion due essentiellement à un souhait de s'orienter vers une vision Web sémantique. Le but commun, dans ces différents secteurs, est d'optimiser la représentation des connaissances et leurs exploitation.

Le partage de l'information au niveau du SEMIDE ne peut se faire que si les acteurs d'un tel système partagent des concepts communs, d'un côté pour décrire le domaine et d'un autre coté pour décrire l'information, et ainsi faciliter son utilisation. Afin de répondre aux besoins du SEMIDE avec les différentes contraintes d'hétérogénéité et de multilinguisme des ressources, nous proposons une approche d'annotation des documents qui a comme résultat un ensemble de concepts issus de l'ontologie du domaine. Cette ontologie est un ensemble de termes utilisés par la communauté, ces termes étant définis par rapport à l'information existante.

Parce que cette annotation est basée sur une ontologie, nous avons également traité le problème de l'enrichissement de l'ontologie. En effet, la masse de documents évolue dans le temps, utiliser la même ontologie ne suffit pas, car l'apparition de nouveaux documents entraîne forcément l'apparition de nouveaux concepts pour décrire le domaine. Cet enrichissement servira également à affiner ou à enrichir l'annotation existante.

Dans ce chapitre, nous proposons une solution pour enrichir l'ontologie du domaine. Notons que cet enrichissement ne comprend pas la suppression et la transformation de concepts. L'ontologie enrichie ici est un ensemble de termes reliés par la relation de spécialisation/généralisation et d'autres relations comme la synonymie (section 5.3).

Le traitement des ontologies est un domaine très actif où on cherche à automatiser au mieux ce processus car il n'existe pas d'approche unique. La construction/enrichissement des ontologies dépend dans un premier lieu des besoins des acteurs ainsi que de l'existant. L'approche que nous proposons est adaptée au contexte du SEMIDE, elle se base essentiellement sur la phase de recherche effectuée par les utilisateurs. Le SEMIDE souhaitant faire participer les différents acteurs du système pour la mise en œuvre de l'ontologie du domaine, nous avons traduit ceci par la saisie des utilisateurs de termes n'appartenant pas à l'ontologie afin d'effectuer une recherche sur la base documentaire. Cette recherche est guidée par l'ontologie du domaine, tout en laissant la liberté à l'utilisateur de saisir d'autres mots clés s'il considère que les termes existants ne suffisent pas pour constituer la requête. Notre approche consiste à exploiter ces termes, trouver les relations lorsqu'elles existent avec les termes de l'ontologie. Cette première approche est complétée par une analyse linguistique sur les termes composés.

Ce chapitre est composé des parties suivantes :

- quelques définitions utiles sont données dans la section 5.2 ;
- dans la section 5.3, nous présentons l'ontologie du SEMIDE ;
- dans la section 5.4, nous développons la partie principale de notre approche qui consiste à utiliser les requêtes des utilisateurs afin de déterminer de nouveaux termes. Elle est complétée par une analyse linguistique dans la section 5.5 ;
- la section 5.6 étudie l'impact de l'enrichissement sur la phase d'annotation.

5.2 Prérequis

Dans cette section sont rappelées quelques définitions utiles pour la suite de ce chapitre sur les treillis de concepts.

5.2.0.1 Notions sur les treillis de concepts

Le treillis de Galois ou treillis de concepts est une structure mathématique permettant de représenter les classes non disjointes sous-jacentes à un ensemble d'objets [NN05].

Contexte Un contexte est un triplet $K = (O, A, \zeta)$ où O est un ensemble d'objets ou d'individus, A est un ensemble d'attributs ou de propriétés et ζ est une relation binaire entre O et A.

Un contexte $K = (O, A, \zeta)$ peut être représenté sous forme d'un tableau, où une ligne correspond à un objet avec ses attributs.

Treillis Un treillis est un ensemble ordonné dans lequel deux éléments quelconques ont une borne supérieure et une borne inférieure. Un treillis complet est un treillis pour lequel tout élément possède une borne supérieure et une borne inférieure.

Correspondance de Galois Soit le contexte $K = (O, A, \zeta)$, f une application $P(O)$ dans $P(A)$ et g une application $P(A)$ dans $P(O)$, f et g étant défnies de la manière suivante :

- $f : P(O) \rightarrow P(A) f(O_i) = \{a \in A | (o, a) \in A, \forall o \in O_i\}$ intention ;
- $g : P(A) \rightarrow P(O) g(A_i) = \{o \in O | (o, a) \in A, \forall a \in A_i\}$ extention ;

Le couple (f, g) est appelé la correspondance de Galois sur K.

Concept formel Soient $O_i \subseteq O$ et $A_i \subseteq A$, (O_i, A_i) est un concept ssi :

- O_i est l'extention de A_i ;
- A_i est l'intention de O_i ;

$O_i = g(A_i)$ et $A_i = f(O_i)$

Treillis de galois Soient $f : O \rightarrow A$ et $g : A \rightarrow O$ deux fonctions définies sur les treillis $(O, \leq_O)$ et $(A, \leq_A)$, telles que (f, g) est une correspondance de Galois.

Soit $G = \{(o, a)$, où o est un élément de O et où a est un élément de A, tel que $o = g(a)$ et $a = f(o)\}$. Soit $\leq$ la relation d'ordre définie par : $(o1, a1) \leq (o2, a2)$ ssi $a1 \leq_A a2$. $(G, \leq)$ est un *treillis de Galois*.

5.3 Ontologie du SEMIDE

Nous nous sommes intéressés dans le cadre de notre travail à mettre en œuvre une ontologie pour le SEMIDE. Dans le chapitre 3, nous avons rapproché notre ontologie d'un thésaurus dans les relations existantes, cependant, ces relations peuvent évoluer dans le temps. L'ontologie existante est basée sur les termes et les relations du thésaurus de l'office Internationale de l'Eau (OIEau[19]). Ceci a été réalisé principalement afin de :

1. partager le savoir commun entre la communauté ou les agents logiciels : par exemple chaque *PFN*[20] possède des informations sur l'eau et nous souhaitons partager et publier les mêmes termes. Le système sera capable d'extraire et de fusionner les informations des différents *PFN* fournisseurs, et l'utiliser pour répondre à la requête ;

2. pouvoir utiliser des ontologies d'autres domaines afin d'enrichir notre ontologie (notre vocabulaire) : par exemple si on souhaite ajouter le terme *informatique* dans notre ontologie, on peut réutiliser l'ontologie de ce domaine pour enrichir la nôtre ;

3. faire en sorte de rendre le domaine explicite.

L'ontologie du SEMIDE est décrite par les concepts du domaine et les relations qui les relient. La figure 5.1 illustre un extrait cette ontologie :

- un nœud représente un concept, représenté par un cercle dans la figure (par exemple le concept C_1) ;
- les concepts sont reliés par des arcs orientés définissant la relation de spécialisation/généralisation, ici le concept C_2 est une spécialisation du concept C_1 ;
- à chaque concept sont reliés des termes, un terme pouvant avoir des termes synonymes ;
- les termes sont dans différentes langues, actuellement arabe, anglais et français et bientôt espagnol. le terme *Hydraulique agricole* est associé au concept C_1.

[19] http ://www.oieau.fr/
[20] Point Focal National

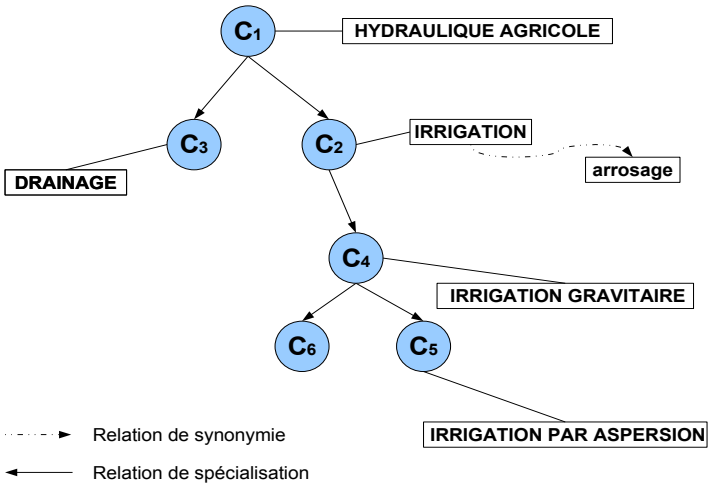

FIG. 5.1 – Extrait de l'ontologie du Semide

5.4 Enrichissement de l'ontologie par exploitation des requêtes

5.4.1 Principe de la méthode

L'idée de notre approche est basée sur l'exploitation des requêtes des utilisateurs afin d'enrichir l'ontologie du SEMIDE. La figure 5.2 donne un aperçu des différentes étapes pour l'enrichissement de l'ontologie.

A partir du moteur de recherche, les utilisateurs soumettent leurs requêtes. Le moteur de recherche renvoie des réponses et l'utilisateur peut modifier sa requête (ici la requête est un ensemble de termes issus ou non de l'ontologie). Toutes ces requêtes sont sauvegardées dans un entrepôt de requêtes. Le processus complet pour chaque utilisateur constitue une session de recherche.

Les différentes étapes sont les suivantes :

1. identification des sessions de recherche par utilisateur, c'est-à-dire des ensembles de requêtes qui sont petit à petit affinées par les utilisateurs qui y rajoutent ou suppriment des termes. Nous utilisons le moteur de recherche du SEMIDE ;
2. représentation des sessions de recherche ;
3. ajout de nouveaux termes apparaissant dans les requêtes. Ces termes sont utilisés par les experts dans leur recherche et n'existent pas au préalable dans l'ontologie ;
4. enrichissement de l'ontologie avec ces termes.

Exemple de session de recherche Illustrons la constitution d'une session de recherche par un exemple. Un utilisateur U effectue ses recherches selon les étapes suivantes :

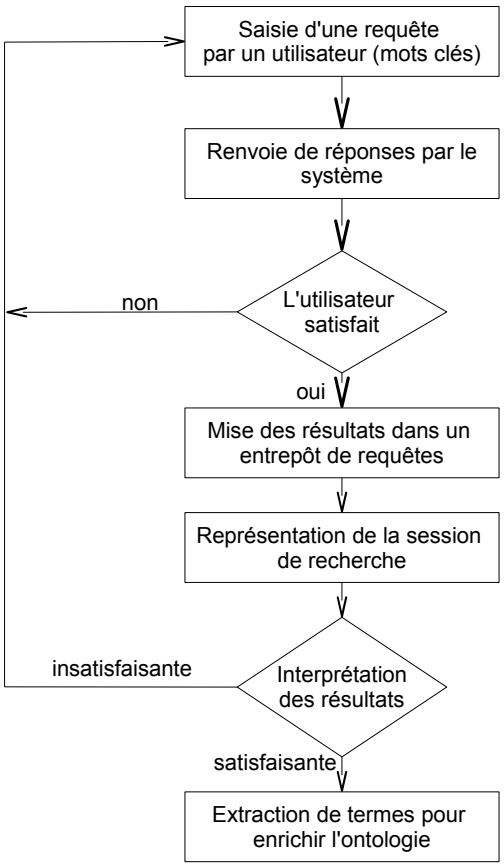

FIG. 5.2 – Enrichissement à partir de la phase d'interrogation

1. dans un premier temps, U saisie la requête *hydrogéologie*, il obtient comme réponse les documents $D1, D2, ..., D6$;
2. U essaie d'affiner sa requête en recherchant différentes combinaisons *(hydrogéologie, forage)*, *(hydrogéologie, alimentation de nappe)*, ou bien *(hydrogéologie, aquifère)* ;
3. l'utilisateur peut ne pas être satisfait et ajouter un terme qui n'est pas dans l'ontologie, comme par exemple la combinaison *(hydrogéologie, aquifère, barrage souterrain)* ;
4. l'utilisateur termine sa recherche quand il est satisfait du résultat, ou bien il refait une autre recherche s'il combine un ensemble de termes disjoints des précédents.

La représentation de ce type de structure comme une matrice d'objets et de verbes ou d'autres relations binaires entre deux ensembles se fait depuis les années 80 avec l'AFC (Analyse Formelle des Concepts), utilisée en analyse de données et extraction de connaissances. L'AFC utilise la structure du treillis.

Les treillis de Galois (ou treillis de concepts) fournissent un processus de classification formel et efficace pour découvrir et représenter des hiérarchies de concepts [LA04]. La représentation des sessions de recherche est basée sur cette structure.

5.4.2 Représentation des sessions de recherche par le treillis de concepts

Afin de représenter l'association (requête/documents), nous utilisons le treillis de concepts à partir d'un entrepôt de requêtes par utilisateur. La représentation des couples (termes de requêtes/documents réponses) sert à extraire les termes non présents dans l'ontologie, mais répondant à la recherche (besoin) des utilisateurs. Les treillis de concepts sont utilisés dans la recherche d'informations pour affiner ou généraliser la requête de l'utilisateur [Mes04]. L'utilisation des treillis est ici différente, l'utilisateur pose sa requête, l'affine ou la généralise en fonction des réponses et de sa recherche. Les différentes requêtes constituent une session de recherche représentée par un treillis de concepts.

Les deux étapes de notre processus sont les suivantes :
- la construction du treillis ;
- la construction de la hiérarchie.

5.4.2.1 Exemple de représentation de requête

Notre approche consiste à construire un entrepôt de requêtes à partir des recherches des utilisateurs. La représentation des réponses se fera par utilisateur. Le tableau 5.1 représente l'exemple précédent de correspondance entre 6 documents réponses des 5 termes $\{t_1, t_2, t_3, t_4, t_5\}$.
Les termes sont les suivants : (t_1, hydrogéologie), (t_2, aquifère), (t_3, forage), (t_4, alimentation de nappe), (t_5, barrage souterrain).

	t_1	t_2	t_3	t_4	t_5
D_1	X		X		X
D_2	X	X		X	
D_3	X		X	X	
D_4	X		X		
D_5	X			X	
D_6	X	X			

TAB. 5.1 – Exemple de correspondance

Nous avons ici, par exemple, la correspondance de Galois :
- $O_1 = \{D_3, D_4\} \Rightarrow f(O_1) = \{t_1, t_3\}$
- $A_1 = \{t_1, t_3\} \Rightarrow g(A_1) = \{D_1, D_3, D_4\}$

Dans cet exemple, on a le couple $(\{t_3\},\{D_1,D_3,D_4\})$ qui signifie que le résultat de la requête avec le terme t_3 donne pour réponse les documents D_1, D_3 et D_4.

Après avoir donné des définitions sur le contexte des treillis de Galois et expliqué notre approche de représentation de requêtes et documents réponses, nous présentons l'approche de construction du treillis de Galois avant de détailler ensuite l'étape d'enrichissement de l'ontologie générale.

5.4.2.2 Construction du treillis de Galois

On distingue deux types d'algorithmes pour la construction de treillis :

- les algorithmes non incrémentaux utilisés lorsque le texte est entièrement connu (on construit le treillis) ;
- les algorithmes incrémentaux où l'ajout d'un élément ne nécessite pas de nouveau le calcul du treillis ;

Nous nous basons sur l'algorithme de Bordat, non incrémental (algorithme 3) afin de construire le treillis de Galois et ceci pour chaque session de recherche. Cet algorithme est adapté dans le cas où le diagramme de Hasse[21] est à générer [NN05].

Illustration

Nous illustrons le résultat de l'algorithme sur l'exemple du tableau 5.1

$c = (\emptyset, \{t_1,t_2,t_3,t_4,t_5\})$.

$\delta_c = max\{f_c(D_1), f_c(D_2), f_c(D_3), f_c(D_4), f_c(D_5), f_c(D_6)\}$
$= max\{\{t_1,t_3,t_5\},\{t_1,t_2,t_4\},\{t_1,t_3,t_4\},\{t_1,t_3\},\{t_1,t_4\},\{t_1,t_2\}\}$
$= max\{\{t_1,t_3,t_5\},\{t_1,t_2,t_4\},\{t_1,t_3,t_4\}\}$

Dans ce cas les successeurs directs de c sont :

$c_1 = (\{D_1\},\{t_1,t_3,t_5\})$
$c_2 = (\{D_2\},\{t_1,t_2,t_4\})$
$c_3 = (\{D_3\},\{t_1,t_3,t_4\})$

De la même façon, on calcule les successeurs directs de c_1 :

$\delta_{c_1} = max\{f_{c_1}(D_2), f_{c_1}(D_3), f_{c_1}(D_4), f_{c_1}(D_5), f_{c_1}(D_6)\} = max\{\{t_1\},\{t_1,t_3\},\{t_1,t_3\},\{t_1\},\{t_1\}\}$
$= max\{\{t_1,t_3\},\{t_1,t_3\}\}$

Les successeurs directs de c_1 sont :

$c_4 = (\{D_1,D_4,D_3\},\{t_1,t_3\})$

La suite du calcul des successeurs directs se fait de la même manière. Le résultat de l'exemple du tableau 5.1 est le suivant (figure 5.3)

[21] Le diagramme de Hasse de $(G,\leq)$ est une représentation graphique qui contient toutes le informations concernant la relation d'ordre représentée.

Algorithme 3 : Algorithme de Bordat

Données :
un contexte $K = (O, A, \zeta)$
un concept $c = (X, Y)$
$X \longleftarrow \{x_1, x_2, ..., x_p\}$
δ l'ensemble des parties maximales
initialiser deux listes L_0 et L_1
suc_c l'ensemble des successeurs directs de c
début
  Insérer c dans L_0
  répéter
    c est le premier élément dans L_0
    supprimer c
    // Calcul des successeurs de c
    pour *i de 1 à p* **faire** $max(x_i) = vrai$
    $suc_c = \emptyset$
    pour *i de 1 à p* **faire**
      $X' = \{x_i\}$
      si $max(x_i)$ **alors**
        pour *j de* $i+1$ *à p* **faire**
          si $f_c(x_j) \subset f_c(x_i)$ **alors** $max(x_i) = faux$
          si $f_c(x_j) = f_c(x_i)$ **alors** $X' = X' \cup \{x_i\}$
        si $max(x_i)$ **alors** $suc_c = suc_c \cup (X', f_c(x_i))$
    Retourner suc_c
    Insérer les éléments de suc_c dans L_0
    $L_1 = L_1 \cup \{c\}$
  jusqu'à $(L_0 = \emptyset)$;
fin

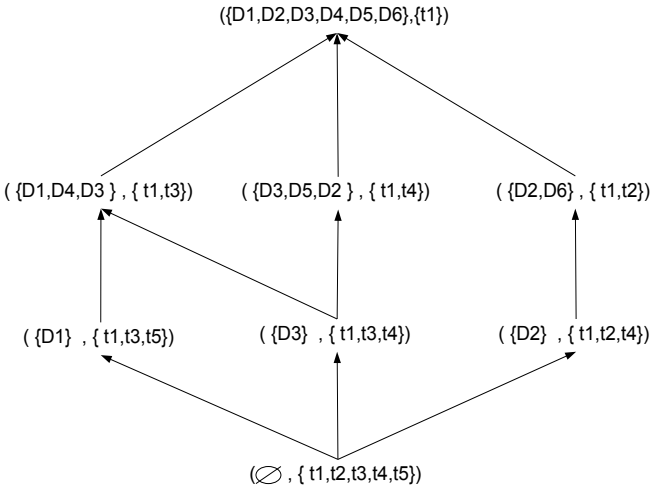

FIG. 5.3 – Treillis de Galois

5.4.3 Construction de la hiérarchie

La structure du treillis est utilisée afin d'extraire les relations candidates entre les termes. nous ne nous s'intéresserons dans le treillis qu'aux termes. Nous construisons la hiérarchie des termes afin de ne garder qu'une seule occurrence des termes. Nous partons de l'ensemble des termes du plus haut niveau et on élimine les occurrences de chaque élément dans les niveaux inférieurs.

Définition 5.1 *On définit l'ordre partiel P_t comme étant la restriction du treillis aux nœuds qui contiennent les éléments maximaux de t.*

Le résultat de la hiérarchisation des concepts est illustré dans la figure 5.4.

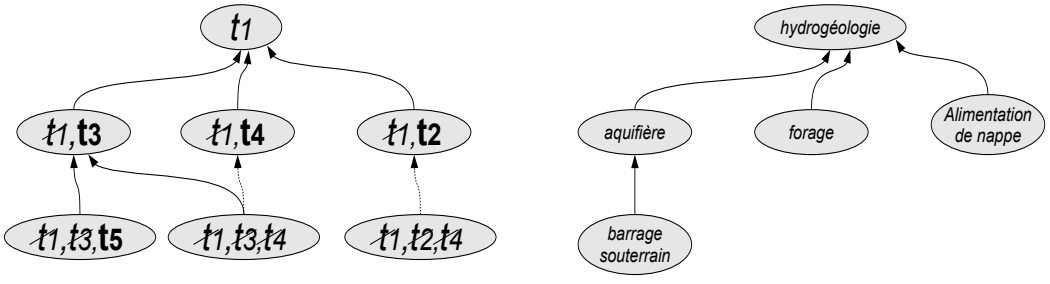

FIG. 5.4 – La hiérarchie des concepts

La première approche que nous avons réalisée afin d'exploiter ce treillis rejoignait celle de Sandersonet Croft, laquelle se base sur les co-occurences des termes pour déduire que x et un terme générique de y si les documents où y apparaît sont un sous-sensembe de ceux où x apparaît.

Cependant, dans la majorité des corpus que nous avons étudiés, cette inclusion ensembliste n'est jamais vérifiée, car le terme y apparaît régulièrement dans des documents où x n'apparaît pas, ce qui nous empêche d'utiliser cette propriété pour conclure sur le type de la relation entre x et y.

Cette hypothèse peut toutefois s'avérer pour certains corpus à un instant donné, mais les corpus de documents à exploiter sont en continuelle évolution, et il s'avère que la propriété ne reste jamais longtemps vérifiée. D'autre part, cette hypothèse ne peut pas parfaitement s'appliquer sur notre approche car le treillis ne prend pas en compte les documents qui répondent uniquement au terme y. Par exemple, dans la figure 5.4, le terme $t4$ n'est jamais considéré seul.

Les méthodes et outils étudiés dans le chapitre 3 de construction et d'enrichissement d'ontologies nous ont amenés à la constatation que le choix d'un corpus représentatif est une tâche difficile, et penser que l'on peut se baser sur le même corpus sans prendre en compte l'évolution de la masse d'information peut entraîner un manque dans l'ontologie.

Etre dans un système tel que le SEMIDE ne donne pas accès au contenu des documents. Ces documents sont chez les fournisseurs et sont décrits par un ensemble de métadonnées (titre, résumé, annotation...) et dans ce cas une solution se basant sur le contenu est impossible.

Le treillis construit à partir des sessions de recherche défini bien une relation thématique entre les deux termes mais qui ne se traduit pas forcément en une relation de généralisation. Pour cette raison, la relation de hiérarchie entre un terme x et un terme y dans le treillis est transformé en relation ontologique *related to*.

Définition 5.2 *Nous définissons la relation "related to" comme la relation déduite à partir du treillis décrivant deux termes qui sont liés par inclusion thématique. La thématique de y est incluse dans celle de x si x est ancêtre de y dans la restriction du treillis.*

Une phase complémentaire à l'enrichissement de l'ontologie basée sur le texte est présentée dans la section suivante. Cette analyse linguistique est basée sur l'existant qui est le nombre important de termes composés.

5.5 Analyse linguistique des termes composés

L'enrichissement de l'ontologie a été motivé en premier lieu par le besoin de la communauté du SEMIDE. Se baser sur les requêtes des utilisateurs nous permet d'extraire de nouveaux termes et de construire des relations entre des termes qui sont proches sémantiquement ou dans les mêmes champs sémantiques. Une approche complémentaire a été réalisée suite au constat que dans ce domaine il existait un nombre important de termes composés. Nous avons utilisé cette spécificité afin d'extraire des relations ontologiques avec les termes existants.

Cette phase a consisté à analyser le corpus du SEMIDE, et ceci afin d'extraire des termes pour définir des règles de construction de relation de spécialisation à partir de termes composés [AL05], [AL06].

Les étapes de l'analyse sont les suivantes :

- à partir d'un corpus de documents, nous obtenons un ensemble de termes composés. Ces termes sont des groupes[22] nominaux ;
- une analyse linguistique est faite sur ces termes ;
- nous avons retenu deux types de groupes nominaux : celui qui se compose d'un nom et d'un groupe prépositionnel qui dans notre cas se compose d'un nom et d'un adjectif.
- à partir de ces deux cas, nous avons défini des règles pour l'enrichissement de l'ontologie que nous présentons dans la suite

5.5.1 Règle sur le groupe prépositionnel

La première règle d'enrichissement s'applique quand le descripteur est composé d'un nom et d'un groupe prépositionnel de la façon suivante (algorithme 7) :

Algorithme 4 : Règle sur le groupe préposionnel

Données :
$C \in D = \{AB_i / \textit{ensemble de descripteurs}\}$
C=AB_i
n =*nombre des descripteurs*
$OG = \{\textit{ensemble de termes de l'ontologie}\}$
début
  répéter
    si *($A \in OG \wedge B$ est un goupe préposionnel)* **alors**
      Insérer C dans OG avec C terme spécifique de A
    $i = i + 1$
  jusqu'à $(i = n)$;
fin

1. Le descripteur C est un groupe nominal, C est de la forme AB
2. Si A appartient à l'ontologie du Semide, alors C est une spécialisation de A dans l'ontologie.

Exemple

L'expression *Directeur de l'eau* donne comme résultat :

[22]Groupe (de mots). Ensemble de mots ayant une unité sémantique, fonctionnelle ou rythmique. Synon. Syntagme. Groupe complément, sujet ; groupe verbal, nominal, prépositionnel.

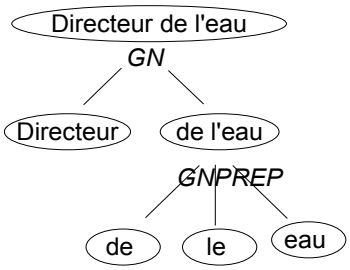

FIG. 5.5 – Règle sur le groupe prépositionnel

Directeur de l'eau

terme spécifique de *Directeur* si ce terme existe dans l'ontologie (figure 5.5).

5.5.2 Règle sur le groupe adjectival

La première règle d'enrichissement s'applique quand le descripteur est composé d'un nom et d'un groupe adjectival de la façon suivante :

Algorithme 5 : Règle sur le groupe adjectival

Données :
$C \in D = \{AB_i / \textit{ensemble de descripteurs}\}$
$C = AB_i$
$n =$ *nombre des descripteurs*
$OG = \{$*ensemble de termes de l'ontologie*$\}$
début
  répéter
    si $(A \in OG \wedge B$ *est un groupe adjectival)* **alors**
      Insérer C dans OG avec C terme spécifique de A
    $i = i + 1$
  jusqu'à $(i = n)$;
fin

Dans notre cas le groupe adjectival peut se composer soit d'un adjectif ou bien d'un adverbe et d'un adjectif, comme l'illustre la figure 5.6.

Exemple

L'expression *Zone artisanale* donne comme résultat :

Zone artisanale

terme spécifique de *Zone* si ce terme existe dans l'ontologie.

L'expression *Zone non saturée* donne comme résultat :

Zone non saturée

terme spécifique de *Zone* si ce terme existe dans l'ontologie.

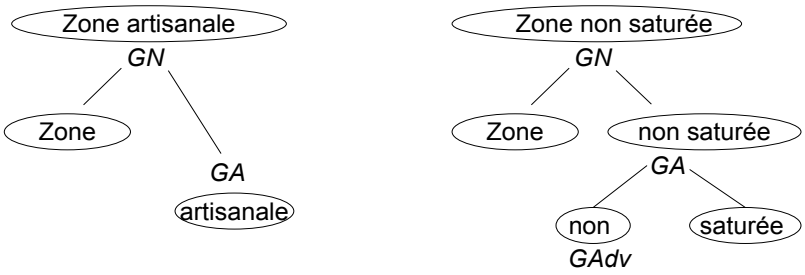

FIG. 5.6 – Règle sur le groupe adjectival

Le moteur de recherche du SEMIDE n'étant pas encore en ce moment, nous n'avons malheureusement pas encore pu construire les sessions de recherche. Cette approche est au stade théorique en attendant la mise en place du nouveau site du SEMIDE

L'analyse linguistique que nous faisons est basée sur la langue française. Cette analyse peut être assez facilement généralisée à d'autres langues en utilisant d'autres patrons syntaxiques. Nous utilisons l'analyseur morpho-syntaxique du français SYGFRAN, basé sur le système opérationnel SYGMART (Système Grammatical de Manipulation Algorithmique et Récursive de Texte) [Cha99]. SYGFRAN utilise un ensemble de règles de transformations d'éléments structurés, mettant en œuvre les règles de la grammaire française. Ces règles transforment une phrase en un arbre syntaxique.

SYGFRAN prend en entrée du texte et donne comme résultat une structure parenthésée qui correspond à l'arbre morpho-syntaxique de la phrase [YMP06]. L'analyse de la phrase "Le SEMIDE est un système d'information dans le domaine de l'eau" donne l'arbre syntaxique présenté en figure 5.7

Les noms des nœuds internes (rectangles) correspondent aux natures des constituants : ***PH*** pour PHrase, ***GN*** pour Groupe Nominal, ***GV*** pour Groupe Verbal, ***GA*** pour Groupe Adjectival, ***GNPREP*** pour Groupe Nominal PREPositionnel. Les noms des feuilles (ellipses) sont les formes canoniques des lexies (masculin, singulier, infinitif).

Comme nous l'avons expliqué dans notre approche, nous nous intéressons aux ***GA*** et ***GNPREP***. Ainsi pour les termes composés nous utilisons SYGFRAN pour extraire les relations, nous avons par exemple comme résultat : *aride* est un ***GA***. On obtient donc : *zone aride* terme spécifique *zone*.

Les patrons utilisés pour extraire les relations sont simples et peuvent être appliqués dans les différentes langues du système. En effet, pour cela, il faudrait mettre en place des traitements linguistiques plus lourds, à savoir une analyse-morphosyntaxique des termes composés. Pour le moment, nous ne disposons

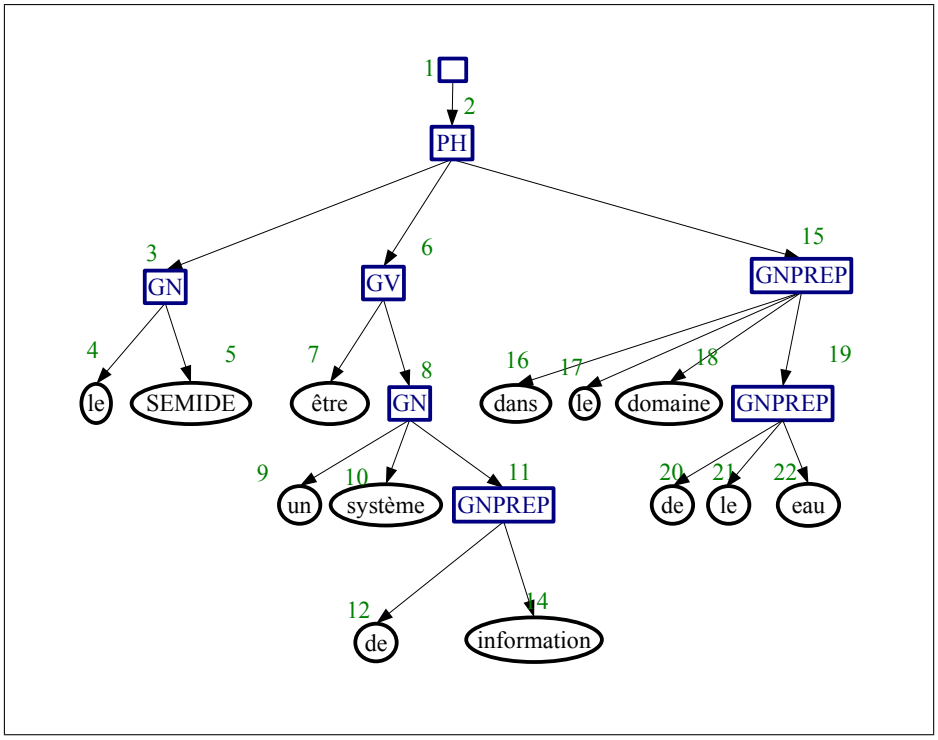

FIG. 5.7 – Exemple d'analyse de SYGFRAN

pas de tels analyseurs pour toutes les langues concernées. Pour ce que est des termes généraux hors du domaine, nous ne les incluons pas dans l'ontologie, mais nous gardons une trace des relations qu'ils entretiennent avec les termes du domaine.

L'utilisation d'une telle approche nous permet de hiérarchiser un certain nombre de termes composés dans l'ontologie. Cette approche est semi-automatique car elle nécessite une validation de la part d'un expert.

5.6 Etude de l'impact sur l'annotation d'un document

Dans cette section, nous présentons la méthodologie générale incluant la phase d'annotation des documents. La partie 5.6.2 décrit l'impact de l'enrichissement de l'ontologie générale sur la phase d'annotation des documents et comment ceci est géré par le système.

5.6.1 Résumé des étapes de notre approche

La figure 5.8 illustre les différentes étapes de notre approche.

Partant d'un nouveau document d, nous extrayons ses références. L'annotation de ce document basée sur l'ontologie est faite à partir des annotations de

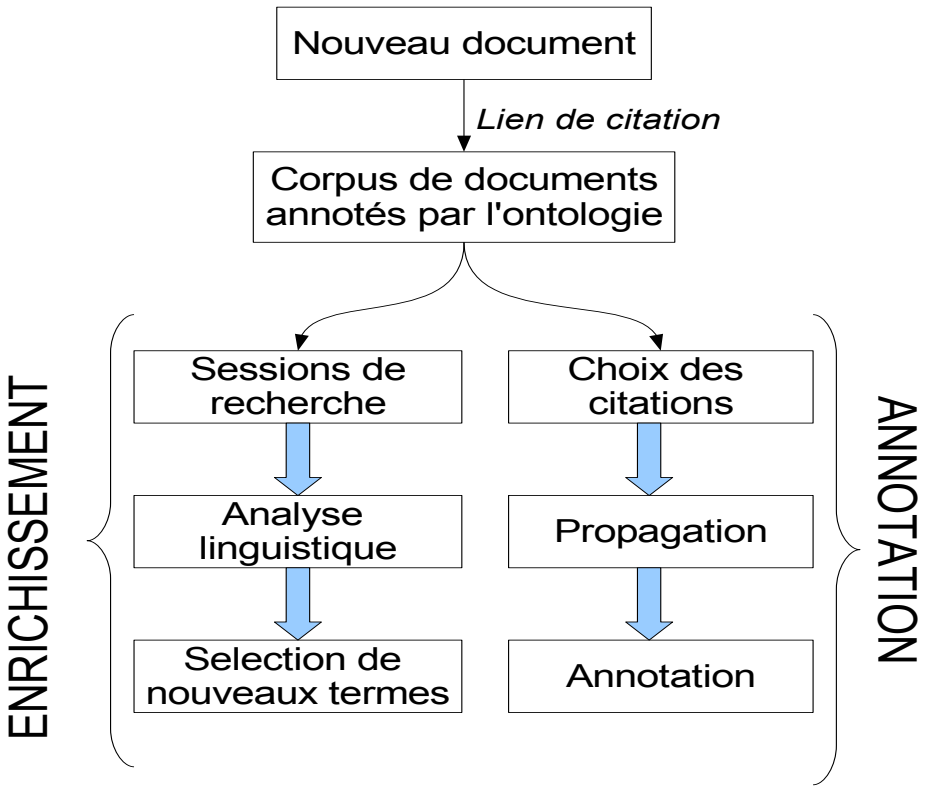

FIG. 5.8 – Les étapes de l'approche retenue

ses références *sans connaissance préalable de son contenu* selon un regroupement thématique. L'annotation du document effectuée, un autre problème est apparu, celui de l'apparition de nouveaux concepts dans le domaine. Pour cela nous nous basons sur les sessions de recherche des utilisateurs et une analyse linguistique.

5.6.2 Approche pour la révision de la phase d'annotation

L'enrichissement de l'ontologie peut agir sur l'annotation d'un document d existante de deux façons :

- la première est que le nouveau terme apparaissant dans l'ontologie est parent d'un ou plusieurs termes du document d. Dans ce cas, cela ne change pas l'annotation, étant donné qu'un document est annoté avec un ensemble de termes qui sont définis par un sous arbre de l'ontologie partant de la racine. Le nouveau concept viendra juste s'ajouter au chemin ;
- la deuxième, qui est moins simple, est qu'un nouveau terme apparaisse dans l'ontologie, que ce terme serve d'annotation à un document déjà

annoté et que l'ensemble de l'annotation existante n'ait pas le lien de spécialisation avec le nouveau terme.

Pour pallier ce problème d'apparition de nouveaux termes, nous avons défini une approche incrémentale qui se base sur les dates d'annotation, avec l'idée que si la date d'annotation d'un document référence est supérieure à la date d'annotation du document cible, alors nous mettons à jour l'annotation de ce dernier.

Algorithme 6 : mise à jour

Données :
- un graphe unidirectionnel dont l'ensemble des nœuds est :
$R_d = \{R_{1d}, R_{2d}, R_{3d}, ...\}$
- un document d annoté.
- $Date_d$ date de mise à jour du document d.
- R_d ensemble des références de d

début
  Définir une période de mise à jour
  répéter
    si $Date_d > Date_{R_{id}}$ **alors**
      Réannoter d
    sinon
      Prendre le nœud suivant
  jusqu'à *parcourir tous les nœuds du graphe* ;
fin

Notons qu'un document qui a plusieurs versions est considéré comme plusieurs documents avec des dates différentes.

5.7 Conclusion

Il n'existe pas de méthode générale pour la construction et l'enrichissement d'une ontologie. Une ontologie représentant un domaine, sa construction dépend des besoins des utilisateurs. Dans le chapitre 3, trois remarques essentielles avaient été faites :

– la difficulté du choix d'un corpus représentatif ;
– l'apport d'une analyse linguistique ;
– le rôle important des acteurs du domaine dans la construction d'une ontologie, ces acteurs étant utilisateurs de cette dernière.

Notre approche combine ces trois remarques en répondant aux besoins de la communauté. L'extraction de nouveaux termes issus de la recherche utilisateur consiste à combler le manque de l'ontologie, représentée par le besoin des utilisateurs. La structure des treillis de concepts est utilisée afin de représenter les sessions de recherche par utilisateur et extraire la hiérarchie des concepts. Une perspective intéressante à cette représentation serait de se baser sur les

treillis construits à partir des sessions de recherche, et ceci afin de construire des classes d'utilisateurs par rapport à leur recherche. Effectivement, une ontologie ne peut pas satisfaire tous les utilisateurs, surtout s'ils ont des niveaux de connaissance différents. D'un autre coté, plus une ontologie sera grande et plus elle perdra du sens.

La classification de treillis (session de recherche) permettrait d'enrichir l'ontologie par classe d'utilisateur et ainsi répondre aux besoins des acteurs en fonction de leur niveau.

L'analyse linguistique nous a aidés à identifier les termes composés, et les relier aux termes existants, et ceci en définissant des règles propres au corpus du domaine.

Troisième partie

Expérimentation et évaluation

6

L'outil RAS (Reference Annotation System)

Dans ce chapitre nous donnons une vue générale de l'outil réalisé dans le cadre de nos travaux. Cet outil traduit notre approche d'annotation.

Sommaire

6.1 Introduction 109
6.2 Les étapes d'annotation 109
6.3 Les fonctionnalités de l'outil 110
6.4 Le résultat de l'annotation 112

6.1 Introduction

Un outil a été réalisé pour montrer la faisabilité de notre approche. C'est un outil d'annotation de documents basée sur les liens de citation. Il utilise les technologies suivantes :

- Python[23] comme langage de script ;
- la base documentaire Citeseer[24] ;
- L'ontologie dmoz[25] (informatique) ;
- l'algorithme de classification fuzzy C-means [Dun73].

Cet outil a été réalisé dans le contexte d'un besoin réel, celui d'une communauté souhaitant partager l'information existante et ceci sous certaines contraintes, la plus importante étant celle de l'absence de contenu des documents à partager. Rappelons que cette raison a motivé notre approche d'annotation en nous basant sur le contexte de citation, les différents organismes de la communauté fournissant uniquement un ensemble de métadonnées.

La base du SEMIDE n'étant pas encore complète, l'expérimentation présentée dans le chapitre 7 à été effectuée sur la base documentaire Citeseer. Pour cette raison, l'illustration des différentes étapes du fonctionnement de l'outil a été réalisée sur la même base. A plus long terme, l'outil sera intégré dans le SEMIDE. L'outil que nous avons nommé RAS est disponible en ligne à l'adresse suivante : $http://www.lirmm.fr/annotation$

De manière générale l'outil permet de réaliser une annotation sur un document existant dans la base. Les nouveaux documents insérés dans la base sont sous forme d'un fichier XML. Enfin, l'outil permet de visualiser le résultat de l'annotation sous forme d'une liste de concepts de l'ontologie présentés sous la forme d'une hiérarchie.

Le chapitre suivant est structuré de la manière suivante :

- la section 6.2 décrit les différentes étapes d'annotation dans RAS ;
- les différentes fonctionnalités de l'outil sont présentées dans la section 6.3 ;
- le résultat du processus d'annotation est illustré dans la section 7.

6.2 Les étapes d'annotation

Afin d'annoter un nouveau document, l'outil regroupe les différentes étapes de notre approche. Elles sont illustrées sur la figure 6.1 :

1. un nouveau document d est sélectionné pour être annoté ;

[23] http ://www.python.org/
[24] http ://citeseer.ist.psu.edu/
[25] http ://www.dmoz.org/

2. les indices de co-citations calculés lors de la constitution de la base documentaire (cf. chapitre 7) des références du document d sont sélectionnées ;
3. les distances représentant les similarités thématiques entre les références sont calculées ;
4. les annotations des documents références sont importées et ordonnées ;
5. Enfin, le résultat de l'annotation est propagé sur le document d.

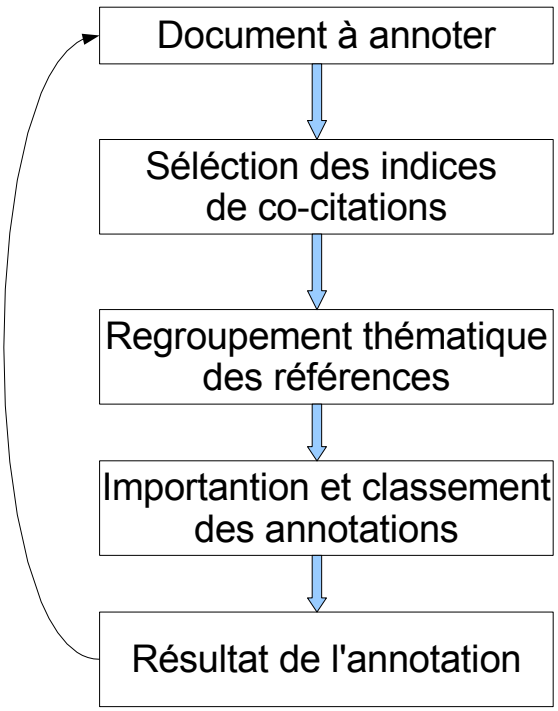

FIG. 6.1 – Etapes d'annotation dans RAS

6.3 Les fonctionnalités de l'outil

Dans cette section, nous présentons les différentes fonctionnalités de l'outil RAS, pour une meilleure compréhension, nous présentons ceci sous forme de captures d'écran.

6.3.1 Visualisation d'un document existant

On peut visualiser un document existant dans la base en affichant différentes informations qui sont : le titre, le résumé et les documents cités. Chaque document cité pointe vers sa propre description. Cette fonctionnalité nous permet d'avoir une vision générale sur un document que l'on souhaite annoter. Ceci est illustré sur les figures 6.3 et 6.2.

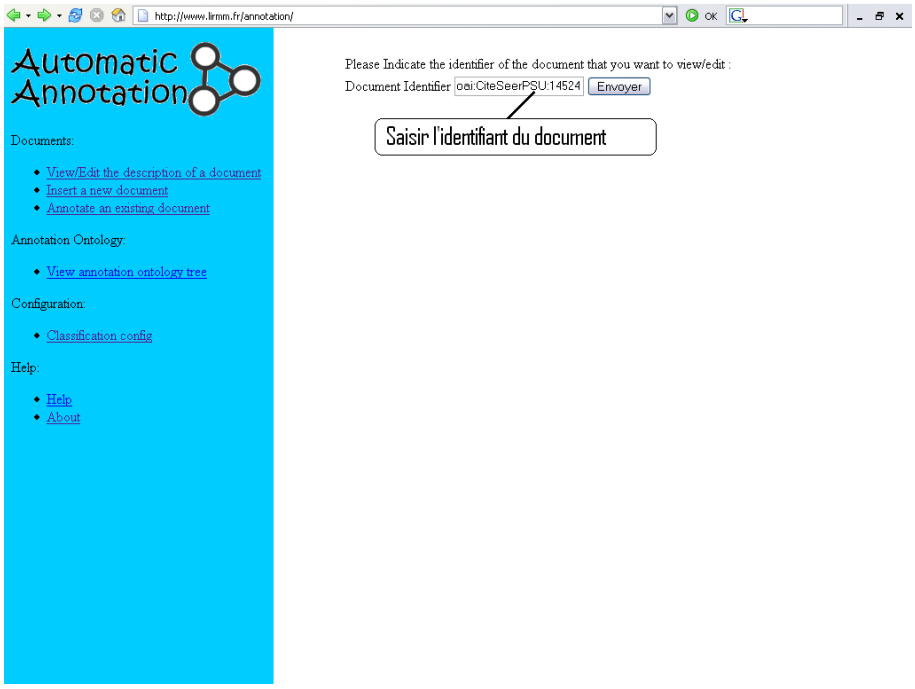

FIG. 6.2 – Sélection d'un document pour visualisation

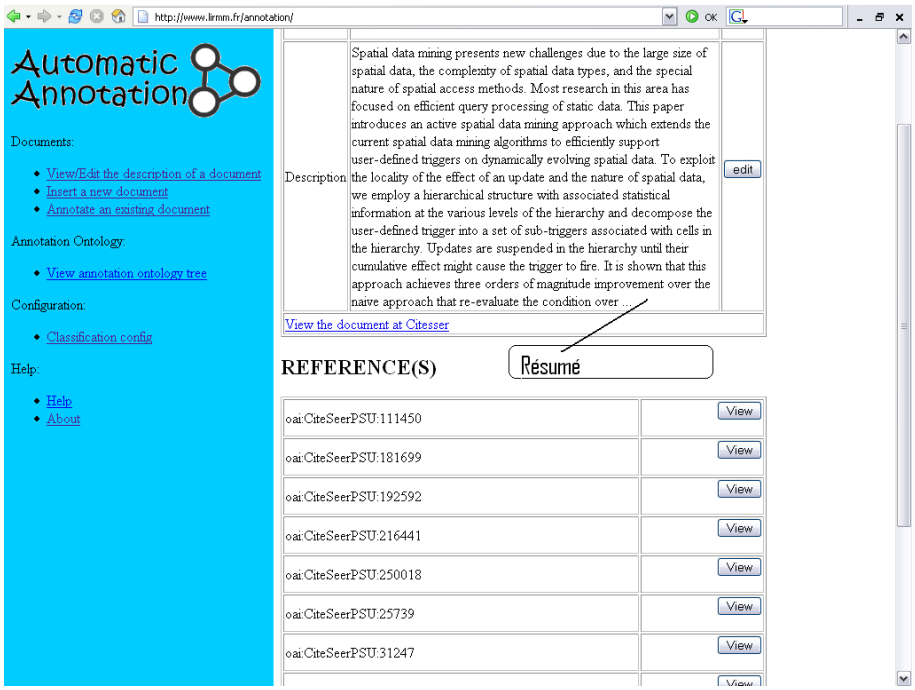

FIG. 6.3 – Visualisation d'un document

6.3.2 Insertion d'un nouveau document

L'insertion d'un nouveau document se fait en sélectionnant un document XML ou les différentes balises contiennent la description du document. Voici un exemple d'un document à insérer :

```
<document>
<id>oai:CiteSeerPSU:131886</id>
<title>Set-based Analysis of Reactive...</title>
<abstract>We present an automated abstract ...</abstract>
<references>
<ref>oai:CiteSeerPSU:175303</ref>
<ref>oai:CiteSeerPSU:205081</ref>
<ref>oai:CiteSeerPSU:208210</ref>
<ref>oai:CiteSeerPSU:215202</ref>
</references>
</document>
```

6.3.3 Ontologie pour l'annotation

L'ontologie utilisée pour l'annotation des documents de notre base est dmoz. Nous avons sélectionné le thème *Science Computer* et intégrer la hiérarchie dans notre base, afin d'annoter les documents avec cette ontologie. L'outil permet de visualiser l'ontologie utilisée, dans le figure 6.4, nous avons sélectionné les concepts fils du concept *Artificial intelligence*.

6.4 Le résultat de l'annotation

6.4.1 Calcul des similarités thématiques entre les références

La distance représentant la similarité thématique entre les documents est basée sur la méthode de co-citations. Afin de calculer cette similarité, notre première approche a été de nous baser sur la fonction de similarité de Prime qui prend en considération les degrés de citations. Pour des raisons exposées précédemment (chapitre 4), nous avons utilisé une autre mesure de similarité qui est :

$$S_{i,j} = \frac{1}{C_{(i,j)}{}^2} \tag{6.1}$$

Avec C_{ij} le nombre de fois où le document i est cité avec le document j.

Les différentes étapes du calcul de la matrice de distances sont les suivantes :

- séléction de la matrice de co-citations des documents références ;
- calcul de la matrice de distance ;
- application de l'algorithme de Djijkstra.

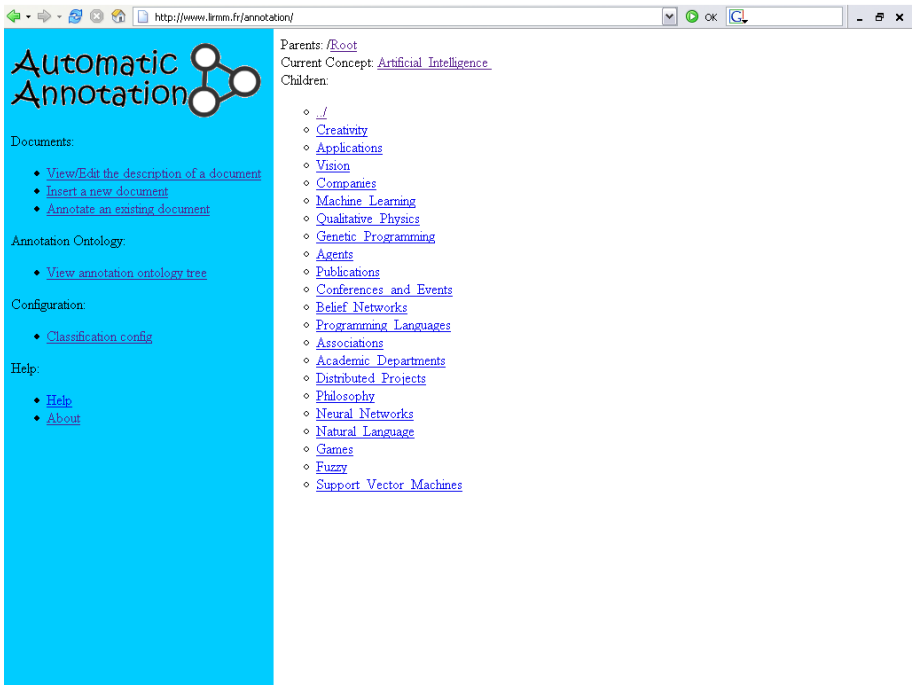

FIG. 6.4 – Ontologie dmoz utilisée pour l'annotation

La figure 6.5 illustre le graphe de distances généré par l'outil. Ceci permet de visualiser la similarité thématique entre les documents références.

6.4.2 Regroupement thématique des références

L'annotation d'un document se fait à partir d'un regroupement thématique des références, en utilisant une classification floue non supervisée.

Les paramètres de l'algorithme de classification utilisé *fuzzy c-means* peuvent être modifiées à partir de l'interface de l'outil, par exemple le nombre de classes.

6.4.3 Importation et propagation des annotations

A partir du résultat de regroupement, le système utilise l'ordre décrit dans notre approche afin d'annoter le nouveau document :

- l'importance du cluster ;
- le degré d'appartenance au cluster ;
- le nombre de fois où l'annotation apparaît.

Ces propriétés ont été implémentées et le résultat est un ensemble de concepts de l'ontologie ordonnés par degré d'importance. La figure 6.7 illustre le résultat de l'annotation d'un nouveau document. Nous remarquons sur cette capture d'écran qu'un choix est laissé à l'expert pour supprimer ou ajouter si besoin des concepts. Les concepts sont présentés sous la forme d'un chemin de l'ontologie globale commençant par la racine.

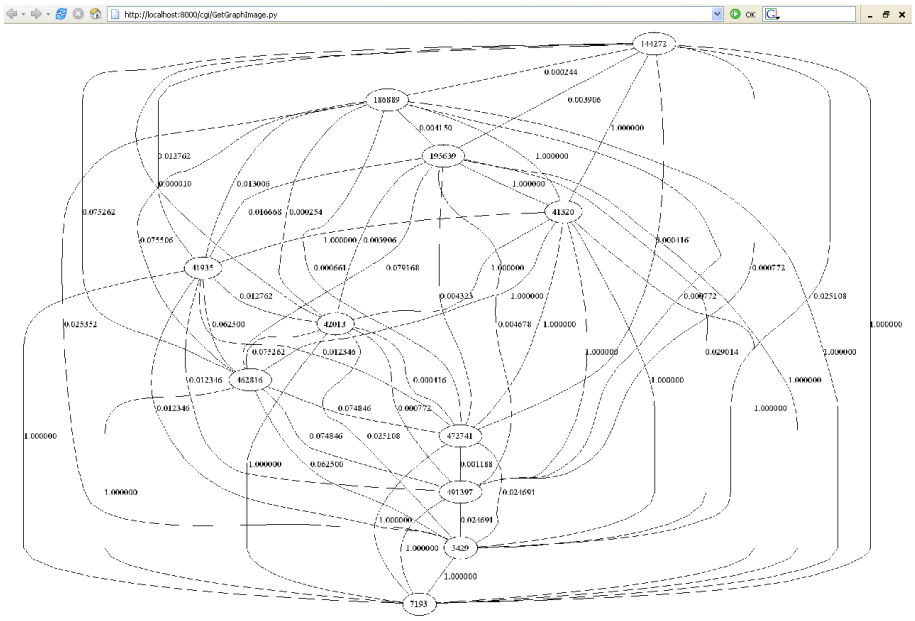

FIG. 6.5 – Graphe de distance avec RAS

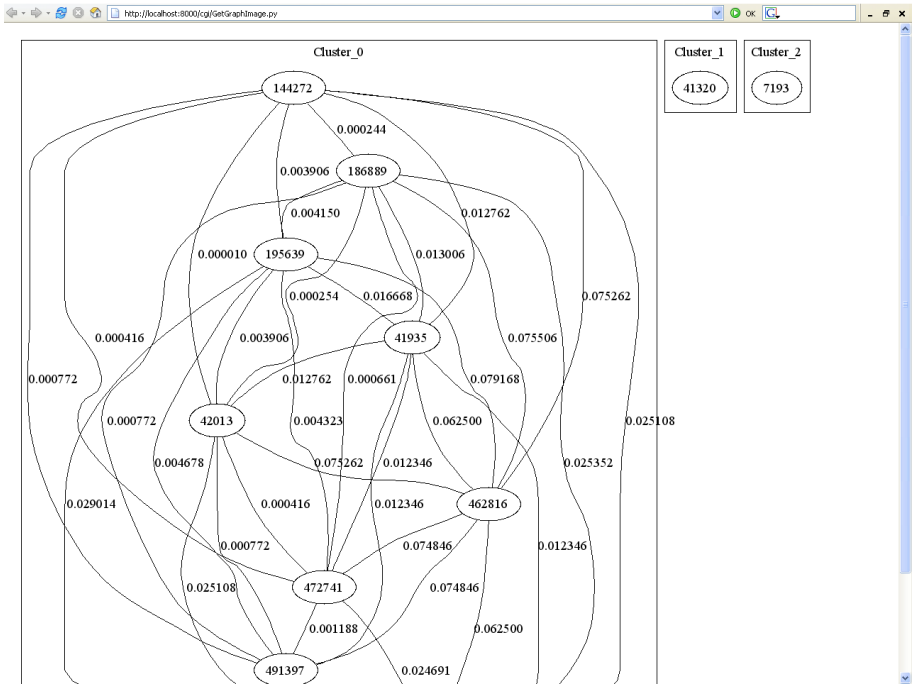

FIG. 6.6 – Regroupement thématique des références

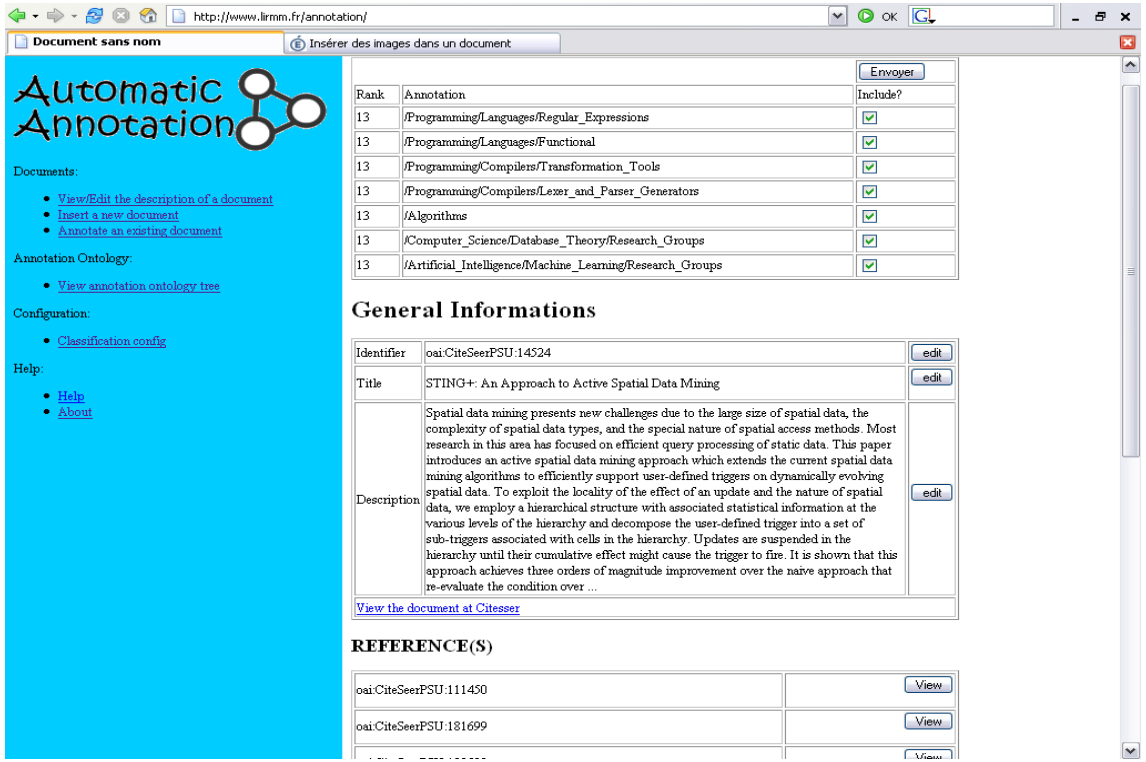

FIG. 6.7 – Résultat de l'annotation par RAS

7

Constitution du corpus de tests et évaluation

Ce chapitre, présente la partie expérimentale de notre travail. Nous décrivons la série de tests utilisée dans nos expérimentations. Nous analysons ensuite les résultats obtenus à partir de notre approche d'annotation.

Sommaire

7.1 Introduction 119
7.2 Descriptif du protocole d'expérimentation . . 119
7.3 Constitution du corpus de tests 120
7.4 Evaluation 124
7.5 Conclusion et discussion 132

7.1 Introduction

Nous décrivons dans ce chapitre l'expérimentation que nous avons menée au cours de cette thèse. Son objectif est l'évaluation de la qualité des résultats de l'annotation de notre approche (chapitre 4). L'annotation d'un document est réalisée par une propagation des annotations des documents cités, et ceci en regroupement thématiquement ces documents.

La taille de la base du SEMIDE n'étant encore importante, et afin de valider notre approche, nous avons choisi une base de taille importante de documents techniques qui s'inter-référencent. L'évaluation de notre approche s'est réalisée en deux phases :

- la construction du corpus de test ;
- l'évaluation du résultat de l'annotation sur le corpus.

La démarche adoptée pour chaque étape et les choix associés sont détaillés dans la suite de ce chapitre.

7.2 Descriptif du protocole d'expérimentation

La figure 7.1 illustre le protocole d'expérimentation :

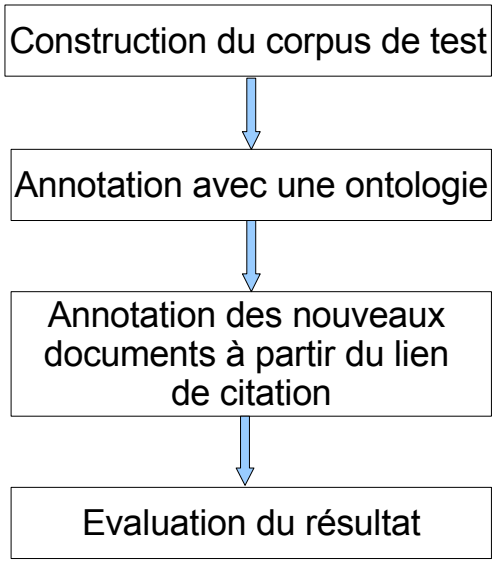

Fig. 7.1 – Le protocole d'expérimentation

Les différentes étapes de notre expérimentation sont les suivantes (figure 7.1) :

1. constitution d'un corpus qui répond à nos besoins, en l'occurrence des documents qui s'inter-référencent ;

2. utilisation d'une ontologie pour l'annotation des documents ;
3. annotation des nouveaux documents ;
4. évaluation du résultat de l'annotation.

L'évaluation du résultat de l'annotation des documents est basée sur deux méthodes : la comparaison avec une annotation existante, faite généralement par l'auteur du document et l'évaluation par des experts du domaine. La deuxième méthode est réalisée à partir d'un formulaire d'évaluation sur lequel nous reviendrons plus tard.

La suite de ce chapitre est organisée de la manière suivante :

– la section 7.3 présente la constitution de la base de tests, ici Citeseer, ainsi que l'ontologie utilisée ;
– les méthodes d'évaluation ainsi que les résultats obtenus sont décrits dans la section 7.4.

7.3 Constitution du corpus de tests

Rappelons que nos travaux s'inscrivent dans le contexte du projet euro-méditérannéen SEMIDE. L'annotation des documents constitue un problème majeur dans la mission de partage et de diffusion de l'information. D'après le constat que les documents sont généralement techniques et citent des documents de la base, nous avons choisi l'approche d'annotation en nous basant sur ces références.

La base du SEMIDE ne constitue pas pour le moment une taille suffisamment importante pour que l'on puisse tester notre approche. Pour cela notre choix s'est porté sur une base avec un nombre important de documents qui s'inter-référencent.

7.3.1 Collection Citeseer

Dans le cadre de nos expérimentations, nous avons choisi comme collection de tests la base de Citeseer[26]. Cette collection a été choisie pour deux raisons : la première est le nombre important de documents et par conséquent l'augmentation de la validité de nos tests, et la deuxième encore plus importante est l'inter-référencement des articles, ce qui convient exactement à nos expérimentations.

Citeseer [SCL+05],[GHN05] est une bibliothèque numérique sur la littérature scientifique. Son but est la propagation du savoir, la récupération et l'accès à la littérature scientifique [LBG99]. Citeseer localise les articles scientifiques sur le Web, extrait différentes informations tels que les citations et le titre des articles.

La bibliothèque nous a permis de construire une base de plus de 550 000 documents qui se référencent. La figure 7.2 illustre le schéma de la base :

[26] http ://citeseer.ist.psu.edu/

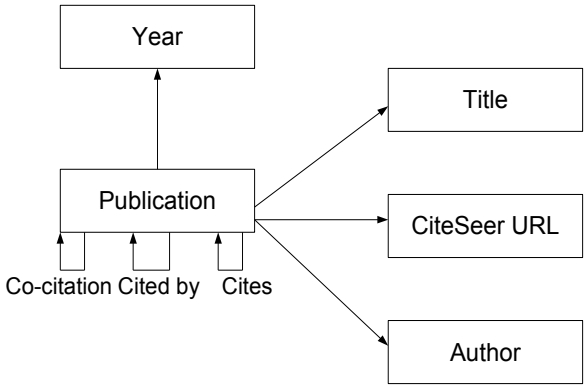

FIG. 7.2 – Le schéma de citeseer

- une publication a un titre, une URL et un ou plusieurs auteurs,
- une publication est citée par au moins une publication,
- une publication cite au moins une publication
- une publication est co-citée avec au moins une autre publication.

7.3.2 Construction de la base

A partir de toutes ces informations, nous avons construit notre base de tests. Au cours de notre phase d'expérimentation, nous avons travaillé sur une machine qui a les caractéristiques suivantes : un processeur Pentium 4 et une RAM de $512Mo$.

La base de données *MySQL* a une taille de $1G$ et avait le schéma :

- la table $Article(\underline{id}, titre, des)$: des métadonnées trouvées sur des articles de la base Citeseer, nous avons retenu l'identifiant de l'article, son titre et sa description ;
- la table $keywords(\underline{id}, keyword)$: après avoir constitué la description des articles, nous avons eu besoin d'une base de documents déjà annotés. Nous avons pu obtenir un ensemble d'articles indexés avec des mots clés (base de Citeseer). La table contient l'identifiant du document et le mot clé associé ;
- la table $Reference(\underline{source}, \underline{destination})$: *source* est l'identifiant du document qui référence le document qui a pour identifiant *destination* ;
- la table $Reference_par(\underline{source}, \underline{destination})$: *source* est l'identifiant du document qui est référencé par le document qui a pour identifiant *destination* ;.

7.3.3 Utilisation d'une ontologie

La propagation des annotations se fait à partir d'un ensemble de documents annotés avec des termes de l'ontologie. Dans notre corpus de tests, nous disposons de mots clés associés aux documents de la base.

Dans cette section, nous présentons l'ontologie Dmoz utilisée, ainsi que l'annotation avec celle-ci.

7.3.3.1 L'ontologie choisie

Dans la base de Citeseer, pour chaque document on ne retrouve que son titre et sa description. Dans notre approche, on ne s'intéresse qu'à un domaine particulier qui suit un vocabulaire contrôlé, et où l'on suppose que pour propager les annotations, nous disposons d'une base de références déjà annotées. Dans la suite, nous présentons l'ontologie Dmoz.

7.3.3.2 Dmoz

L'Open Directory Project connu sous l'acronyme Dmoz[27] pour Directory Mozilla, est un annuaire web créé en 1998, sous licence Netscape Open Directory License. Dmoz est géré par une équipe importante d'éditeurs volontaires, plus de 60 000 éditeurs ont participé au projet depuis son lancement. Il est librement utilisé par de nombreux sites et moteurs comme google[28]. En avril 2005, l'annuaire contenait plus de 4 millions d'adresses et de sites classés dans plus de 590 000 catégories. Pour annoter les documents nous avions besoin d'une ontologie que nous avons limitée au domaine de l'informatique *Computer*. Nous avons obtenu le fichier rdf[29] à partir de l'ODP[30] et extrait les concepts liés au domaine de l'informatique. Voici un extrait du fichier *rdf* récupéré.

```
<Topic r:id="Top/Computers">
  <tag catid="4"/>
  <d:Title>Computers</d:Title>
  <narrow r:resource="Top/Computers/Hacking"/>
  <narrow r:resource="Top/Computers/Graphics"/>
```

La figure 7.3 est un extrait de l'ontologie Dmoz. Un travail de filtrage a été nécessaire pour éliminer les termes tels que : *Conference2001*. Notre ontologie finale contient 3750 concepts.

7.3.3.3 Annotation avec Dmoz

Les premiers documents qui servent de base pour la propagation des annotations étaient indexés avec des mots clés libres. Après avoir récupéré l'ontologie

[27] http ://www.Dmoz.org/
[28] www.google.com
[29] http ://rdf.Dmoz.org/
[30] Open Directory Project

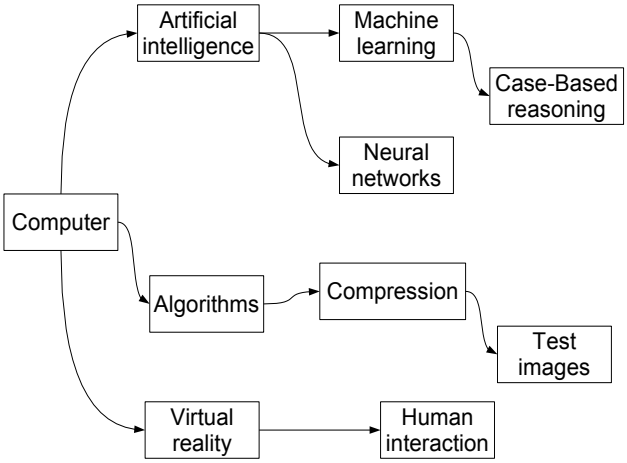

FIG. 7.3 – Extrait de l'ontologie de Dmoz

pour *Computer* et avoir parsé le fichier pour intégrer la hiérarchie de notre ontologie dans la base, nous avons fait un passage des mot clés à l'ontologie en associant chaque terme à un nœud de l'ontologie. Pour cela on utilise le moteur de recherche de Dmoz et le passage se fait comme suit :

1. on construit l'ensemble de tous les mots clés présents dans la base
2. pour chaque mot clé, on effectue la recherche sur le moteur de recherche de Dmoz à partir de l'adresse suivante :
   <http ://search.Dmoz.org/cgi-bin/search ?search=**mot_clé**&all=no &cat=**Computers**&t=b

Par exemple, on effectue une recherche pour les mots clés ***Robotics*** et ***classification***, on trouve ces deux résultats :

```
Computers: Robotics    (423 matches)
Computers: Artificial Intelligence: Machine Learning:    (8 matches)
```

Dans ce cas, on associe les mots clés aux concepts de l'ontologie et les documents sont annotés avec notre ontologie.

7.3.4 Calcul des indices de co-citations

Comme nous l'avons expliqué dans le chapitre 4, la propagation des annotations se fait avec un sous-ensemble de citations par regroupement thématique. La similarité thématique est basée sur la méthode de co-citations.

La méthode de co-citations a pour objectif de créer à partir d'articles scientifiques d'un même domaine de recherche, et à partir de leurs références bibliographiques, des cartes relationnelles de documents ou d'auteurs qui reflètent à la fois les liens sociologiques et thématiques de ce domaine [PCBL02].

A partir des tables ***Reference_par*** et ***Reference***, nous avons calculé les indices de co-citations de tous les documents de la base. Le calcul des indices de co-citations à l'apparition d'un nouveau document se fait en incrémentant l'indice de chaque couple de référence si ces deux documents ont déjà été cités ensemble ou bien en le mettant à un sinon. Le calcul se fait de manière **incrémentale** (algorithme 7)

Algorithme 7 : Calcul des indices de co-citations

Données :
d document à annoter
$x, y \in R = \{$*ensemble des références de* $d\}$ et $(x \neq y)$
C_{xy}=indice de co-citations des références x et y
début
  répéter
    si *(*$\exists C_{xy}$*)* **alors**
      $C_{xy} = C_{xy} + 1$
    sinon
      $C_{xy} = 1$
  jusqu'à *(Calcul de tous les* C_{xy}*)* ;
fin

7.4 Evaluation

Dans cette section, nous présentons les résultats de l'évaluation obtenus. Pour évaluer notre approche, deux ensembles d'expérimentations ont été effectués en se basant sur deux méthodes d'évaluation :

1. la première consiste à comparer l'annotation obtenue avec celle existante dans la base Citeseer ;
2. l'évaluation d'un petit nombre de documents lors de la première méthode a nécessité une deuxième méthode. Elle est basée sur l'avis d'experts à travers un formulaire d'évaluation.

3000 documents ayant des références déjà annotées ont été sélectionnés de manière aléatoire dans la base et annotés avec notre outil (cf. chapitre 6). Les deux méthodes d'évaluation ainsi que les résultats obtenus sont présentés dans la suite.

7.4.1 Evaluation basée sur la comparaison d'annotations

7.4.1.1 Description de la méthode

La première méthode d'évaluation consiste à comparer le résultat avec une annotation existante. Parmi les 3000 documents annotés avec notre outil, nous

avons sélectionné les documents déjà indexés dans la base Citeseer, généralement par l'auteur du document.

Les documents sélectionnés ici devaient répondre aux contraintes suivantes :
- les documents dans la base doivent être annotés afin de pouvoir les comparer avec le résultat de notre approche ;
- les documents doivent avoir au moins trois références annotées, et ceci afin de pouvoir appliquer notre approche de propagation d'annotation. Nous avons remarqué de façon empirique que moins de 3 références, la propagation n'est pas pertinente.

Le tableau 7.1 illustre le nombre de documents utilisés :
- le nombre de documents annotés est de 17417. Notre base contient 550000 documents, mais nous n'avons pas pu obtenir tous les mots clés associés aux documents. Les documents de Citeseer étant décrit avec un ensemble de mots clés, un passage à l'annotation avec l'ontologie a été effectué (section 7.3.3.3) ;
- afin de propager les annotations, nous avons sélectionné les documents ayant plus de 3 références annotées ;
- l'intersection des deux premiers ensembles de documents nous donne 66 documents. Cet ensemble sert de comparaison avec l'annotation générée par notre approche. La première méthode a donc été effectuée sur un petit corpus de documents.

Documents dans la base	550 000
Documents annotés par les experts	17 417
Documents ayant plus de 3 références annotées	940
Documents déjà annotés et ayant plus de 3 références annotées	66
Nombre moyen de termes annotant par document dans Citeseer	6,77

TAB. 7.1 – comparaison des annotations

7.4.1.2 Résultat

La comparaison des annotations avec les annotations récupérées sur quelques documents à partir de la base de Citeseer nous ont donné le résultat illustré sur la figure 7.4. Les annotations existantes sont réalisées généralement par les auteurs des documents. Afin de représenter le résultat, nous avons défini un indice de qualité :

$$I_q = \frac{A_c}{A_d} \in [0,1] \tag{7.1}$$

- A_c : nombre d'annotations correctes par document ;
- A_d : nombre d'annotations par document.

Concernant les 66 documents évalués nous obtenons un indice de qualité I_q qui est égale à 0,732 avec :

- une moyenne de 6,77 concepts par document ;
- une moyenne de 4,96 concepts de documents correspondant à l'annotation existante.

Chaque document est annoté avec une moyenne de 6,77 concepts. Notons qu'un index a été fourni par l'administrateur de la base Citeseer sous forme de mots clés associés à un ensemble de documents. Afin de pouvoir comparer avec le résultat de notre approche, nous avons transformé ces mots clés en une annotation en utilisant les concepts de Dmoz. Parmi les résultats de notre approche 4,96 correspondent à l'annotation existante, ce qui correspond à 73% d'annotations correctes (figure 7.4).

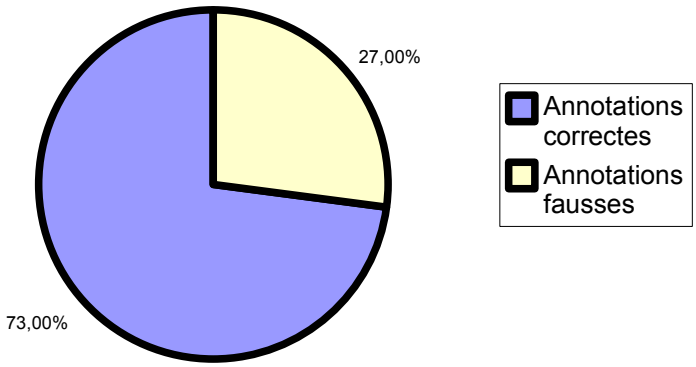

FIG. 7.4 – Résultat de la première méthode d'évaluation

Nous avons pu constater du bruit dans l'annotation générée par notre approche, et ceci est dû principalement aux raisons suivantes :

- la liste des termes qui nous a été fournie par l'administrateur de la base Citeseer n'est pas complète. Effectivement, plusieurs documents sont associés à moins de 4 termes ;
- le passage de mots clés vers l'ontologie Dmoz génère du bruit. Malgré la phase de filtrage que nous avons effectuée sur l'ensemble des concepts de l'ontologie Dmoz, nous avons quelques concepts qui n'appartiennent pas au domaine de travail *Science computer*.

Les résultats de la première méthode sont encourageants mais incomplets, d'une part à cause du petit nombre de documents 66 et d'autre part le nombre insuffisant de termes associés aux documents qui nous ont servis pour notre

étude comparative. Pour cette raison, nous complétons notre validation par une deuxième méthode d'évaluation basée sur l'avis d'experts.

7.4.2 Evaluation basée sur l'avis d'experts

Les résultats de la première méthode n'étant pas suffisants pour valider notre approche, une seconde méthode a été effectuée.

7.4.2.1 Description de la méthode

La deuxième méthode d'évaluation est basée sur l'évaluation des experts. Afin de d'évaluer notre approche, nous avons mis en ligne un formulaire d'évaluation des annotations obtenues par notre approche. Voici quelques questions du formulaire :

1. est-ce que l'annotation correspond au titre du document ?
2. est-ce que l'annotation correspond à la description du document ?
3. est-ce qu'il ya des termes correspondant parfaitement au document et d'autres pas du tout (incohérence) ?
4. avis général sur l'annotation (très satisfaisant, satisfaisant, peu satisfaisant, pas satisfaisant)

Le formulaire[31] est illustré dans la figure 7.5. Il est divisé en deux parties :
- une partie pour la visualisation du résultat de l'annotation, ainsi que des informations sur le document pour l'évaluateur. Il existe aussi la possibilité d'aller consulter le document directement sur le site Citeseer.
- une deuxième partie pour les questions concernant l'annotation.

Une sélection aléatoire est faite au début de l'évaluation. Nous avons créé dans notre base une table ***A_evaluer*** où on sélectionne le document à annoter pour l'évaluation et une autre table ***evaluer*** où on met les réponses des experts. Chaque fois qu'un document est évalué, il est inséré dans la table ***evaluer*** et supprimé de ***A_evaluer***. Ceci est fait pour évaluer le maximum de documents étant donné leur nombre important, et éviter d'évaluer toujours le même document. 30 experts ont participé à l'évaluation du résultat de l'annotation avec notre approche.

7.4.2.2 Calcul de l'échantillon minimal de test

Nous calculons dans cette section le nombre de documents nécessaires afin de valider notre résultat. Pour cela, nous calculons la taille minimale de l'échantillon utilisé pour l'évaluation.

Afin de déterminer le nombre de documents nécessaires pour calculer le pourcentage de documents bien annotés, il faut partir d'un pourcentage de départ sur un petit ensemble de documents (établi à partir d'une pré-étude).

[31] http ://www.lirmm.fr/annotation/evaluation/

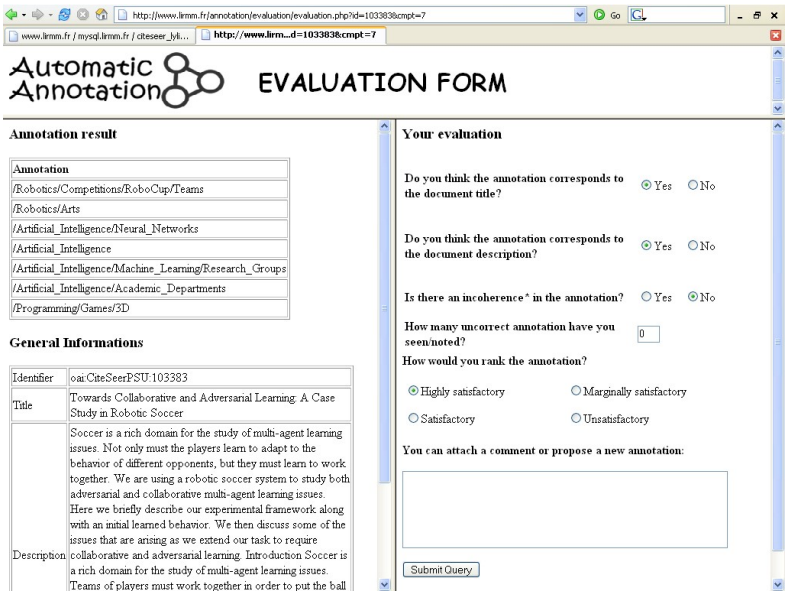

FIG. 7.5 – formulaire d'évaluation

La pré-étude a été effectuée sur 50 documents, nous avons obtenu le résultat suivant :

- 30 documents bien annotés (très satisfaisant, satisfaisant)
- 20 documents mal annotés (peu satisfaisant, pas satisfaisant)

Pour calculer la taille minimale de l'échantillon, il faut partir de la formule de l'intervalle de confiance d'un pourcentage, avec un risque de se tromper qui est : $\alpha = 5\%$. La formule est donnée par :

$$p_0 \pm 1.96\sqrt{\frac{p_0 q_0}{n}} \tag{7.2}$$

1. $1,96$ est un chiffre qui se lit sur la table de la loi normale centrée réduite, avec $\alpha = 5\%$;
2. p_0 représente le pourcentage de documents bien annotés lors de la pré-étude ;
3. q_0 représente le pourcentage de documents mal annotés lors de la pré-étude ;
4. n représente la taille minimale de l'échantillon.

La nombre de documents dans l'échantillon est calculé de la façon suivante :

$$n = \frac{(1.96)^2 p_0 q_0}{i^2} \tag{7.3}$$

i représente le degré de précision qui est égale à $1.96\sqrt{\frac{p_0 q_0}{n}}$. Généralement $i = \frac{p_0}{10}$.

Pour notre étude nous avons les résultats suivants : $p_0 = 0.6$, $q_0 = 1 - p_0 = 0.4$ et $i = 0.06$. Donc :

$$n = \frac{(1.96)^2(0.6)(0.4)}{(0.06)^2} = 256.1 \tag{7.4}$$

Il faut donc avoir au moins 256 documents évalués.

Une fois les résultats de l'évaluation obtenus, nous calculerons l'intervalle de confiance du pourcentage de documents bien annotés.

7.4.2.3 Résultat

En pratique, l'évaluation a été effectuée sur 322 documents. Les articles de Citeseer touchant plusieurs sous domaines, le choix des évaluateurs a été fait en utilisant différentes listes de diffusion de la communauté du domaine informatique.

Afin d'augmenter la pertinence de l'évaluation et éviter les incohérences, nous avons choisi d'évaluer la correspondance d'une annotation avec le titre et la description (résumé) d'un document, pour ensuite évaluer l'annotation de manière générale. Les termes qui sont dans le titre et le résumé sont les plus importants en terme de poids dans les méthodes d'indexation classique. Effectivement, un terme qui se trouve dans le titre et un autre en fin de document n'ont pas la même importance dans le document.

En ce qui concerne les deux premières questions portant sur le niveau de correspondance (très satisfaisant $\longrightarrow$ pas satisfaisant) entre les annotations et respectivement le titre et la description des document, l'évaluation a donné le résultat illustré sur le tableau 7.2.

	très sat	sat	peu sat	pas sat
titre	70	121	88	43
description	71	125	86	40

Tab. 7.2 – Résultat du formulaire d'évaluation

L'évaluation sur le résultat de l'annotation de manière générale est illustré sur la figure 7.6.

Dans un premier temps, on compare l'évaluation de l'annotation de façon générale avec celle du titre et de la description. Ceci est effectué dans le but de recherche d'incohérence (par exemple correspondance avec le titre et pas avec la description). La figure 7.7 illustre cette comparaison.

Nous remarquons sur la figure 7.7, que l'évaluation de la correspondance de l'annotation avec le titre et la description correspond avec celle de l'annotation de manière générale. Cependant, il existe quelques évaluations où il y a une incohérence et ceci dans deux cas :

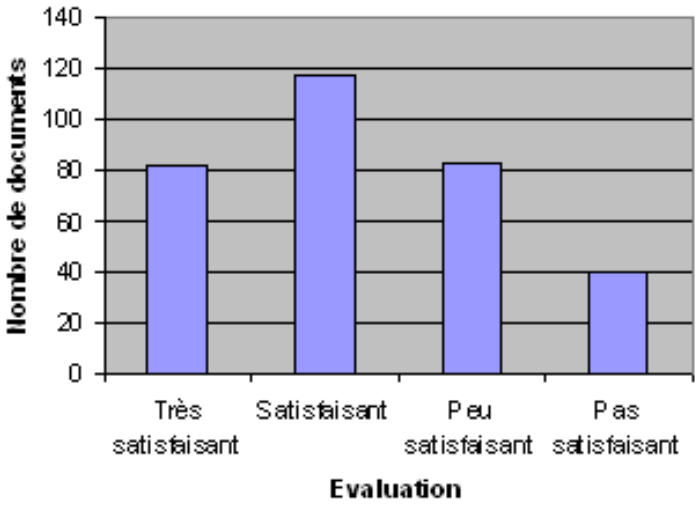

FIG. 7.6 – Avis sur l'annotation en général

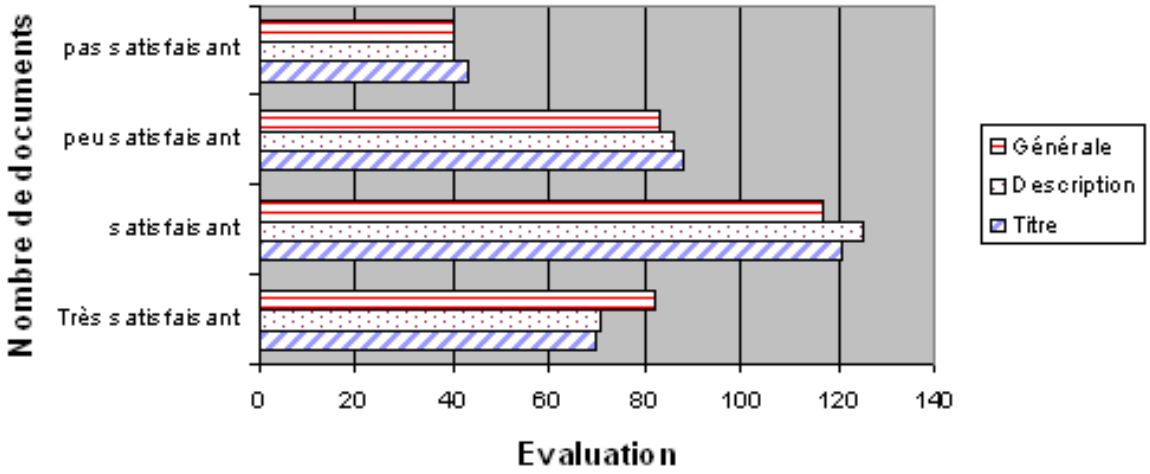

FIG. 7.7 – Comparaison du résultat de l'évaluation (titre, description, générale)

– le cas où on est satisfait de l'annotation par rapport au titre et résumé et pas tout à fait de l'annotation générale ;
– l'inverse, càd satisfait de l'annotation de manière générale et pas forcément du titre et du résumé.

Le premier cas est apparu lorsque le document est annoté avec un nombre important de concepts dont ceux qui correspondent aux termes apparaissant dans le titre et d'autres qui ne correspondent pas à la thématique du document.

Le second cas est apparu lorsque l'annotation est satisfaisante de manière générale mais pas assez spécialisé. Les termes apparaissant dans le titre sont

très spécialisés, le titre contenant les termes clés du document.

Nous estimons que l'évaluation est satisfaisante dans le cas de choix (très satisfaisant, satisfaisant). Sur les 322 documents évalués par les experts, 62% annotations générées par notre approche sont estimées satisfaisantes. Le résultat est validé sur un échantillon, nous calculons à présent l'intervalle de confiance de ce pourcentage. p_0 est égale à 0.62, la formule de l'intervalle de confiance d'un pourcentage est la suivante :

$$p_0 - 1.96\sqrt{\frac{p_0 q_0}{n}} < p < p_0 + 1.96\sqrt{\frac{p_0 q_0}{n}} \tag{7.5}$$

Nous obtenons comme résultat un pourcentage de documents annotés correctement p avec :

$$0.57 < p < 0.67$$

7.4.2.4 Analyse

Dans cette section, nous tentons d'analyser les mauvais résultats et ceci en nous basant sur les résultats et sur les commentaires saisis par les évaluateurs insatisfaits du résultat de l'annotation.

Voici un extrait des commentaires :

- "Les annotations sont en rapport mais de loin avec l'article".
- "Artificial intellgence oui mais je ne vois pas pourquoi il y a machine leanring alors que l article parle de natural language processing".
- "A voir le titre et la description de l'artile, l'annotation automatique obtenue est un peu trop large et ne cerne pas de près le sujet du texte".
- "Les annotations sont dans le domaine, mais aucune ne précise le sujet précis à savoir la compression de vidéos".
- "une seule annotation et elle ne correspond pas au document".
- "quelques bruits dans les annotations".
- "le mot partitioning semble avoir été pris dans le sens musical, causant une incohérence avec l'annotation".

Nous avons catégorisé les commentaires par critère qui sont les suivants :

1. le nombre des concepts de l'annotation est insuffisant ;
2. le dernier niveau des concepts est trop spécialisé ;
3. les concepts spécifiques au document n'apparaissent pas (annotation correcte mais pas assez spécialisée) ;
4. incohérence entre les concepts, dû au passage de mots clés vers l'ontologie Dmoz.

L'annotation d'un document est basée sur des annotations existantes des documents cités, le regroupement de ces références est complètement indépendant de leurs annotations étant donné que ce regroupement est basé sur les co-citations.

Cette propagation nécessite alors une base déjà annotée, ce qui n'est pas toujours le cas. Quelques documents évalués avaient un petit nombre de références annotées, ce qui a généré dans certains cas une annotation se composant de un ou deux concepts et dans ce cas insuffisants.

La deuxième raison de la mauvaise annotation est la propagation des concepts de références spécifiques au document référencé. Prenons par exemple une référence qui traite du SGBD "MySQL", ce document est une définition des bases de données. Ne faisant pas la distinction entre la nature des références (dans une définition ou dans le corps du document), nous nous basons sur la proximité thématique des documents, dans ce nous propageons le concept "MySQL" qui représente un niveau trop spécifique pour le document à annoter. Une solution serait de propager un niveau plus haut dans la hiérarchie du concept (ex : SGBD).

Le contraire peut générer une annotation insuffisante en n'ayant pas de concepts spécifiques au document à annoter. En propageant des références proches thématiquement, qui sont dans le même domaine que le document à annoter mais qui ne traitent pas forcément du même sujet (de façon détaillée), génère une annotation correcte et cohérente mais pas assez fine (spécifique).

Enfin, lors de nos expérimentations, quelques incohérences sont apparues et ceci lors du passage de mots clés vers l'ontologie Dmoz. Un traitement manuel de filtrage a été effectué afin d'éliminer les concepts n'appartenant au domaine.

7.5 Conclusion et discussion

Dans ce chapitre, nous avons présenté l'expérimentation que nous avons menée au cours de notre travail en décrivant en premier lieu notre protocole d'expérimentation. Notre expérimentation s'est déroulée en deux étapes :

- la constitution du corpus de tests pour l'annotation ;
- l'évaluation de l'annotation.

D'un point de vue pratique, nous avons implémenté cette approche et procédé ensuite à l'évaluation sur un corpus de la base de test, ici Citeseer. Nous avons présenté deux méthodes d'évaluation, la première basée sur la comparaison d'autres annotations effectuées par les auteurs des documents, et la deuxième sur l'avis d'experts à travers un formulaire d'évaluation.

Les résultats obtenus pour les deux méthodes sont trés encourageants. Effectivement, en ce qui concerne la première méthode, 73% des annotations sont correctes. En ce qui concerne la deuxième méthode, les experts estiment 62%

des annotations sont satisfaisantes ou très satisfaisantes.

L'annotation des documents, en se basant sur le contexte de citation, est une approche très intéressante spécialement dans des systèmes où le contenu des documents n'est pas fourni généralement pour des raisons de confidentialité ou des raisons commerciales. Cependant, on peut noter que notre approche peut être améliorée afin d'avoir de meilleurs résultats. Pour cela l'annotation a besoin d'être affinée. Nous proposons deux pistes dans ce sens :

- utiliser d'autres éléments du document afin de spécialiser l'annotation d'une part, et de créer un "champs d'annotation" d'autre part ;
- utiliser la structure du document afin de pondérer la propagation des annotations.

Une première approche pour éliminer le bruit serait d'utiliser quelques éléments du document à annoter, par exemple le titre du document afin d'extraire des termes qui serviront de filtre aux concepts propagés. Ceci peut être effectué en définissant une distance entre les concepts de l'ontologie qui peut être définie par une distance ultramétrique entre deux concepts. Il s'agit du chemin minimal à parcourir dans l'arbre des concepts pour aller d'un concept à un autre.

Cependant, les termes extraits du titre n'appartiennent pas forcément à l'ontologie du domaine. Ceci peut être résolu par une correspondance entre le titre et les concepts de l'ontologie, mais qui génère dans certains cas des erreurs ou des incohérences, comme nous l'avons vu pour le cas du passage à l'ontologie Dmoz et rejoindrait le processus d'enrichissement d'ontologie. Des termes importants dans les documents sont proposés comme termes candidats pour l'enrichissement de l'ontologie.

La deuxième solution est d'utiliser la structure du document afin de restreindre le niveau de spécialisation d'un concept propagé. La figure 7.8 illustre la structure d'un document.

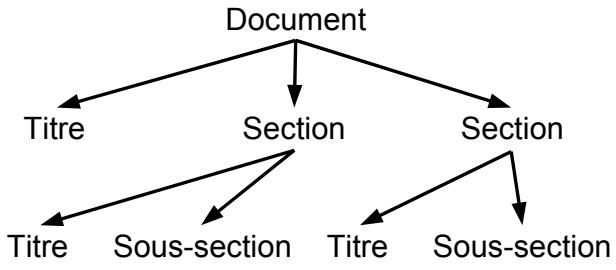

Fig. 7.8 – Structure d'un document

L'exploitation de cette structure est intéressante dans le sens où le niveau de propagation des annotations ne serait pas le même d'une section

à une autre. Par exemple pour une référence qui apparaît au niveau d'une sous section, on ne prendrait que le niveau le plus haut. Par exemple dans le concept `/Artificial_Intelligence/Agents/Tools`, on ne prendrait que `/Artificial_Intelligence`. La structure du document serait un critère de plus lors de l'importation des annotations.

Conclusion et perspectives

8

Conclusion

8.1 Synthèse

Les travaux présentés dans cette thèse se situent dans le contexte de la recherche d'informations en utilisant les technologies du Web sémantique.

L'eau devient un enjeu national et international de plus en plus important, la gestion de l'eau nécessite des connaissances multiples qui proviennent de plusieurs domaines. Le partage et l'échange d'informations dans le domaine de l'eau est la raison d'être du SEMIDE qui vise à faciliter l'accès et la mise en commun de l'information existante. Tout au long de cette thèse, nous montrons la nécessité de l'annotation de documents pour la recherche et la découverte de ressources, et l'utilisation de l'ontologie pour la mise en commun.

Dans ce travail de thèse, nous abordons cette problématique, en introduisant tout d'abord dans le chapitre 2 deux approches d'annotation que nous avons retenues, la première se basant sur le contenu des documents et la deuxième sur le contexte représentant les relations inter-documents. La première consiste à annoter un document à partir de son contenu. Nous avons montré dans cette partie l'apport de l'utilisation d'une ontologie sur deux niveaux :

- l'indexation des documents en utilisant un vocabulaire contrôlé permet le limiter le bruit lors de la recherche ;
- la recherche, où les relations inter-concepts sont utilisées afin de l'améliorer.

Parce que n'utiliser que le contenu d'un document pour l'annoter peut générer un manque, et surtout n'est pas toujours possible (car on n'a pas toujours le contenu), nous avons présenté le deuxième type d'annotation basé sur le contexte. Nous nous sommes intéressés aux travaux traitant du contexte de citation/référencement, notre approche se basant sur les citations. L'étude de travaux traitant de l'utilisation des citations nous a motivé, à prendre en considération la pertinence des références dans un document. Etant dans un contexte technique, spécialisé et multilingue, cette annotation nécessite l'utilisation d'une ontologie qui est définie par les acteurs du domaine. Cette on-

tologie doit évoluer en même temps que la masse d'informations et en tenant compte du besoin des utilisateurs.

Une fois ce constat mis en évidence, nous présentons dans le chapitre 3 quelques approches de construction et d'enrichissement d'ontologies, nous situons notre définition de l'ontologie par rapport aux autres vocabulaires contrôlés. Sont décrits dans ce chapitre, dans un premier temps, des méthodologies et des outils classiques. Compte tenu du nombre important de travaux s'intéressant à cette thématique, nous avons retenu dans un second temps quelques travaux se rapprochant le plus de nos travaux. Cet état de l'art nous a amenés à la conclusion qu'il n'existait pas de méthode générale pour l'enrichissement d'une ontologie. En effet, l'enrichissement d'une ontologie dépend du domaine d'application et des besoins d'une communauté. Les approches présentées passent principalement par une phase de constitution de corpus et d'analyse linguistique pour la détermination des termes du domaine. Cette étape n'est pas évidente dans le sens où il est difficile de constituer un corpus pour décrire un domaine, ce dernier étant aussi souvent en évolution et par cela dynamique.

A partir de nos besoins et des limites des travaux présentés, une nouvelle approche d'annotation basée sur le contexte de citation et utilisant l'ontologie du domaine est présentée dans le chapitre 4. Elle consiste à propager les annotations existantes de références sur le document citant. Les citations utilisées sont regroupées par une classification floue non supervisée. Nous positionnons dans un premier lieu notre approche par rapport aux manques des approches décrites dans l'état de l'art. Les différentes étapes ainsi que nos choix sont détaillés dans la seconde et principale partie de ce chapitre. Les étapes de l'approche sont illustrées par de petits exemples afin de faciliter la compréhension des différentes étapes.

L'ontologie utilisée pour l'annotation étant en continuelle évolution, notre approche est complétée dans le chapitre 5 par une deuxième partie traitant de l'enrichissement de l'ontologie. Le processus d'enrichissement est basé sur les sessions de recherche effectuées par les utilisateurs. Les sessions de recherche sont construites et représentées par des treillis de concepts afin d'extraire les nouveaux termes dans l'ontologie. Ceci est complété par une analyse linguistique réalisée à l'aide du système SYGFRAN. Ce chapitre est complété par l'algorithme de mise à jour d'annotation lors de l'ajout de concepts dans l'ontologie.

L'approche d'annotation a été implémentée pour ensuite être validée par différentes méthodes d'évaluation.

Nous mettons en évidence l'intérêt de l'approche d'annotation de documents par des expérimentations menées sur une base de documents scientifiques. La base du SEMIDE ne constituant pas pour le moment une taille suffisamment importante afin d'évaluer notre approche, nous avons effectué nos expérimentations sur la base Citeseer, d'un côté pour le nombre important de documents (550 000) et d'un autre côté pour l'inter-référencement des documents.

8.2 Résumé des principales contributions

8.2.1 Annotation de documents

1. **Considérer uniquement le contexte d'un document :** lorsqu'un auteur cite un autre document c'est qu'il estime que ce dernier est important pour la compréhension du contenu. Nous avons montré dans ce travail de thèse l'importance du contexte d'un document et plus particulièrement celui de citation dans le processus d'annotation. Ce travail a été effectué dans le contexte du SEMIDE où on ne dispose pas forcément du contenu des documents mais uniquement un ensemble de métadonnées. A l'issu de cela, nous avons développé une approche complètement indépendante du contenu.

2. **Utilisation des citations :** tout le contexte de citation n'étant pas toujours pertinent, l'annotation d'un document en propageant les annotations des citations, nous a poussé à ne sélectionner que les références pertinentes. Nous avons défini dans ce travail une approche pour la sélection de références pertinentes dans un document par un regroupement thématique.

3. **Approche pour le regroupement :** basée sur l'hypothèse de co-citations qui consiste à rapprocher thématiquement deux documents qui sont souvent cités ensemble, nous avons utilisé des algorithmes issus de la fouille de données ainsi que des approches de classification non supervisée afin de regrouper des documents proches. Ceci nous a amenés à choisir un algorithme autorisant la multi-classification partant du constat qu'un document peut traiter de plusieurs thèmes.

4. **Outil d'annotation :** partant de notre approche d'annotation, nous avons implémenté un outil d'annotation se basant uniquement sur les citations. Les résultats de l'annotation ont été évalués sur la base de documents techniques Citeseer.

8.2.2 Ontologies

1. **Utilisation d'une ontologie pour l'annotation :** nous avons montré dans ce travail l'apport de l'utilisation d'une ontologie dans le processus d'annotation et de recherche d'informations. La description d'un document avec des termes libres ne contient de lien sémantique entre les termes.

2. **Enrichissement de l'ontologie :** une approche pour l'enrichissement de l'ontologie est proposée dans cette thèse. Elle ne nécessite pas une

constitution d'un corpus de base et est basée sur les besoins des acteurs du système. La représentation du treillis de concept de concept est utilisée.

8.2.3 Communauté du SEMIDE

Notre travail de thèse a été effectué dans un contexte bien défini et selon le besoin d'une communauté spécialisée. Dans un domaine aussi important que celui de l'eau, une collaboration entre les acteurs de différents pays est nécessaire, incluant ainsi, le partage de l'information. C'est dans cette optique que le SEMIDE a été crée avec des objectifs générant quelques problèmes que nous avons essayé de résoudre. Pour le SEMIDE nos contributions ont été :

1. **Faciliter l'accès à l'information existante :** en utilisant les technologies du Web sémantique et ainsi répondre au problème de la difficulté de l'exploitation des ressources en utilisant des informations structurées.

2. **Dispersion de l'information :** pour des questions de confidentialité, les ressources sont souvent chez le fournisseur, lequel ne met à disposition qu'un ensemble de métadonnées. Nous avons répondu à cette contrainte en développant une approche d'annotation indépendante du contenu.

3. **Développer une ontologie spécifique aux connaissances dans le domaine de l'eau :** ce travail qui est basé sur une ontologie existante a consisté à l'enrichir, tout en respectant les besoins de la communauté.

8.3 Perspectives

Les perspectives de nos travaux sont nombreuses et portent principalement sur quatre volets.

Utilisation de la structure pour l'annotation la propagation d'annotation à partir des références a démontré sa pertinence à travers l'évaluation effectuée au cours notre travail. Une perspective intéressante à cette approche serait d'utiliser la structure du document (i.e. section, paragraphes ...) afin de délimiter l'impact d'une référence sur un document. Cependant, sur les documents existants ce serait un travail trop long à faire surtout en étant dans un contexte où le contenu des documents n'est pas fourni. Ceci est facilement envisageable sur des documents XML où le contenu est clairement séparé de la forme. Le principe ici serait de propager les annotations sur les différentes parties du document, et par cela faire une association entre les concepts de l'ontologie et la structure du document. Ceci améliorerait le processus de recherche dans le sens où, la réponse à un terme, cible bien une partie d'un document et non pas tout le document.

Similarité thématique des documents Afin de regrouper les références des documents à annoter, nous nous sommes basés sur la méthode de co-citations qui rapproche thématiquement des documents qui sont souvent cités par les mêmes documents. D'autres mesures de similarité pourraient être testées :

- rapprocher les documents qui répondent aux mêmes requêtes ;
- comparer les annotations des documents et calculer une similarité thématique par rapport à la proximité des termes dans l'ontologie ;
- utiliser le principe que si deux documents citent un document rarement cité inclut qu'ils partagent la même thématique.

Utilisations d'autres algorithmes de classification Une perspective à court terme est d'étudier davantage l'impact du choix des algorithmes de classification. Nous projetons d'une part d'améliorer les approches de fouille de données en utilisant les *itemsets* flous pour construire un *clustering* permettant le recouvrement de plusieurs classes. Nous projetons d'autre part de comparer l'ensemble des méthodes de classification (et de classification floue en particulier). Pour accomplir cette comparaison, nous souhaitons proposer des mesures permettant d'estimer la qualité de la classification en vue d'une tâche d'annotation.

Ontologie par profil d'utilisateur l'utilisation de la structure du treillis afin de représenter les sessions de recherche nous a permis d'extraire des termes servant à l'enrichissement de l'ontologie. Une classification des treillis pourrait servir à identifier des catégories d'utilisateurs et par cela avoir une ontologie associée aux niveaux des acteurs du système, ce qui améliorerait la recherche d'informations.

Table des figures

1.1 Organisation du semide 4
1.2 organisation des documents événements 8
1.3 Méthodologie 9

2.1 Processus de RI 15
2.2 Indexation classique vs indexation sémantique 17
2.3 Processus d'indexation dans [DJ02] 18
2.4 Le graphe de citation 24
2.5 Les relations entre les documents 24
2.6 Graphe de couplage 25
2.7 Graphe de co-citations 27
2.8 Propagation de métadonnées (Marchiori) 29
2.9 Dendogramme de la méthode de classification 30
2.10 Propagation avant et arrière 31
2.11 Propagation selon Marchiori 34

3.1 Les types d'ontologies, extrait de [Gua98] 43
3.2 Les différentes composantes de la BCT, extrait de [CR97] 47
3.3 Méthode de construction de Latiri [LMB05] 53
3.4 Enrichissement d'ontologie d'après [PGF04] 56

4.1 Les différentes étapes de la propagation 65
4.2 Extrait du graphe de citation des documents 68
4.3 Le graphe de co-citations 68
4.4 graphe de distance 74
4.5 graphe de distance final 75
4.6 Résultat de la classification ascendante 76
4.7 Résultat de l'algorithme k-means 77
4.8 Le résultat du regroupement des références 79
4.9 Annotation des documents cibles 81

5.1 Extrait de l'ontologie du Semide 90
5.2 Enrichissement à partir de la phase d'interrogation 91
5.3 Treillis de Galois 95
5.4 La hiérarchie des concepts 95
5.5 Règle sur le groupe prépositionnel 98

5.6 Règle sur le groupe adjectival . . . 99
5.7 Exemple d'analyse de SYGFRAN . . . 100
5.8 Les étapes de l'approche retenue . . . 101

6.1 Etapes d'annotation dans RAS . . . 110
6.2 Sélection d'un document pour visualisation . . . 111
6.3 Visualisation d'un document . . . 111
6.4 Ontologie dmoz utilisée pour l'annotation . . . 113
6.5 Graphe de distance avec RAS . . . 114
6.6 Regroupement thématique des références . . . 114
6.7 Résultat de l'annotation par RAS . . . 115

7.1 Le protocole d'expérimentation . . . 119
7.2 Le schéma de citeseer . . . 121
7.3 Extrait de l'ontologie de Dmoz . . . 123
7.4 Résultat de la première méthode d'évaluation . . . 126
7.5 formulaire d'évaluation . . . 128
7.6 Avis sur l'annotation en général . . . 130
7.7 Comparaison du résultat de l'évaluation (titre, description, générale) . . . 130
7.8 Structure d'un document . . . 133

A.1 modélisation entité/association du corpus de test . . . 147

C.1 Utilisation des métadonnées dans le Web sémantique, extrait de [CLR03] . . . 164
C.2 Quelques métadonnées du type événement . . . 165
C.3 Cas pratique de recherche de documents . . . 165

Annexes

A

Schéma de la base de données de documents test

A.1 Modèlisation EA du corpus

La figure A.1 illustre la modélisation entité/association du corpus de test.

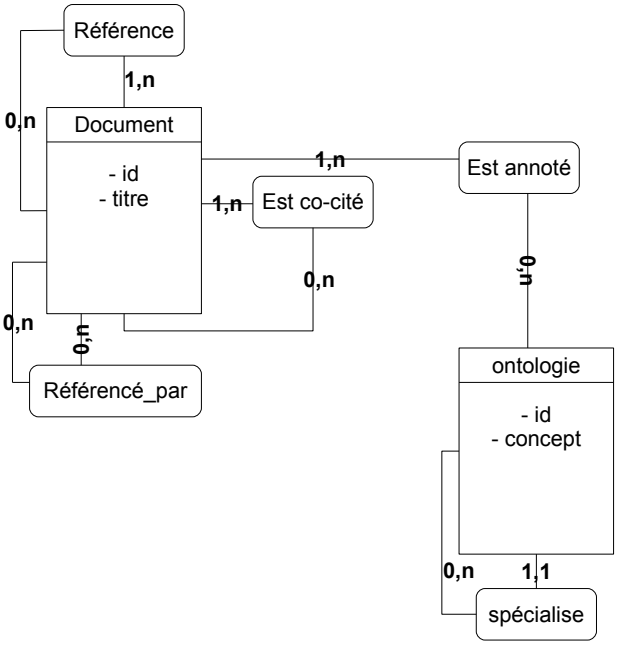

FIG. A.1 – modélisation entité/association du corpus de test

– un document peut référencer un ou plusieurs documents ;
– un document peut être référencé par un ou plusieurs documents ;

– un document peut être co-cité avec d'autres documents ;
– un document est annoté avec les concepts de l'ontologie ;
– un concept dans l'ontologie peut spécialiser un ou plusieurs concepts.

A.2 Schéma relationnel

Cette section présente les tables obtenues à la transformation du schéma E/A en modèle relationnel.

A.3 Structure de la table *article*

Cette table regroupe les différents articles ainsi que les informations les concernant, on retrouve l'identifiant, le titre et le résumé du document.

TAB. A.1: Structure de la table article

Champ	Type	Null	Défaut
ID	varchar(30)	Non	0
TITLE	text	Non	
DESCR	text	Non	

A.4 Structure de la table *reference*

Un document *Source* référence un document *Cible*, avec la table *Reference_par*, cette table nous sert à calculer les indices de co-citations.

TAB. A.2: Structure de la table reference

Champ	Type	Null	Défaut
SOURCE	varchar(30)	Non	0
DESTINATION	varchar(30)	Non	0

A.5 Structure de la table *reference_par*

Un document *Source* est Référencé par un document *Destination*.

TAB. A.3: Structure de la table reference_par

Champ	Type	Null	Défaut
SOURCE	char(30)	Non	0
DESTINATION	char(30)	Non	0

A.6 Structure de la table *indicecocitation*

La table *indicecocitation* contient les degrés de co-citations entre les documents de la base. Elle est calculée de manière incrémentale.

Tab. A.4: Structure de la table indicecocitation

Champ	Type	Null	Défaut
i	int(11)	Non	0
j	int(11)	Non	0
frequence	int(11)	Non	0

A.7 Structure de la table *ontologie*

Cette table regroupe tous les concepts de l'ontologie.

Tab. A.5: Structure de la table ontologie

Champ	Type	Null	Défaut
id	int(11)	Non	
concept	text	Non	

A.8 Structure de la table *ontologie_is_a*

Dans cette table, nous avons deux champs, le concept ***Parent*** est une généralisation du concept ***Child***.

Tab. A.6: Structure de la table ontologie_is_a

Champ	Type	Null	Défaut
child	int(11)	Non	0
parent	int(11)	Non	0

A.9 Structure de la table *article_ontology_annotation*

Cette table regroupe les correspondance entre les articles et leurs annotations (concept dans l'ontologie)

Tab. A.7: Structure de la table article_ontology_annotation

Champ	Type	Null	Défaut
article	int(11)	Non	0
annotation	int(11)	Non	0

A.10 Structure de la table *resultat*

Cette table regroupe les évaluations des experts à partir du formulaire d'évaluation. Elle regroupe les champs suivants :

- *id_doc* : l'identifiant du document évalué ;
- *cor_titre* : l'évaluation sur la correspondance de l'annotation avec le titre du document ;
- *cor_desc* : l'évaluation sur la correspondance de l'annotation avec la description du document ;
- *incoherence* : avis s'il existe une incohérence entre les concepts de l'annotation ;
- *general* : l'évaluation sur l'annotation de façon générale (trés satisfaisante, satisfaisante,...)
- *nbre_incorrect* : nombre de concepts incorrects dans l'annotation ;
- *nbre_annotation* : nombre de concepts généré par notre outil d'annotation ;
- *comment* : commentaire de l'évaluateur.

Tab. A.8: Structure de la table resultat

Champ	Type	Null	Défaut
id_doc	varchar(100)	Non	
cor_titre	varchar(10)	Non	
cor_desc	varchar(10)	Non	
incoherence	varchar(10)	Non	
general	varchar(200)	Non	
nbre_incorrect	int(11)	Non	0
nbre_annotation	int(11)	Non	0
comment	varchar(255)	Non	

B

Extraits de scripts RAS (Python ™)

B.1 Calcul des annotations d'un document

Extrait du fichier de script ras.py

```
1  #
       ************************************************************************
2  # * RAS Library
3  # * module name: ras.py
4  # * Author: L. Abrouk
5  # * Description: exporte la classe AutomaticAnnotator pour annoter
6  # * automatiquement un document de la base en utilisant un regroupement
7  # * thematique de ses references.
8  #
       ************************************************************************/
9
10 ##"""
11 #Attention : pour rendre le code plus lisible, toutes les classes et
       fonctions utilitaires
12 #ne sont pas affichees dans ce script. Vous pouvez vous referer au code
       source
13 #pour une version complete de ce fichier.
14 ##"""
15
16 from Cmeans import *
17
18 ...
19
20 ##"""
21 #Classe principale pour calculer automatiquement les annotations d'un
       document
22 #par regroupement thematique des references.
23 #
24 #Cette classe s utilise comme suit:
25 #> t = AutomaticAnnotator(<nom de la base de donnees>)
26 #> annotation_list = t.getAnnotation(<identifiant du document a annoter
       >)
```

```
27 #
28 ##"""
29 class AutomaticAnnotator:
30     def __init__(self,basename,num_cluster=3,power=3,epsilon=0.05,
       maxstep=10):
31         self.matrix_builder = CitationMatrixBuilder(basename)
32         self.num_cluster = num_cluster
33         self.power = power
34         self.epsilon = epsilon
35         self.maxstep = maxstep
36         self.resolver = {}
37         self.citation_graph = 0
38         self.clustering_matrix = 0
39         self.clusters = {}
40     ##"""
41     #Retroune les annotations du document doc a partir de la base de
       donnees
42     ##"""
43     def getDefinedDocumentAnnotation(self,doc):
44         return self.matrix_builder.getDefinedDocumentAnnotation(doc)
45 
46     ##"""
47     #Retourne a partir de la matrice d'appartenance (matrix) l ensemble
       des
48     #clusters avec les documents contenus dans chaque cluster.
49     #Le resultat est une table { cluster->{documents} }
50     ##"""
51     def getClusters(self,matrix):
52         num_lines = len(matrix)
53         num_cols = len(matrix[0])
54         result = {}
55         for c in range(num_cols):
56             result[c] = []
57 
58         for i in range(num_lines):
59             result_document = matrix[i]
60             num_cols = len(matrix[i])
61             max_prob = 0
62             max_cluster = 0
63             for j in range(num_cols):
64                 if(matrix[i][j]>max_prob):
65                     max_prob = matrix[i][j]
66                     max_cluster = j
67 
68             result[max_cluster].append(i)
69         return result
70     #"""
71     #Calcule et retourne les annotations correspondant a chaque cluster
72     #a partir de la matrice d appartenance passee en parametre :
       clusters
73     #"""
74     def getClustersAnnotation(self,clusters):
75         num_cols = len(clusters)
76         result = {}
77         for c in range(num_cols):
78             result[c] = []
79         for c in clusters:
80             doclst = clusters[c]
81             cluster_annotation = []
```

```
82              for i in doclst:
83                  #print "looking for the annotation of this document %i"%
                        self.resolver[i]
84                  annotation = self.getDefinedDocumentAnnotation(self.
                        resolver[i])
85                  for n in annotation:
86                      #print "got this annotation %s"%n
87                      if (n not in cluster_annotation):
88                          cluster_annotation.append(n)
89
90              result[c].append(cluster_annotation)
91
92          return result
93
94      #"""
95      #Cette fonction retourne les annotations d'un document
96      #@param documentid: identifiant du document dans la base.
97      #"""
98      def getAnnotation(self,documentid):
99          #""" Cluster est une table de correspondance #"""""
100         cluster_annotation = {}
101         #""" Recuperer la matrice de co-citation a partir de la base
102         #concernant le documentid
103         #"""
104         res = self.matrix_builder.getCoCitationMatrix(documentid)
105         #""" Sauvegarde des resultats dans des variables internes
106         self.citation_graph : graphe de citations de documentid
107         self.resolver : table de correspondance
108         #"""
109         self.citation_graph = res[0]
110         self.resolver = res[1]
111         #""" ----- #"""
112
113         dim    = len(self.citation_graph)
114         if(dim == 0):
115             #"""ensemble de references vide pour ce document#"""
116             return cluster_annotation
117
118         #""" ici l ensemble des references n est pas vide !#"""
119
120         #""" Transformer le graphe de co-citation pour avoir une
                distance Euclidienne#"""
121         self.matrix_builder.transformIntoEucledienneMatrix(self.
                citation_graph)
122         #""" Appliquer la fonction fuzzyCmeans #"""
123         result = fuzzyCmeans(self.citation_graph,self.num_cluster,dim,
                self.power,self.epsilon,self.maxstep)
124         #""" result est un couple:
125         #    - ( matrice de degre d'appartenance au cluster
126         #       ,
127         #       matrice des centroides)
128         #"""
129         matrix = result[0]
130         centroids = result[1]
131         #"""Assigner la reference interne de la matrice d appartenance
                #"""
132         self.clustering_matrix = matrix
133         #""" Recuperer l ensemble des clusters a partir de la matrice d'
                appartenance#"""
```

```
134         self.clusters = self.getClusters(matrix)
135         #""" Avoir les annotations d'un cluster#"""
136         cluster_annotation = self.getClustersAnnotation(self.clusters)
137         #""" Preparer la liste des annotations a retourner#"""
138         toret = []
139         for cluster in cluster_annotation:
140             cluster_rank = len(self.clusters[cluster])
141             for annotation in cluster_annotation[cluster]:
142                 for y in annotation:
143                     item = AnnotationItem(y,cluster_rank)
144                     toret.append(item)
145 
146         toret.sort(reverse=True)
147         #""" Retourner les annotations #"""
148         return toret
149 
150 class AnnotationItem:
151     def __init__(self,annotation,rank):
152         self.annotation = annotation
153         self.rank = rank
154 
155     def __eq__(self,other):
156         return self.rank == other.rank
157 
158     def __lt__(self,other):
159         return self.rank < other.rank
```

B.2 Regroupement thématique par l'algorithme Cmeans

Extrait du fichier de script cmeans.py

```
#
    ***************************************************************************

# * RAS Library
# * module name: Cmeans.py
# * Author: L. Abrouk
# * Description: exporte la fonction qui implemente l algorithme
# * fuzzy cmeans.
#
    ***************************************************************************/


#"""
#Attention : pour rendre le code plus lisible, toutes les classes et
    fonctions utilitaires
#ne sont pas affichees dans ce script. Vous pouvez vous referer au code
    source
#pour une version complete de ce fichier.
#"""

from random import *

#"""
#Instance generatrice aleatoire
#"""
rand = Random()

#"""
#Fonction util. pour copier une matrice
#"""
def copyMatrix(oldmatrix):
    newmatrix = []
    for i in range(len(oldmatrix)):
        newline = []
        for j in range(len(oldmatrix[i])):
            newline.append(oldmatrix[i][j])

        newmatrix.append(newline)

    return newmatrix

#"""
#Difference/distance entre deux matrices en utilisant la formule
#definie par le Cmeans (fonction 4.7, Chapitre 4 de la these)
#"""
def diffMatrix(matrix1,matrix2):
    diff = 0
    dim_i = len(matrix1)
    dim_j = len(matrix1[0])

    for i in range(dim_i):
        for j in range(dim_j):
```

```
47              diff += pow(matrix1[i][j] - matrix2[i][j] , 2)
48
49      diff = pow(diff,0.5)
50      return diff
51
52  #"""
53  #Somme simple de deux vecteurs
54  #"""
55  def vectSum(vect1,vect2):
56      #print "sum of vect1 %s and vect2 %s"%(vect1,vect2)
57      res = []
58      if(vect1 == 0):
59          res = list (vect2)
60          return res
61
62      for i in range(len(vect1)):
63          res.append(vect1[i] + vect2[i])
64
65      return res
66
67  #"""
68  #Division element par element de deux vecteurs
69  #"""
70  def vectDiv(vect1,vect2):
71      res = []
72      for i in range(len(vect1)):
73          if(vect1[i] == 0):
74              res.append(0)
75          else:
76              res.append(float(vect1[i]) / float(vect2[i]))
77
78      return res
79
80  #"""
81  #Division element par elemeent de deux vecteurs en faisant attention
82  #a mettre zero si le denominateur = 0
83  #"""
84  def vectDivZero(vect1,vect2):
85      res = []
86      for i in range(len(vect1)):
87          if( vect2[i] == 0):
88              res.append(0)
89          else:
90              res.append(float(vect1[i]) / float(vect2[i]))
91      return res
92
93  #"""
94  #Vecteur a la puissance element par element
95  #"""
96  def vectPower(vect,power):
97      res = []
98      for i in range(len(vect)):
99          res.append(pow(vect[i],power))
100
101     return res
102
103 #"""
104 #Multiplication element par element de deux vecteurs
105 #"""
```

```
106 def vectMult(vect1,vect2):
107     res = []
108     for i in range(len(vect1)):
109         res.append(float(vect1[i]) * float(vect2[i]))
110 
111     return res
112 #"""
113 #Multiplication d un vecteur par un scalaire
114 #"""
115 def vectScalareMult(scalar,vect):
116     res = []
117     for i in vect:
118         res.append(i*scalar)
119     return res
120 
121 #"""
122 #Inverse d un vecteur element par element
123 #"""
124 def vectDivInverse(vect):
125     res = []
126     for i in vect:
127         res.append(float(1)/float(i))
128     return res
129 
130 #"""
131 #Vecteur de disctance entre deux vecteurs: sqrt(somme (xi - yi)^2)
132 #"""
133 def vectDistance(vect1,vect2):
134     res = []
135     #print "distance between %s and %s"%(vect1,vect2)
136     dist_local = 0
137     dist = 0
138     for i in range(len(vect1)):
139         dist_local = abs( float(vect1[i]) - float(vect2[i]))
140         dist_local = pow(dist_local,2)
141         dist += dist_local
142 
143     dist = pow(dist,0.5)
144 
145     return dist
146 #"""
147 #Valeur absolue d un vecteur : sqrt(somme(x_i^2))
148 #"""
149 def getAbsoluteValue(vect):
150     s = 0
151     for i in vect :
152         s += pow(i,2)
153     return pow(s,0.5)
154 #"""
155 #Calcule nouvelle matrice d appartenance des points au clusters
156 #parametres:
157 #centroidList : liste des centroides
158 #membershipMatrix: l ancienne matrice d appartenance a l etape
        precedante
159 #data: ensemble des points a regrouper
160 #power: parametre de Cmeans
161 #"""
162 def calculateNewMembershipMatrix(centroidList,membershipMatrix,data,
        power):
```

```
163     for i in range(len(centroidList)):
164         for j in range ( len(data) ):
165 
166             inner_sum = 0
167             for k in range(len(centroidList) ):
168                 d_i_j = vectDistance(centroidList[i],data[j])
169                 d_k_j = vectDistance(centroidList[k],data[j])
170                 if(d_k_j == 0):
171                     pass
172                 if(d_k_j != 0):
173                     div_res = (d_i_j/d_k_j)
174                     new_power = float(2) / float(power-1)
175                     div_res = pow(div_res,new_power)
176                     inner_sum = inner_sum + div_res
177 
178             if(inner_sum != 0):
179                 inner_sum = float(1)/ inner_sum
180             membershipMatrix[j][i] = inner_sum
181 
182 #"""
183 #Calcule la dissimilarite pour arreter l algo.
184 #U(k+1) - U(k) < epsilon
185 #"""
186 def caculateDissimilarity(centroidList,membershipMatrix,data,dataDim,
        power):
187     J = 0
188 
189     for i in range(len(centroidList)):
190         inner_sum = 0
191         for j in range(len(data)):
192             u_i_j_m = pow( membershipMatrix[j][i], power)
193             d_i_j = vectDistance(data[j],centroidList[i])
194             mult = d_i_j*u_i_j_m
195             inner_sum = inner_sum + mult
196 
197         J = J+inner_sum
198 
199     return J
200 
201 #"""
202 #Calcule les centroides
203 #"""
204 def caculateCentroid(centroidList,membershipMatrix,data,powerParam):
205     for i in range(len(centroidList)):
206         #caluclate the sum
207         mysum = 0
208         mysum2 = 0
209         for j in range(len(data)):
210             u_i_j = membershipMatrix[j][i]
211             u_i_j_m = pow(u_i_j,powerParam)
212             mysum2 = mysum2+u_i_j_m
213             u_i_j_m__x_j = vectScalareMult( u_i_j_m, data[j])
214             mysum = vectSum(mysum,u_i_j_m__x_j)
215 
216         if(mysum2 == 0):
217             pass
218         else:
219             mysum2 = float(1)/float(mysum2)
220 
```

```
221         centroidList[i] = vectScalareMult(mysum2,mysum)
222
223 #"""
224 #Cree une matrice vide nrows x ncols
225 #"""
226 def createMatrix(nrows,ncols):
227     matrix = []
228
229     i = 0
230     while (i < nrows):
231         inner_matrix = range(ncols)
232         for k in range(ncols) :
233             inner_matrix[k] = 0
234         matrix.append(inner_matrix)
235         i += 1
236
237     return matrix
238
239 #"""
240 #Initialisation de la matrice d appartenance
241 #"""
242 def initializeMemberShipMatrix(matrix,nrows,ncols,dataDim, factor=1000):
243     for line in range(nrows):
244         maxvalue = ncols * factor
245         jcols = range(ncols)
246         for j in range (ncols - 1):
247             #print "i = %i, j = %i"%(line,j)
248             currentvalue_int = rand.randint(0,maxvalue)
249             maxvalue = maxvalue - currentvalue_int
250             current_value_float = float(currentvalue_int)/float(ncols*
                    factor)
251             currentindex = rand.choice(jcols)
252             jcols.remove(currentindex)
253             matrix[line][currentindex] = current_value_float
254
255         current_value_float = float(maxvalue)/float(ncols*factor)
256         matrix[line][jcols[0]] = current_value_float
257
258 #"""
259 #Calculer les degrés d'appartenance au clusters ainsi que les centroïdes
260 #"""
261 def fuzzyCmeans(data,number_cluster,dataDim,power,epsilon,maxstep):
262     i_max = len(data)
263     j_max = number_cluster
264     centroid_list = range(j_max)
265
266     membershipMatrix = createMatrix(i_max,j_max)
267     initializeMemberShipMatrix(membershipMatrix,i_max,j_max,dataDim)
268
269     diff = 1000 # valeur initiale epsilon tres grande.
270     step = 0
271     old_J = 0
272     old_matrix = 0
273
274     while(not (diff < epsilon) and step < maxstep):
275         #""" Etape k#"""
276         #""" Calculer les centroides #"""
277         caculateCentroid(centroid_list,membershipMatrix,data,power)
278         #""" Calculer la nouvelle matrice d appartenance #"""
```

```
279         calculateNewMembershipMatrix(centroid_list,membershipMatrix,data
                ,power)
280         step += 1
281 
282         if(old_matrix != 0):
283             diff = diffMatrix(old_matrix,membershipMatrix)
284             #""" avoir le facteur U(k) - U(k+1)#"""
285         #""" Sauver la matrice a l etape k et passer a l etape k+1#"""
286         old_matrix = copyMatrix(membershipMatrix)
287 
288     return (membershipMatrix,centroid_list)
```

C

Système d'information du SEMIDE

C.1 Introdction

Le SEMIDE est un instrument de l'échange d'informations et de connaissances dans le domaine de l'eau entre tous les pays du Partenariat Euro-Méditerranéen. Dans ce cadre, nous avons besoin d'accéder aux informations disponibles sur l'eau issues de sources distantes et hétérogènes pour :

- identifier les connaissances disponibles dans le domaine Euro-Med ;
- rendre les informations accessibles ;
- permettre à la communauté internationale d'accéder aux services développés par le SEMIDE.

L'accès aux informations mises à disposition se fait à partir du portail du SEMIDE. Elles sont fournies et utilisées par les différents acteurs du système, qui sont :

1. des *fournisseurs d'information*, qui gèrent des ressources informationnelles dans le secteur de l'eau et diffusent les résultats des différents projets (Projets MEDA-Eau,CE, etc) ;
2. des *intermédiaires*, qui facilitent l'accès à l'information (UT) ;
3. des *utilisateurs finaux*, qui accèdent aux informations pour réaliser leur recherche (professionnels, étudiants).

Le SEMIDE gère des données non quantitatives. Les documents peuvent être : *(i)* des nouvelles, *(ii)* des évènements, *(iii)* des textes de loi, *(iv)* des multimedia, *(v)* des projets et *(vi)* d'autres documents plein texte.

Afin de faciliter l'exploitation de ces ressources, nous avons défini des métadonnées communes pour chaque type.

C.2 Les métadonnées dans le SEMIDE

Rappelons que les métadonnées sont des "données sur des données". Elles représentent une solution à la recherche d'informations. Il s'agit d'un ensemble standard et structuré d'informations, traitable par un logiciel, décrivant une ressource sur support électronique ou papier. Elles comprennent un certain nombre d'éléments qui permettent de les rendre plus facilement identifiables (accessibles) et plus manipulables (interopérables, réutilisables, durables, adaptables).

Les métadonnées sont utilisées depuis très longtemps dans les bibliothèques pour classer les documents. Les bibliothécaires parlent de catalogue [Dhé05]. On retrouve aussi des métadonnées sous forme de méta étiquettes (meta-tag) dans les pages HTML des sites Web, ces étiquettes sont utilisées par les moteurs de recherche (ex. Google[32]).

Il existe quatre types de métadonnées :

- *les métadonnées de description*, qui représentent la ressource et son apport informationnel (titre, date de publication, auteur) ;
- *les métadonnées administratives*, liées à la gestion des ressources (droits d'auteurs, propriété intellectuelle,...) ;
- *des informations sur le contenu* : il peut s'agir des mots-clés qui sont issus d'un thésaurus ou d'un glossaire ou bien du résumé. Ce type de métadonnées nous intéressera plus particulièrement dans la suite de notre travail ;
- *des informations sur l'accès* : localisation, protocole d'accès, structuration des données.

Pour des ressources informationnelles, les métadonnées facilitent :

- *la description du contenu* et des relations entre divers contenus ;
- *la classification* des contenus ;
- *la recherche* par extension des critères de recherche étendus au-delà des mots clés ;
- *la localisation* ;
- *la gestion* et la *préservation des ressources* ou collections de ressources ;
- *l'interopérabilité* pour l'échange, le partage et l'intégration de contenu ;
- *la gestion des droits* et l'authentification des contenus ;
- *la personnalisation du contenu* d'une interface utilisateur (ex. My Yahoo[33]).

C.2.1 Métadonnées et Web

Un des principes du Web sémantique est d'associer aux ressources du Web des informations qui peuvent être exploitées par des logiciels afin d'utiliser ces ressources. On peut par exemple associer un index de mots clés, une notice comprenant des informations sur la ressource telles que l'auteur, la date de

[32] www.google.fr
[33] fr.my.yahoo.com/

création, etc. D'après le *Rapport Action Spécifique 'Web sémantique'* [CLR03], associer une information exploitable à une ressource signifie que : *(i)* l'information doit être structurée, utilisable et descriptive afin de faciliter et d'améliorer l'accès à une ressource directement visualisée par un utilisateur, et l'exploitation d'une ressource dans le cadre d'un service à l'utilisateur, *(ii)* la ressource doit exister indépendamment de ces informations ajoutées, lesquelles doivent être utiles mais non nécessaires à l'utilisation et l'exploitation de la ressource.

Les métadonnées facilitent le travail des moteurs d'indexation et de recherche en leur permettant d'extraire de manière automatique des informations sur la ressource. Une définition commune des métadonnées est nécessaire par les acteurs de la communauté[34]. C'est pour cela que la communauté du Web sémantique a défini des standards, tels que Dublin Core, adaptés aux ressources électroniques. Ces standards proposent un schéma de métadonnées liées aux pages Web.

Les standards utilisés pour les métadonnées sont de trois types :

- *logique pour le protocole d'interopérabilité*. Par exemple, le RDF (Ressource Description Framework) peut être considéré comme un métalangage qui donne un cadre formel aux métadonnées sans toutefois préciser la sémantique des ressources ;
- *sémantique* pour des jeux d'éléments descriptifs (Dublin Core) ;
- *syntaxique*, qui est le format d'écriture (HTML, XML).

La normalisation des documents, des métadonnées et de la sémantique des documents suit des formalismes allant du plus simple au plus complexe. Certains formats de métadonnées sont simples, de type document HTML, généralement pauvres, et d'autres plus structurés comme le Dublin Core, ou le Resource Description Framework (RDF).

La figure C.1 illustre l'utilisation des métadonnées dans le Web sémantique. Les pages sont annotées en utilisant des ontologies, et ces annotations servent à des agents pour la recherche et la découverte d'informations.

C.2.2 Métadonnées dans le SEMIDE

Le SEMIDE étant un projet à long terme, nous souhaitons intégrer les évolutions à venir, notamment vers une architecture orientée Web sémantique. L'objectif est de permettre aux machines d'exploiter automatiquement les contenus de sources d'information accessibles par le Web pour accomplir des tâches variées. La réalisation de cet objectif repose sur l'existence de données, accessibles par le Web, structurées ou semi structurées, représentées dans un formalisme autorisant des processus automatisés allant au delà des traitements liés à la présentation des données et mettant en œuvre des mécanismes d'inférence puissants. A partir de cela, des métadonnées communes sont nécessaires pour l'échange et l'intégration de contenu ainsi que interopérabilité des

[34] Ensemble de personnes ayant un, ou des intérêts communs.

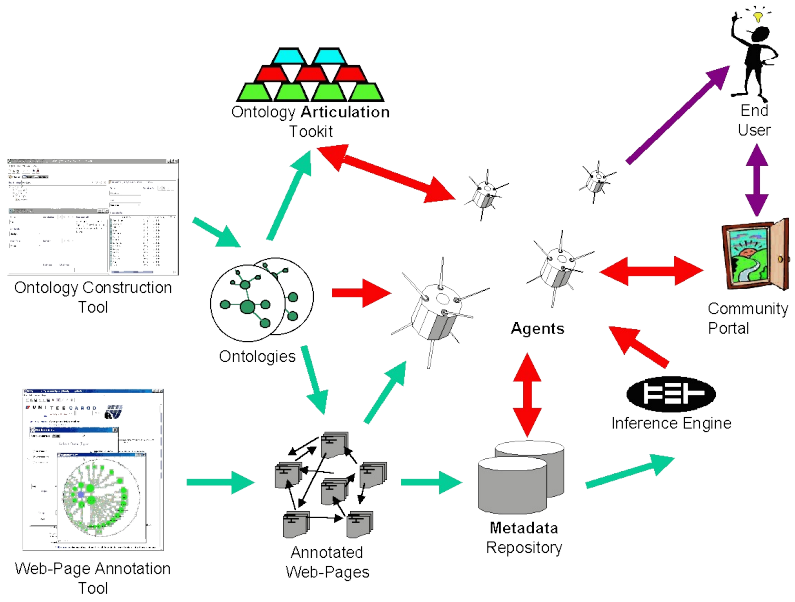

FIG. C.1 – Utilisation des métadonnées dans le Web sémantique, extrait de [CLR03]

services fournis. Nous avons défini, avec les différents points focaux[35], des métadonnées pour chaque type de ressource géré par le SEMIDE. La figure C.2 représente quelques métadonnées du type événement. A chaque événement, nous associons un liste d'éléments (titre, résumé, références...)

C.3 Processus de recherche d'information existant au SEMIDE

La figure C.3 présente un cas pratique, où un utilisateur cherche de l'information dans le domaine de l'eau. Les documents sont distribués sur les différents points focaux. Cet utilisateur peut :

- soit interroger un point focal national particulier ;
- soit faire une recherche sur l'ensemble des points focaux.

Notre démarche a été justifiée par les contraintes suivantes :

- l'information peut être sur les différents points focaux,
- les points focaux ne peuvent nous fournir que quelques métadonnées sur leurs documents. Effectivement, pour une question de confidentialité, il

[35]Les points focaux sont essentiellement chargés de tâches de mobilisation et de diffusion de la documentation et de l'information relative au secteur de l'eau.

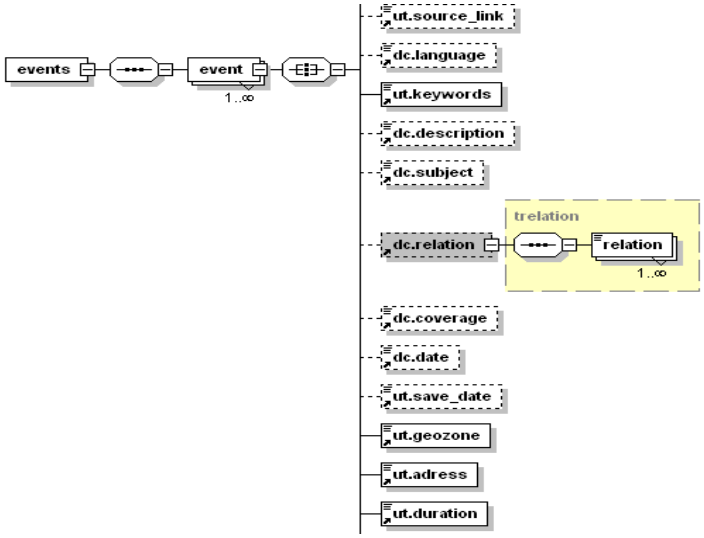

FIG. C.2 – Quelques métadonnées du type événement

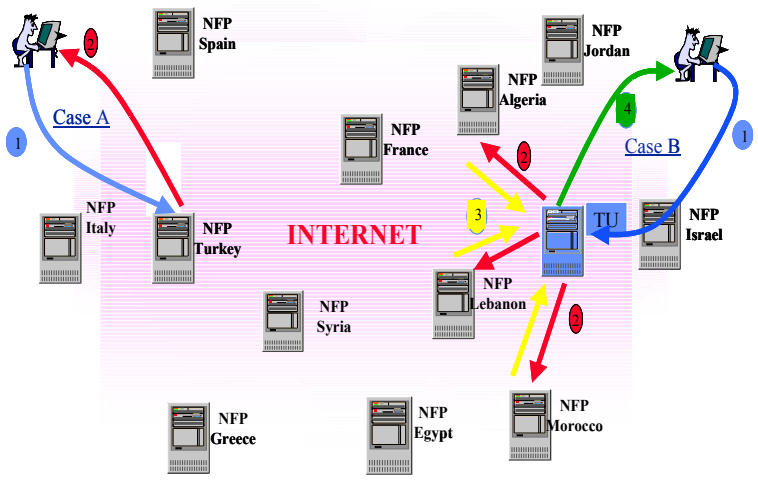

FIG. C.3 – Cas pratique de recherche de documents

est rare que le PFN fournisse le contenu complet ;
– les différents acteurs du système ne parlent pas la même langue.

D

L'ontologie du SEMIDE

Nous illustrons dans ce chapitre un extrait de l'ontologie du SEMIDE, les termes sont reliés par la relation de généralisation/spécialisation. Le premier niveau qui correspond au thème est également appelé *Terme de Tête*.

Elle est basée sur les termes et relations du thésaurus de l'OIEau (Office Internationale de l'Eau), leur thésaurus étant opérationnel. Un travail de traduction est en cours dans différentes langues (italien, espagnol, arabe) par les différents points focaux et l'unité technique du SEMIDE.

Les termes existent aussi en anglais, cette traduction a été effectuée en se basant sur différents ressources (dictionnaires, glossaires, thésaurus...) existantes, parmi lequels : le thésaurus GEMET[36] et le glossaire international d'hydrologie (UNESCO[37])

```
DEMANDE EN EAU (Terme de Tête)

        DEMANDE EN EAU
           GESTION DE LA DEMANDE EN EAU
           SYSTEME D'OFFRE ET DE DEMANDE EN EAU
        RESSOURCE EN EAU
           ETAT QUANTITATIF
              GESTION DE NAPPE
                 NAPPE INTENSEMENT EXPLOITEE
        GESTION DE LA RESSOURCE EN EAU
           GESTION INTEGREE DE LA RESSOURCE EN EAU
        RESSOURCE EN EAU NON RENOUVELABLE
           NAPPE FOSSILE
        RESSOURCE EN EAU RENOUVELABLE
        PENURIE D'EAU
           PENURIE D'EAU CONJONCTURELLE
           PENURIE D'EAU STRUCTURELLE
        ECONOMIE D'EAU
```

[36]http ://www.eionet.europa.eu/GEMET/
[37]http ://portal.unesco.org/

```
            RECYCLAGE DE L'EAU
            REUTILISATION DE L'EAU
            PRELEVEMENT D'EAU
               CONSOMMATION D'EAU
         RECHERCHE D'EAU
               ZONE DE REPARTITION DES EAUX


AGRICULTURE (Terme de Tête)
         AGRICULTURE
            AGRICULTEUR
            AGRICULTURE BIOLOGIQUE
            AGRICULTURE INTENSIVE
            AGRICULTURE RAISONNEE
            AGROMETEOROLOGIE
            AGRONOMIE
            ARBORICULTURE
            COOPERATIVE AGRICOLE
            CULTURE CEREALIERE
            CULTURE ENERGETIQUE
            CULTURE FOURRAGERE
            CULTURE INDUSTRIELLE
            CULTURE INTERMEDIAIRE
            CULTURE MARAICHERE
            DESHERBAGE
            ELEVAGE
            EXPLOITATION AGRICOLE
               MATERIEL AGRICOLE
            FERTILISATION
               FERTILISATION RAISONNEE
            HORTICULTURE
            HYDRAULIQUE AGRICOLE
               DRAINAGE
               IRRIGATION
                  CONDUITE DE L'IRRIGATION
                  IRRIGATION DE SURFACE
                    IRRIGATION GRAVITAIRE
                    IRRIGATION PAR ASPERSION
                  IRRIGATION SOUTERRAINE
                  MATERIEL D'IRRIGATION
                  MICROIRRIGATION
                  RETENUE COLLINAIRE
            JARDINAGE
            LUTTE BIOLOGIQUE
```

LUTTE CONTRE L'EROSION
PAILLE
PRATIQUE AGRICOLE
CULTURE ASSOLEE
JACHERE
SILO
SOL
APTITUDE DU SOL
DECONTAMINATION DU SOL
EROSION DU SOL
HUMUS

TOURISME - SPORT - LOISIR (Terme de Tête)
LOISIR
CHASSE
EAU DE BAIGNADE
PECHE
THALASSOTHERAPIE
ZONE DE LOISIR
SPORT
GOLF
NAVIGATION DE PLAISANCE
SPORT NAUTIQUE
TOURISME
AMENAGEMENT TOURISTIQUE
PISCINE
PLAGE
STATION BALNEAIRE
STATION DE MONTAGNE
CAMPING
TOURISME DURABLE

EAU POTABLE (Terme de tête)
PROCEDES DE TRAITEMENT STRICTS
ANTITARTRE
DESSALEMENT DE L'EAU
ELECTRODIALYSE
OSMOSE INVERSE
DISTILLATION DE L'EAU
TRAITEMENT DE L'EAU
TRAITEMENT DOMESTIQUE
USINE DE TRAITEMENT

```
            FILIERE DE TRAITEMENT
    RESSOURCE CAPTAGE ET DISTRIBUTION DE L'EAU
        ALIMENTATION EN EAU
           ADDUCTION D'EAU
              AQUEDUC
           DISTRIBUTION D'EAU
              STOCKAGE D'EAU
                 RESERVOIR D'EAU
        CAPTAGE
           CAPTAGE D'EAU SOUTERRAINE
           CAPTAGE D'EAU SUPERFICIELLE
           PROTECTION DE CAPTAGE
           STATION DE POMPAGE
           ALIMENTATION DE NAPPE
              ALIMENTATION ARTIFICIELLE DE NAPPE
              ALIMENTATION NATURELLE DE NAPPE
           AQUIFERE
              BARRAGE SOUTERRAIN
              KARST
                 ENGOUFFREMENT
                 RIVIERE SOUTERRAINE
                   RESURGENCE
              NAPPE ALLUVIALE
                 NAPPE CAPTIVE
                 NAPPE LIBRE
           CAPILLARITE
           DECONTAMINATION DE NAPPE
           ECHANGE NAPPE RIVIERE
           FORAGE
           PUITS DE DEPOLLUTION
           MILIEU FISSURE
           MILIEU NON SATURE
           MILIEU POREUX
           MILIEU SATURE
           PROSPECTION GEOPHYSIQUE
           PROTECTION DE NAPPE
           PUITS
              PUITS ARTESIEN
           RABATTEMENT DE NAPPE
           RESTAURATION DE NAPPE
           SOURCE
           VARIATION DE NIVEAU DE NAPPE
           VULNERABILITE DE NAPPE
```

ASSAINISSEMENT -PROCEDES D'EPURATION STRICTS (terme de tête)
- ANOXIE
- ASSAINISSEMENT
  - ASSAINISSEMENT COLLECTIF
  - ASSAINISSEMENT NON COLLECTIF
- EPURATION DE L'EAU
  - STATION D'EPURATION
    - FILIERE D'EPURATION
      - DEPHOSPHATATION
      - LIT FLUIDISE
      - PROCEDE BIOLOGIQUE
        - BOUE ACTIVEE
        - CULTURE FIXEE
          - DISQUE BIOLOGIQUE
          - EPURATION PAR LE SOL
          - LIT BACTERIEN
        - CULTURE LIBRE
        - EPURATION ANAEROBIE
        - LAGUNAGE
        - LIT A MACROPHYTE
      - TRAITEMENT TERTIAIRE
    - RENDEMENT D'EPURATION
  - EPANDAGE
    - CHAMP D'EPANDAGE
    - EPANDAGE D'EAU USEE
    - EPANDAGE DE BOUE
    - MATERIEL D'EPANDAGE
    - PLAN D'EPANDAGE

Bibliographie

[AAMH00] E. Agirre, O. Ansa, D. Martinez, and E. Hovy. Enriching very large ontologies using the www. In *Proceedings of Ontology Learning Workshop*, 2000.

[AAMH01] E. Agirre, O. Ansa, D. Martinez, and E. Hovy. Enriching wordnet concepts with topic signatures. In *Procceedings of the SIGLEX workshop on "WordNet and Other Lexical Resources : Applications, Extensions and Customizations". In conjunction with NAACL*, 2001.

[ABB00] M. Ashburner, CA. Ball, and JA. Blake. Gene ontology : tool for the unification of biology. *The gene ontology consortium*, 2000.

[AGBS00] N. Aussenac-Gilles, B. Biébow, and S. Szulman. Modélisation du domaine par une méthode fondée sur l'analyse de corpus. In *Proceedings of Journées francophones d'ingénierie des connaissances (IC'2000)*, pages 97–104, Toulouse, France, Mai 2000.

[Age99] European Environmental Agency. Gemet - general multilingual environmental thesaurus. *Technical report, European Topic Centre on Catalogue of Data Sources (ETC/CDS) European Environmental Agency (Version 2.0)*, 1999.

[Agu02] F. Aguiar. *Modélisation d'un système de recherche d'information pour les systèmes hypertextes. Application à la recherche d'information sur le World Wide Web.* PhD thesis, Ecole supérieure des Mines de Saint-Etienne, 2002.

[Ahm05] W. Ben Ahmed. *Safe-Next : Une approche systémique pour l'extraction de connaissances de données. Application à la construction et à l'interprétation de scénarios d'accidents de la route.* PhD thesis, Ecole centrale des arts et manufactures 'école centrale Paris', 2005.

[AL05] L. Abrouk and M. Lafourcade. Application of the papillon project to ontology management. In *Proceedings of PAPILLON-SNLP-05, 6th Symposium on Natural Language Processing*, Chiang Rai, Thailande, 12-14 décembre 2005.

[AL06] L. Abrouk and M. Lafourcade. Enrichissement d'ontologies dans le secteur de l'eau douce en environnement internet distribué et

multilingue. In *Proceedings of EGC'2006*, Lille, France, 25-27 janvier 2006.

[All87] R. Allier. *Nouveau Larousse illustré. Dictionnaire universel encyclopédique. Chapter ontologie*. Larousse, 1987.

[ALM05a] F. Amardeilh, P. Laublet, and J. L. Minel. Document annotation and ontology population from lingusitic extraction. In *Proceedings of Third International Conference on Knowledge Capture*, 2005.

[ALM05b] F. Amardeilh, P. Laublet, and J.L. Minel. Annotation documentaire et peuplement d'ontologies à partir d'extractions linguistiques. In *Proceedings of IC 2005*, Août 2005.

[AMS04] S. Aupetit, N. Monmarché, and M. Slimane. Utilisation des chaînes de markov cachées à substitution de symboles pour l'apprentissage et la reconnaissance robuste d'images. In *Majestic*, 2004.

[ANTT01] A. Arasu, J. Novak, A. Tomkins, and J. Tomlin. Pagerank computation and the structure of the web : Experiments and algorithms, 2001.

[Ass98] H. Assadi. *Construction d'ontologies à partir de textes techniques - Application aux systèmes documentaires*. PhD thesis, Université Paris6, 1998.

[BAG03] D. Bourigault and N. Aussenac-Gilles. Construction d'ontologies à partir de textes. In *Proceedings of 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2003)*, pages 27–50, Batz-sur-Mer, Juin 2003.

[Baz05] M. Baziz. *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. PhD thesis, Institut de recherche en informatique de Toulouse, université Paul Sabatier, 2005.

[BCGM+99] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of web pages. In *Proceedings of WebDB (Informal Proceedings)*, 1999.

[BCW02] C. Brewster, F. Ciravegna, and Y. Wilks. User-centred ontology learning for knowledge management. In *Proceedings of NLDB '02 : 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers*, pages 203–207, London, UK, 2002. Springer-Verlag.

[Bez81] J. C. Bezdek. Pattern recognition with fuzzy objective function algoritms. *Plenum Press*, 1981.

[BKMW89] J. Bateman, R. Kasper, J. Moore, and R. Whitney. A general organization of knowledge for natural language processing : The

penman upper model. technical report. In *Proceedings of Information Sciences Institute*, 1989.

[BL06] M. Bouklit and M. Lafourcade. Propagation de signatures lexicales dans le graphe du web. In *Proceedings of RFIA 2006, 15e congrès francophone AFRIF-AFIA, Reconnaissance des Formes et Intelligence Artificielle*, Janvier 2006.

[Bla90] DC. Blair. Language and representation in information retrieval. In *Proceedings of Elesevier Science Publications*, Amsterdam, 1990.

[Bou94] D. Bourigault. Lexter, un logiciel d'extraction de terminologie. application à l'acquisition des connaissances à partir de textes., 1994.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of seventh international conference on World Wide Web 7*, pages 107–117, Australia, 1998.

[BS99] B. Biebow and S. Szulman. Terminae : A linguistics-based tool for building of a domain ontology. In *D. Fensel and R. Studer, editors, the 11th European Workshop (EKAW'99), LNAI 1621*, pages 49–66, Heraklion, Crete., 1999.

[Cha99] J. Chauché. Un outil multidimensionnel de l'analyse du discours. In *Coling-84*, pages 11–15, Standford University, California, 1999.

[CLR03] J. Charlet, P. Laublet, and C. Reynaud. Action spécifique 32 web sémantique. rapport final. Technical report, CNRS/STIC, Octobre 2003.

[CR97] A. Condamines and J. Rebeyrolle. Construction d'une base de connaissances terminologique à partir de textes. In *Proceedings of Journées Ingénierie des Connaissances et Apprentissage Automatique, JICAA'97*, pages 191–206, Roscoff, mai 1997.

[CST03] P. Cimiano, S. Staab, and J. Tane. Automatic acquisition of taxonomies : Fca meets nlp. In *Proceedings of the ECML/PKKD Workshop on Adaptive Text Extraction and Mining*, 2003.

[Dam03] O. Dameron. Modélisation, représentation et partage de connaissances anatomiques sur le cortex cérébral, 2003.

[DDL+90] C.S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6) :391–407, 1990.

[Dhé05] C. Dhérent. Les métadonnées, à quoi ça sert ?, 2005.

[Dij59] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1 :269–271, 1959.

[DJ02] E. Desmontils and C. Jacquin. Indexing a web site with a terminology oriented ontology. *The Emerging Semantic Web, I.F. Cruz, S. Decker, J. Euzenat and D. L. McGuinness Ed*, pages 181–197, 2002.

[DJM02] E. Desmontils, C. Jacquin, and E. Morin. Indexation sémantique de documents sur le web : application aux ressources humaines. In *Proceedings of Journées de l'AS-CNRS Web sémantique*, Octobre 2002.

[Dub04] R. Dubois. Application des nouvelles méthodes d'apprentissage à la détection précoce d'anomalies en électrocardiographie, 2004.

[Dun73] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3 :32–57, 1973.

[EMT92] C. Enguehard, P. Malvache, and P. Trigano. Indexation de textes : l'apprentissage des concepts. In *Proceedings of 14th conference on Computational linguistics*, pages 1197–1202, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[Fen00] D. Fensel. *Ontologies : Silver Bullet for Knowledge Management and Electronic Commerce*. Springer, 2000.

[FFHS02] O. Ferret, C. Fluhr, FR. Hans, and JL. Simoni. Building domain specific lexical hierarchies from corpora. In *Proceedings of LREC 2002*, Las Palmas, Espagne, May 29-31 2002.

[Fot04] H. Njike Fotzo. *Structuration Automatique de Corpus Textuels par Apprentissage Automatique*. PhD thesis, Université Paris6, 2004.

[Für02] F. Fürst. L'ingénierie ontologique. Rapport de recherche, Octobre 2002.

[FS02] A. Faatz and R. Steinmetz. Ontology enrichment with texts from the www. In *Proceedings of Semantic Web Mining 2nd Workshop at ECML/PKDD-2002*, Août 2002.

[FWE03] B. C. M. Fung, K. Wang, and M. Ester. Hierarchical document clustering using frequent itemsets. In *Proceedings of the SIAM International Conference on Data Mining*, 2003.

[Gar93] E. Garfield. Co-citation analysis of the scientific literature : Henry small on mapping the collective mind of science. *Essays of an Information Scientist : Of Nobel Class, Women in Science, Citation Classics and Other Essays*, 15(19), 1993.

[Gau98] E. Gauthier. L'analyse bibliométrique de la recherche scientifique et technologique : Guide méthodologique d'utilisation et d'interpréstation. In *Proceedings of Observatoire des Sciences et des Technologies*, 1998.

[GBJJ05] C. Ghaoui, V. Bannore, L.C. Jain, and M. Jain. *Knowledge-Based Virtual Education : User-Centred Paradigms*. Springer, 2005.

[GD02] M. Martinez M. Gordon et al G. Dreyfus, M. Samuelides. Réseaux de neurones. 2002.

[GHN05] S. Ghita, N. Henze, and W. Nejdl. Task specific semantic views : Extracting and integrating contextual metadata from the web. In *Proceedings of the International Semantic Web Conference Workshop on The Semantic Desktop - Next Generation Personal Information Management and Collaboration Infrastructure, ISWC*, Galway, Ireland, November 2005.

[GKR98] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proceedings of UK Conference on Hypertext*, 1998.

[GMV99] N. Guarino, C. Masolo, and G. Vetere. Ontoseek : Content-based access to the web, 1999.

[Gru93] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis*, 5(2) :199–220, 1993.

[Gua98] N. Guarino. Formal ontology in information systems. In *Proceedings of FOIS'98*, pages 3–15, 6-8 June 1998.

[Gua99] N. Guarino. Some organizing principles for a unified top-level ontology. In *Proceedings of AAAI Spring Symposium on Ontological Engineering.*, pages 1.1–1.15, Stanford, 1999.

[GW97] B. Ganter and R. Wille. Applied lattice theory : Formal concept analysis, 1997.

[HCGS99] J. Habrant, A. Corbel, J. Girardot, and J. Savoy. Utilisation des réseaux sémantiques pour la navigation dans l'hypertexte. In *Proceedings of Colloque Multimédia et Construction des Savoirs*, mai 1999.

[HL05] J. S. Hare and P. H. Lewis. Saliency-based models of image content and their application to auto-annotation by semantic propagation. In *Proceedings of Multimedia and the Semantic Web / European Semantic Web Conference 2005*, Heraklion, Crete., 2005.

[HM02] S. Harabagiu and S. Maiorano. Multi-document summarization with gistexter. In *Proceedings of Third LREC Conference 2002 (LREC 2002)*, 2002.

[HMM99] S. Harabagiu, G. Miller, and D. Moldovan. Wordnet 2 - a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX Workshop*, pages 1–8, 1999.

[JFBR92] M. Joubert, M. Fieschi, G. Botti, and J.J. Robert. Représentation de concepts médicaux pour la recherche d'information : réalisation d'une maquette à partir de mmls. *Informatique et Santé.*

Nouvelles Méthodes de Traitement de l'Information en Médecine, 5, 1992.

[Jou93] C. Jouis. Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes, réalisation d'un prototype : le système seek, 1993.

[KCG+01] R. Dieng Kuntz, O. Corby, F. Gandon, A. Giboin, J. Golebiowska, N. Matta, and M. Ribière. *Méthodes et outils pour la gestion des connaissances : une approche pluridisciplinaire du knowledge management*. Deuxième édition. Dunod, 2001.

[KDK04] K. Khelif and R. Dieng-Kuntz. Annotations sémantiques pour le domaine des biopuces. In *Proceedings of 15èmes journées francophones d'ingénierie des connaissances*, 2004.

[Kes65] M.M. Kessler. Comparison of the results of bibliographic coupling and analytic subject indexing. *American documentation*, 14 :10–15, 1965.

[Kha00] L.R. Khan. Ontology-based information selection, 2000.

[Kle99] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Journal of the ACM*, pages 139–146, 1999.

[KRRT99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of Computer Networks*, Amsterdam, Netherlands, 1999.

[LA04] S. Laszlo and N. Amedeo. Les treillis de galois pour l'organisation et la gestion des connaissances. In *Proceedings of 11èmes Rencontres de la Société Francophone de Classification - SFC '04.*, pages 298–301, Bordeaux, France, 2004.

[Lam02] G. Lame. *Construction d'ontologie a partir de textes. Une ontologie du droit dédiée à la recherche d'information sur le Web.* PhD thesis, Ecole des Mines de Paris, 2002.

[Lau97] P. Lauri. The bibliometrics a trend indicator. In *International Journal Information Sciences for Decision Making*, page 2836, 1997.

[LBG99] S. Lawrence, K. Bollacker, and C.L. Giles. Indexing and retrieval of scientific literature. In *Proceedings of eight International Conference on Information and Knowledge Management, CIKM99*, 1999.

[LHL01] T. Berbers Lee, J. Hendler, and O. Lasila. The semantic web. *scientific American*, 2001.

[LMB05] C. Latiri, M. Mtir, and S. Benyahia. Méthode de construction d'ontologie de termes à partir du treillis de l'iceberg de galois. In *Proceedings of Extraction et Gestion des Connaissances (EGC2005)*, pages 365–376, Paris, 19-21 Janvier 2005.

[Mar64] J. Martyn. bibliographic coupling. *Journal of Documentation*, 20(4) :236, 1964.

[Mar98] M. Marchiori. The limits of web metadata, and beyond. In *Proceedings of Seventh International World Wide Web Conference*, pages 1–9, Australia, 1998.

[MBF+90] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to wordnet : An on-line lexical database. *International Journal of Lexicography*, 1990.

[Mes04] N. Messai. Treillis de galois et ontologies de domaine pour la classification et la recherche de sources de données génomiques, Juin 2004.

[Miz04] R. Mizoguchi. Le rôle de l'ingénierie ontologique dans le domaine des eiah. *Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation*, 11 :231–246, 2004.

[MTF04] R. Mihalcea, P. Tarau, and E. Figa. Pagerank on semantic networks, with application toword sense disambiguation. In *Proceedings of 20th international conference on computational linguistics (COLING2004)*, Geneva, Switzerland, 2004.

[NFF+91] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W.R. Swartout. *Enabling Technology For Knowledge Sharing*. AI Magazine, 1991.

[NFF04a] W. Njmogue, D. Fontaine, and P. Fontaine. Identification des thèmes d'un document relativement à un référentiel métier. In *Proceedings of MAJECSTIC'04*, 13-15 Octobre 2004.

[NFF04b] W. Njmogue, D. Fontaine, and P. Fontaine. Indexation des documents dans un référentiel métier. In *Proceedings of Workshop ALCAA 2004, Agents Logiciels - Coopération Apprentissage - Activité humaine*, 7-18 Juin 2004.

[NFG04] H. Njike-Fotzo and P. Gallinari. Learning generalization/specialization relations between concepts - application for automatically building thematic document hierarchies. In *Proceedings of RIAO 2004*, 26-28 Apr 2004.

[NFGL03] H. Njike-Fotzo, P. Gallinari, and N. Lagunas. Génération automatique d'une structure hierarchique de concepts et de documents à partir de corpus. In *Proceedings of CORIA 2004*, pages 57–74, 2003.

[NN05] E. M. Nguifo and P. Njiwoua. Treillis de concepts et classification supervisée. *Technique et Science Informatiques*, 24(4) :449–488, 2005.

[PB99] A.G. Perez and V.R. Benjamins. Overview of knowledge sharing and reuse components : Ontologies and problem-solving methods.

In *Proceedings of IJCAI99 workshop on Ontologies and Problem-Solving Methods (KRR5)*, pages 1.1–1.15, Stockholm, Sweden, August 1999.

[PC04] C. Prime-Claverie. *Vers une prise en compte de plusieurs aspects des besoins d'information dans les modèles de la recherche documentaire : Propagation de métadonnées sur le World Wide Web.* PhD thesis, Ecole supérieure des Mines de Saint-Etienne, 2004.

[PCBL02] C. Prime-Claverie, M. Beigbeder, and T. Lafouge. Clusterisation du web en vue d'extraction de corpus homogènes. In *Proceedings of INFORSID 2002, 20e congrès informatique des organisations et des systèmes d'information et de décision*, Nantes, France, 4-7 Juin 2002.

[PDB02] B. Pouliquen, D. Delamarre, and P. Beux. Indexation de textes médicaux par extraction de concepts, et ses utilisations. In *Proceedings of JADT'02, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, mars 2002.

[PGF04] V. Parekh, J. Gwo, and T. Finin. Mining domain specific texts and glossaries to evaluate and enrich domain ontologies. In *Proceedings of International Conference of Information and Knowledge Engineering*, 2004.

[PK02] D. Phelan and N. Kushmerick. A descendant-based link analysis algorithm for web search. 2002.

[PMB03] V. Psyche, O. Mendes, and J. Bourdeau. *Apport de l'ingénierie ontologique aux environnements de formation à distance*, volume 10 of *Article de recherche.* Sticef : Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation, 2003.

[PSH03] P. Pompidor, M. Sala, and D. Hérin. Une méthode incrémentale d'extraction de connaissances didactiques sur le web. In *Proceedings of EIAH'03 : Environnements Informatiques pour l'Apprentissage Humain*, pages 551–554, Pittsburgh, PA, 2003.

[Qui68] R. Quillian. Semantic informatic processing. *Chapitre Semantic memory*, page 227270, 1968.

[RA93] A. Swami R. Agrawal, T. Imielinski. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington DC, USA, 1993.

[Ras04] F. Rastier. Ontologie(s). *Revue des sciences et technologies de l'information*, 18(1) :15–40, 2004.

[Rij79] V. Rijsbergen. *Information Retrieval, 2nd edition.* Dept. of Computer Science, University of Glasgow, London, 1979.

[Ros96] H. Rostaing, editor. *La bibliométrie et ses techniques*. Sciences de la Société Collection, 1996.

[Sal71] G. Salton. A comparaison between manual and automatic indexing methods. In *Proceedings of Journal of American documentation*, 1971.

[Sal86] G. Salton. Another look at automatic text-retrieval systems. *Commun. ACM*, 29(7) :648–656, 1986.

[SC99] M. Sanderson and W. Bruce Croft. Deriving concept hierarchies from text. In *Research and Development in Information Retrieval*, pages 206–213, 1999.

[Sch05] D. Schwab. Approche hybride - lexicale et thématique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte, 2005.

[SCL+05] J. Stribling, I.G. Councill, J. Li, M. F. Kaashoek, D.R. Karger, R. Morris, and S. Shenker. Overcite : A cooperative digital research library. In *Proceedings of International Workshop on Peer-to-Peer Systems*, 2005.

[SDJ03] L. Simon, E. Desmontils, and C. Jacquin. Utilisation de techniques d'enrichissement d'ontologies pour améliorer le processus d'indexation structurée. In *Proceedings of AFIA 2003*, France, juillet 2003.

[Seg01] P. Seguéla. Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques., 2001.

[Sma73] H.G. Small. Co-citation in the scientific literature. *Society for Information Science*, 24, 1973.

[SSG85] H. Small, E. Sweeney, and E. Greenlee. Clustering the science citation index using co-citation. *8*, 5-6 :321–340, 1985.

[SWdH+94] G. Schreiber, B. Wielinga, R. de Hoog, H. Akkermans, and W. Van de Velde. Commonkads : A comprehensive methodology for kbs development. *IEEE Expert*, 9(6) :28–37, 1994.

[TBL01] O. Lassila T. Berners-Lee, J. Hendler. The semantic web. In *Scientific American*, May 2001.

[TD04] M. Thilliez and T. Delot. Evaluation de requêtes dépendantes de la localisation dans les réseaux mobiles. In *Proceedings of Premières Journées Francophones : Mobilité et Ubiquité 2004*, Nice, Juin 2004.

[Tex05] R. Texier. Taxinomies, thésaurus et ontologies. *EliKya, intelligence des organisations*, 2005.

[the01] Thesauri and controlled vocabularies. *National Library of Canada*, 2001.

[VFD04] V. Vandaele, P. Francq, and A. Delchambre. Analyse d'hyperliens en vue d'une meilleure description des profils. In *Proceedings of JADT 2004, 7es Journées internationales d'Analyse statistique de Données Textuelles*, 2004.

[VSkGP03] A. Valarakos, G. Sigletos, V. karkaletsis, and G. Vouros G. Paliouras. A methodology for enriching a multi-lingual domain ontology using machine learning. In *Proceedings of 6th ICGL workshop on Text Processing for Modern Greek : from Symbolic to Statistical Approaches, held as part of the 6th International Conference in Greek Linguistics*, September 2003.

[VSkP03] A. Valarakos, G. Sigletos, V. karkaletsis, and G. Paliouras. A methodology for semantically annotating a corpus using a domain ontology and machine learning. In *Proceedings of Recent Advances in Natural Language Processing International Conference (RANLP)*, pages 495–499, 2003.

[Wei74] B. H. Weinberg. bibliographic coupling : A review. *Information Storage and Retrieval*, 10 :189–196, 1974.

[WF05] Ian H. Witten and E. Frank. Data mining : Practical machine learning tools and techniques. *National Library of Canada*, 2005.

[Woo97] W. Woods. Conceptual indexing : A better way to organize knowledge, 1997.

[WXL99] Ke Wang, Chu Xu, and Bing Liu. Clustering transactions using large items. In *CIKM '99 : Proceedings of the eighth international conference on Information and knowledge management*, pages 483–490, New York, NY, USA, 1999. ACM Press.

[YMP06] M. Yousfi-Monod and V. Prince. Compression de phrases par élagage de l'arbre morpho-syntaxique. *Technique et Science Informatiques*, 2006.

[ZHL+04] L. Zhang, Y. Hu, M. Li, W. Ma, and H. Zhang. Efficient propagation for face annotation in family albums. In *Proceedings of MULTIMEDIA '04 : 12th annual ACM international conference on Multimedia*, pages 716–723, New York, NY, USA, 2004. ACM Press.

résumé de la thèse :

Cette thèse présente une approche et des outils pour l'annotation de documents en se basant sur des ontologies. Dans notre contexte, ceci se traduit par des documents annotés par un ensemble de concepts clés issus de l'ontologie du domaine. Nous traitons le problème de l'annotation en développant une approche basée sur la relation de citation. Cette relation constitue la base d'une méthode pour affiner la propagation des annotations entre les documents. L'approche est indépendante du contenu et utilise un regroupement thématique des références construit à partir d'une classification floue non-supervisée. L'annotation étant basée sur l'utilisation d'ontologies, nous avons également abordé le problème de l'enrichissement de l'ontologie afin de pouvoir prendre en compte les différentes évolutions des documents et affiner la phase d'annotation. Un outil, nommé RAS, Reference Annotation System, a été développé et des expérimentations ont été réalisées en utilisant la base Citeseer.

abstract :

This thesis presents an approach and tools for annotating documents with ontologies. In our context, this problem is regarded as annotating documents using a set of concepts that belong to a domain ontology. We deal with this problem by developing an approach based on the citation relationship. This relation is the core of a method to refine propagation of annotations between the documents. The approach is independent from documents content and uses clusters of references that are buillt with unsupervised fuzzy classification. We also considered the problem of updating the ontology in order to be able to take into account the various evolutions of documents and to refine the annotation step. A tool, named RAS,Reference System Annotation, has been developed and experiments were realized byusing the Citeseer base. Documents annotation with citation context based on ontology.