



THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Informatique*

Présentée et soutenue par *Fatiha BOUBEKEUR-AMIROUCHE*
Le *01/ 07/ 2008*

Titre : *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets*

JURY

Florence SEDES : Professeur à l'Université Paul Sabatier (Présidente)
Eric GAUSSIER : Professeur à l'Université Joseph Fourier Grenoble I, France (Rapporteur)
Mohand-Saïd HACID : Professeur à l'Université Claude Bernard Lyon 1, France (Rapporteur)
Gabriella PASI : Professeur à l'Université de Milan Bicocca, Italie (Examinatrice)
Mohand BOUGHANEM : Professeur à l'Université Paul Sabatier (Directeur de recherche)
Lynda TAMINE-LECHANI : Maître de Conférences à l'Université Paul Sabatier (Co-encadrante)

Ecole doctorale : *MITT*
Unité de recherche : *CNRS, 5505*
Directeur(s) de Thèse :

Mohand BOUGHANEM : Professeur à l'Université Paul Sabatier (Directeur de recherche)
Lynda TAMINE-LECHANI : Maître de Conférences à l'Université Paul Sabatier (Co-encadrante)
Rapporteurs :

*A mes enfants Amine et Nassim
A Boualem*

Remerciements

Je tiens à remercier en tout premier lieu **M. Mohand Boughanem** qui a dirigé cette thèse d'une main de maître. Tout au long de ces quatre années, il a su orienter mes recherches aux bons moments, toujours dans les bonnes directions. Malgré l'éloignement, il a toujours été disponible pour prodiguer des conseils et des orientations ô combien pertinentes. Pour tout cela, pour m'avoir offert la chance d'en être là aujourd'hui, pour sa confiance et pour sa précieuse aide technique je le remercie du fond du coeur.

Mes plus vifs remerciements vont également à **Mme Lynda Tamine-Lechani** qui a co-dirigé cette thèse avec toute la grandeur et la générosité qui sont les siennes. Tout au long de ces années de thèse, malgré l'éloignement, avec une régularité horlogique elle a toujours été là par ses conseils, par son suivi minutieux de toutes mes propositions dans le cadre de cette thèse, par ses corrections, par ses orientations et suggestions, par de riches et longues discussions. Et même au-delà de cet aspect scientifique, dans les plus pénibles moments de doute et de lassitude, elle a toujours été l'amie qui m'a aidée à me relever, qui m'a encouragée à persévérer. Pour tout cela, pour son aide ô combien précieuse, pour sa générosité, pour son amitié qu'elle trouve ici l'expression de ma plus profonde reconnaissance et de ma sincère amitié.

Je remercie les rapporteurs de cette thèse **M. Mohand-Saïd Hacid** et **M. Eric Gaussier** pour la lecture et la correction de mon manuscrit et pour l'intérêt qu'ils ont porté à mon travail. Je remercie également les autres membres du jury, **Mme Florence Sedes** et **Mme Gabriella Pasi** qui ont accepté de juger ce travail.

Je n'oublierai pas de remercier le premier Responsable de l'équipe SIG, **M. Claude Chriment**, qui sans même me connaître, a fait confiance à ceux qui m'ont fait confiance, à mon Directeur de thèse en l'occurrence que je remercie encore une fois, et qui m'a admise au sein de son équipe, et m'a offert la chance de poursuivre des études doctorales.

Je remercie également **M. Mustapha Baziz** pour sa coopération scientifique, pour son aide précieuse et pour sa gentillesse. Mes remerciements aussi à **Mariam Daoud**, doctorante au sein de l'équipe SIG, pour son aide, sa disponibilité et pour sa gentillesse exemplaire.

Je n'oublierai pas les aides reçues du personnel administratif et du service informatique de l'IRIT, ni les sourires accueillants et la convivialité des membres de l'équipe SIG.

J'ai sûrement oublié de remercier beaucoup d'autres personnes méritantes, des personnes qui m'ont offert leur amitié, qui m'ont ouvert leur cœur, qui m'ont ouvert leur porte, qu'elles trouvent ici l'expression de ma profonde gratitude et de mon amitié la plus sincère.

Je tiens à remercier également mes collègues et amis de l'université de Tizi-Ouzou (Algérie) en les personnes de **M. Samir Redaoui**, **M. Yassine Djouadi** et **Mme Samia Fellag** grâce à qui j'ai pu effectuer un ultime séjour à l'IRIT durant lequel j'ai boosté ce travail de thèse.

Merci également à mes amies **Malika** et **Soraya** pour leur aide et leur soutien. Merci à vous d'avoir été là pour moi à un moment crucial de cette thèse.

Mes remerciements finaux et non les moindres vont à mon mari qui a supporté mes humeurs au gré de cette thèse, qui m'a aidée sur les nombreux fronts de la vie quotidienne, qui m'a encouragée jusqu'au bout, qui m'a remplacée auprès de mes enfants les fois où j'ai du m'absenter, qui a été mon appui tout simplement.

Merci aussi à vous mes enfants pour avoir compris, malgré votre jeune âge, les impératifs de cette thèse pour moi et pour m'avoir aidée à y arriver.

Résumé

Ce travail de thèse adresse deux principaux problèmes en recherche d'information : (1) la formalisation automatique des préférences utilisateur, (ou la pondération automatique de requêtes) et (2) l'indexation sémantique.

Dans notre première contribution, nous proposons une approche de recherche d'information (RI) flexible fondée sur l'utilisation des CP-Nets (*Conditional Preferences Networks*). Le formalisme CP-Net est utilisé d'une part, pour la représentation graphique de requêtes flexibles exprimant des préférences qualitatives et d'autre part pour l'évaluation flexible de la pertinence des documents. Pour l'utilisateur, l'expression de préférences qualitatives est plus simple et plus intuitive que la formulation de poids numériques les quantifiant. Cependant, un système automatisé raisonnerait plus simplement sur des poids ordinaux. Nous proposons alors une approche de pondération automatique des requêtes par quantification des CP-Nets correspondants par des valeurs d'utilité. Cette quantification conduit à un UCP-Net qui correspond à une requête booléenne pondérée. Une utilisation des CP-Nets est également proposée pour la représentation des documents dans la perspective d'une évaluation flexible des requêtes ainsi pondérées.

Dans notre seconde contribution, nous proposons une approche d'indexation conceptuelle basée sur les CP-Nets. Nous proposons d'utiliser le formalisme CP-Net comme langage d'indexation afin de représenter les concepts et les relations conditionnelles entre eux d'une manière relativement compacte. Les noeuds du CP-Net sont les concepts représentatifs du contenu du document et les relations entre ces noeuds expriment les associations conditionnelles qui les lient. Notre contribution porte sur un double aspect : d'une part, nous proposons une approche d'extraction des concepts en utilisant WordNet. Les concepts résultants forment les noeuds du CP-Net. D'autre part, nous proposons d'étendre et d'utiliser la technique de règles d'association afin de découvrir les relations conditionnelles entre les concepts noeuds du CP-Nets. Nous proposons enfin un mécanisme d'évaluation des requêtes basé sur l'appariement de graphes (les CP-Nets document et requête en l'occurrence).

Mots clés : Recherche d'information flexible, pondération des requêtes, indexation sémantique, WordNet, Règles d'association, CP-Nets.

Table des matières

Introduction générale.....	17
Contexte.....	17
Problématique.....	17
Contribution.....	19
Publications dans le cadre de la thèse.....	21
Organisation du mémoire	22
PARTIE 1 De la RI classique à la RI sémantique.....	25
Chapitre 1 Recherche d'information.....	27
1.1 Introduction	27
1.2 La RI classique	27
1.2.1 L'indexation	28
1.2.2 Taxonomie des modèles de RI	32
1.2.3 Reformulation de requêtes.....	40
1.2.4 Conclusion	41
1.3 La RI flexible.....	42
1.3.1 Indexation floue des documents	43
1.3.2 Formulation de requêtes flexibles	43
1.3.3 Evaluation flexible des requêtes.....	46
1.4 Evaluation d'un SRI	48
1.4.1 La campagne d'évaluation TREC.....	50
1.4.2 Autres mesures d'évaluation d'un SRI.....	52
1.4 Conclusion	53
Chapitre 2 Indexation sémantique en RI	55
Introduction	55
2.2 Problématique.....	55
2.3 L'indexation conceptuelle	58
2.4 L'indexation sémantique basée sur la désambiguïisation.....	59
2.4.1 Les approches de désambiguïisation des sens des mots (WSD)	60
2.4.2 Les approches d'indexation sémantique.....	66
2.4 Conclusion.....	71
PARTIE 2 Modèles de RI flexibles basés sur les CP-Nets	73
Chapitre 3 Modèle de RI flexible basé sur les CP-Nets.....	75

3.1 Introduction	75
3.2 Problématique et motivations	76
3.2 Les CP-Nets	78
3.2.1 Notations et définitions préliminaires	78
3.2.2 Le modèle CP-Net	80
3.2.3 Les UCP-Nets	83
3.3 Modèle de RI basé CP-Nets	85
3.3.1 Représentation CP-Net des requêtes préférentielles.....	86
3.3.2 Pondération automatique de la requête.....	87
3.3.3 Evaluation de la requête CP-Net	90
3.4 Conclusion	96
Chapitre 4 Approche de RI sémantique	97
4.1 Introduction	97
4.2 Motivations	98
4.3 Les outils d'aide à l'indexation sémantique	100
4.3.1 WordNet	100
4.3.2 Les règles d'association.....	103
4.4 Approche d'indexation sémantique	110
4.4.1 Aperçu général.....	110
4.4.2 Identification de concepts représentatifs du document.....	112
4.4.3 Découverte des relations entre concepts.....	118
4.4.4 Construction de l'index conceptuel du document	120
4.4.5 Illustration.....	121
4.5 Evaluation des requêtes basée CP-Nets.....	125
4.5.1 Définition formelle	125
4.5.2 Illustration.....	127
4.6 Évaluation expérimentale	130
4.6.1 Collection Muchmore	130
4.6.2 Protocole d'évaluation.....	132
4.6.3 Résultats expérimentaux.....	133
4.6.4 Conclusion	135
Conclusion générale.....	137
Synthèse.....	137
Perspectives	140
Validation expérimentale :	140
Améliorations futures	141
ANNEXES	169
Annexe A Evaluation des techniques de désambiguïsation	169
Annexe B Les CP_Nets.....	171
B.1 Introduction.....	173
B.2 Description avancée	175

B.2.1 Un exemple illustratif	175
B.2.2 La sémantique du CP-Net	176
B.2.3 Raisonner avec les CP-Nets	177
B.2.4 Utilisation des graphes CP-Nets	179
Annexe C Les règles d'association en RI.....	181
C.1 Introduction.....	183
C.2 Extraction de connaissances dans les bases de données (ECBD).....	183
C.2.1 Généralités	183
C.2.2 Le Data Mining (DM).....	184
C.3 Extraction de connaissances dans les bases de données textuelles (ECT).....	185
C.3.1 Introduction.....	185
C.3.2 La fouille de texte	186
C.3.3 Découverte de règles d'association.....	190
C.4 CONCLUSION.....	204

Table des figures

FIGURE 1.1 : Processus en U de la RI	28
FIGURE 1.2 : Taxonomie des modèles en RI	33
FIGURE 1.3 : Distribution des documents dans une collection face à une requête.....	51
FIGURE 2.1 : Un exemple de taxonomie conceptuelle	58
FIGURE 2.2 : Exemple de voisinage du mot house.....	69
FIGURE 3.1: Un exemple de CP-Net	81
FIGURE 3.2 : Graphe de préférences induit.	82
FIGURE 3.3 : Un exemple de UCP-Net.	84
FIGURE 3.4 : Famille étendue de X	84
FIGURE 3.5 : Représentation CP-Net d'une requête booléenne	87
FIGURE 3.6 : L'UCP-Net requête	89
FIGURE 3.7 : D_1 vu comme un UCP-Net.....	92
FIGURE 4.1: Sous hiérarchie de WordNet correspondant au concept "dog"	102
FIGURE 4.2 : Extraction des itemsets fréquents.....	107
FIGURE 4.3 : Les étapes de l'indexation conceptuelle basée CP-Nets	111
FIGURE 4.4 : Identification du contexte relatif d'un mot dans d.....	114
FIGURE 4.5 : Identification des termes	114
FIGURE 4.6 : Le CP-Net document.....	124
FIGURE B.1 : Le CP-Net.....	176
FIGURE B.2 : Exemple de CP-Net	181
FIGURE B.3 : Exemple de reconfiguration du contenu.....	182
FIGURE C.1 : Treillis des parties associé à I.....	191

Liste des tableaux

TABLEAU 1.1 : Distribution de probabilités de pertinence des termes d'un corpus d'apprentissage	39
TABLEAU 3.1 : Document retourné	92
TABLEAU 3.2 : Sous-requêtes conjonctives.....	94
TABLEAU 3.3 : Pertinences partielles et totale du document D1	96
TABLEAU 4.1 : Les concepts de WordNet correspondants au mot dog.....	101
TABLEAU 4.2 : Le nombre de mots et de synsets dans WordNet 3.0	101
TABLEAU 4.3 : pseudo-code de l'algorithme Apriori.....	106
TABLEAU 4.4 : Base transactionnelle D, avec 4 des transactions T_i	107
TABLEAU 4.5 : Règles d'association à 1 item en conséquent.	108
TABLEAU 4.6 : Règles d'association à 1 item en conséquent.	108
TABLEAU 4.7 : Règles d'association à 2 items en conséquent.....	108
TABLEAU 4.8 : Génération des k-itemsets fréquents	122
TABLEAU 4.9 : Règles d'association générées	122
TABLEAU 4.10 : Confiances des règles.....	123
TABLEAU 4.11 : Règles d'association sélectionnées	123
TABLEAU 4.12 : Supports des règles d'association sémantiques	124
TABLEAU 4.13 : Calcul de similarité entre les CP-Nets document et requête.....	130
TABLEAU 4.14 : Résultats d'évaluation de la méthode de détection de concepts	134
TABLEAU 4.15 : Résultats d'évaluation de la méthode de pondération de concepts : impact de la méthode d'indexation par les concepts	136

Introduction générale

Contexte

Le but principal d'un système de recherche d'information (SRI) est de retrouver les documents pertinents en réponse à une requête utilisateur. Ces documents sont typiquement retournés sous forme d'une liste ordonnée, où l'ordre est basé sur des estimations de pertinence. Le modèle de recherche pour un SRI indique comment les documents et requêtes sont représentés et comment ces représentations sont comparées pour évaluer la pertinence. Les SRI classiques représentent les documents et les requêtes par les mots qu'ils contiennent, et basent souvent cette comparaison sur le nombre de mots qu'ils ont en commun, c'est l'appariement lexical. Dans cette approche, des documents pertinents, ne partageant pas de mots avec la requête ne sont pas retrouvés. Tandis que des documents non pertinents, contenant des mots de la requête sont retournés à l'utilisateur. Ces problèmes sont dus au fait que l'appariement lexical ne tient pas compte des sens des mots du document et de la requête. L'indexation sémantique tente de pallier ces problèmes en offrant le moyen de distinguer ces sens, et de les utiliser lors du processus d'appariement. Notre travail s'inscrit principalement dans ce contexte. En particulier, nous proposons une approche de RI sémantique basée sur l'indexation des documents et requêtes, par les sens des mots plutôt que par les mots eux-mêmes. L'approche offre en outre le moyen de prendre en compte les préférences utilisateur sur les critères de recherche, et d'évaluer la pertinence d'un document pour une requête en tenant compte de ces préférences.

Problématique

Dans les SRI classiques, l'évaluation de la pertinence d'un document pour une requête est basée sur la mesure de correspondance du document pour la requête. Plus la requête et le document ont de mots en commun, plus le document est considéré comme étant pertinent. Ces systèmes présentent des insuffisances à différents niveaux : au niveau du langage de requête, de la représentation des documents et requêtes et de l'appariement.

Au niveau du langage de requêtes : une requête traduit le besoin en information de l'utilisateur mais aussi ses préférences sur les informations recherchées. La pondération des termes de la requête par des poids numériques [Buell et al., 81; Bordogna et al., 91; Pasi, 99] a permis d'exprimer les préférences utilisateur sur les critères de recherche. Cependant, les poids numériques des requêtes forcent l'utilisateur à quantifier le concept qualitatif et vague d'importance. Cette tâche n'est pas simple, en particulier si la requête exprime des préférences conditionnelles. D'une part, car il n'existe pas une bonne méthode pour pondérer correctement les termes de la requête, d'autre part, lorsque le nombre de valeurs sur lesquelles portent les préférences est élevé, il est quasiment impossible d'énumérer un poids valide pour tous les termes de la requête. Ces problèmes sont d'autant plus accrus que la requête exprime des préférences conditionnelles. Même si ce type de préférences n'est pas spécifiquement pris en charge par les SRI, il est toujours possible de les traduire en expressions booléennes. Cependant, une pondération aléatoire ou intuitive de telles requêtes préférentielles, peut conduire à des énoncés qui sont complètement contradictoires avec la sémantique qu'elles tentent d'exprimer. De ce fait, pour pallier les difficultés de la pondération numériques des requêtes, des travaux se sont orientés vers l'utilisation de préférences qualitatives plus simples et plus intuitives, formulées à partir de termes linguistiques tels que : *important, très important...* [Bordogna et al., 93; Bordogna et al., 95]. Cependant, le problème de la définition des poids numériques des termes est reporté sur la définition de la sémantique du concept flou *important* et des modulateurs linguistiques *très, peu, moyennement...*

Au niveau de la représentation des documents et requêtes, et de l'appariement correspondant : dans les SRI classiques, documents et requêtes sont représentés comme des listes de mots clés, généralement pondérés. L'appariement document-requête est lexical et se base sur la présence ou l'absence d'un mot de la requête dans le document. Or il est bien connu que les mots de la langue sont ambigus. Un même mot peut désigner différents concepts (et donc exprimer différents sens) et différents mots peuvent avoir une même signification. L'appariement lexical ne considère pas ces aspects. De ce fait, des documents pourtant pertinents, contenant des mots sémantiquement équivalents mais lexicalement différents (synonymes) des mots de la requête, ne seront pas retrouvés. Par ailleurs, des documents non pertinents, contenant des mots lexicalement identiques mais sémantiquement différents (homonymes) des termes de la requête seront retournés à l'utilisateur. L'indexation sémantique (ou indexation par les sens des mots) tente de pallier les problèmes de l'appariement lexical en utilisant pour la recherche, des index conceptuels ou sémantiques au lieu de simples mots clés. De tels index portent sur la sémantique des mots. Ils sont construits à partir (1) des concepts explicites des textes eux-mêmes (indexation conceptuelle), (2) de la sémantique latente des textes des documents (indexation par sémantique latente LSI [Deerwester et al., 90]), ou (3) extraits de la sémantique explicite des mots telle que définie dans les dictionnaires, thésaurus ou ontologies (indexation sémantique). L'approche d'indexation par la sémantique latente résout les sens des mots par un *clustering* des mots

INTRODUCTION GENERALE

sémantiquement proches via une technique de réduction de la dimensionnalité de la matrice termes-documents. L'indexation conceptuelle tente à partir d'une taxonomie conceptuelle extraite du texte, de construire sa sémantique. Les liens entre les différents concepts d'une telle taxonomie sont des liens fonctionnels entre entités lexicales. L'indexation sémantique tente de retrouver, parmi les différents sens possibles d'un mot tels que définis dans les dictionnaires, ontologies et autres ressources linguistiques, le sens correct du mot dans le texte à indexer.

Les mots d'un texte donné définissent implicitement une sémantique orientée sujet (topic), du texte correspondant. Le sujet principal du document est porté par les termes les plus importants. Des sujets connexes secondaires sont portés par des termes moins importants, qui s'agencent dans le document en fonction de la sémantique même du topic du texte. Cet agencement des mots (et surtout des sens) dans le texte des documents définit une dimension sémantique du document orientée topic. Or, les approches d'indexation sémantique classiques ignorent cette dimension.

Nous nous sommes intéressés aux problèmes particuliers posés par la pondération des termes de la requête, la représentation basée mots-clés des documents et requêtes et l'appariement lexical, et avons proposé des techniques et méthodes pour tenter d'y remédier. Notre contribution globale consiste en la définition de deux nouvelles approches de recherche d'information (RI) flexible basées sur les CP-Nets. Chacune des deux approches proposées focalise sur les trois aspects d'un SRI : la représentation des documents (indexation), le langage de requêtes et l'évaluation.

Contribution

Notre première contribution consiste en un nouveau modèle de RI flexible basé sur les CP-Nets. Dans ce modèle, nous avons :

1. introduit un nouveau langage de requêtes exprimant les préférences qualitatives de l'utilisateur. La spécificité de ce langage concerne la prise en charge intuitive des préférences conditionnelles. Pour cela, nous exploitons les CP-Nets pour la représentation de telles requêtes préférentielles conditionnelles.
2. proposé un algorithme pour la pondération automatique de requêtes qualitatives. L'algorithme se base sur le formalisme UCP-Net, extension des CP-Nets par des valeurs numériques de préférences (dites valeurs d'utilités). L'utilisateur est ainsi déchargé de cette lourde et non moins improbable tâche, et les poids produits sont certifiés corrects puisque basés sur les fondements théoriques des UCP-Nets.

3. défini une approche de représentation des documents par des CP-Nets. L'approche est basée sur la projection des documents sur chaque requête soumise. Le document est alors représenté par un CP-Net de même topologie que celui de la requête, facilitant ainsi l'évaluation de la pertinence
4. proposé une approche d'évaluation des requêtes basée sur le paradigme booléen. Nous exploitons pour cela l'interprétation des CP-Nets dans le formalisme booléen, puis l'évaluation de la requête booléenne obtenue, au moyen de l'opérateur d'agrégation du *minimum pondéré* [Dubois et al., 86; Yager, 87].

Notre seconde contribution se rapporte à la RI sémantique. Cette approche est proposée comme amélioration de notre première contribution, au niveau de la représentation des documents et au niveau de l'évaluation. En particulier, le modèle de RI proposé s'affranchit d'une part des limites de la représentation basées mots-clés en proposant une approche d'indexation sémantique, d'autre part des limites de l'appariement lexical et du paradigme booléen en proposant un appariement entre représentations sémantiques des documents et requêtes. En particulier:

1. notre approche d'indexation sémantique a pour objectif d'améliorer la représentation des documents, en se basant sur les sens des mots dans les textes correspondants, et sur les liens entre ces sens dans le texte du document considéré. Notre approche est fondée sur deux étapes principales : une première étape d'extraction des sens des termes d'indexation et une seconde étape de découverte des relations entre ces sens. La première étape se base sur l'utilisation de l'ontologie linguistique WordNet pour identifier, pondérer et désambiguïser les sens des termes d'indexation. La seconde étape est fondée sur l'utilisation des règles d'association pour dériver des relations de dépendance contextuelle entre les concepts menant à une représentation plus expressive des documents. Le principe même de l'approche n'est pas nouveau mais nous avons proposé :

- une nouvelle technique pour identifier les termes d'indexation (simples ou composés) par projection sur l'ontologie WordNet,
- une nouvelle approche de pondération des termes d'indexation tenant compte de la sémantique des termes,
- une approche de désambiguïisation des sens des mots basée sur la notion de distance sémantique et tenant compte de l'importance du mot dans le texte,
- une nouvelle technique pour découvrir des relations entre les concepts correspondants au moyen des règles d'association sémantiques proposées. Les règles d'association sémantiques permettent de découvrir des relations contextuelles entre les concepts amenant à une représentation plus expressive du document.

INTRODUCTION GENERALE

2. Notre approche d'évaluation des requêtes a pour objectif d'évaluer la pertinence des documents et des requêtes représentés par des graphes CP-Nets. La requête CP-Net dérive de notre approche proposée en première contribution. Tandis qu'un CP-Net document est construit à partir de l'index conceptuel du document, issu de l'étape d'indexation sémantique présentée en (1), en organisant l'ensemble formé des concepts représentatifs du document d'une part et des associations correspondantes d'autre part, en un graphe conditionnel, le CP-Net document. L'approche d'évaluation proposée calcule alors la pertinence d'un document pour une requête donnée sur la base d'une mesure proposée de similitude des graphes CP-Nets correspondants.

L'ensemble des techniques ainsi définies constitue alors la base théorique de notre modèle de RI sémantique basée sur les CP-Nets.

3. La validation expérimentale de notre approche concerne principalement deux aspects :

- la validation de l'approche de détection de concepts
- la validation de l'approche de pondération
- la validation du modèle de RI basé sur les CP-Nets
- la validation de l'approche d'indexation sémantique dont :
 - validation de l'approche de désambiguïsation
 - validation de l'approche d'extraction des règles d'association sémantiques

La validation expérimentale de notre modèle de RI proposé dans notre première contribution, nécessite l'utilisation d'un cadre d'évaluation supportant des requêtes CP-Nets. Un tel environnement est à notre connaissance inexistant. Sa construction relève d'un travail de recherche à part entière, qui indépendamment du modèle proposé peut servir de base à la prise en compte des préférences conditionnelles dans le processus de RI.

Nous avons par ailleurs testé notre approche d'indexation sémantique. Les résultats expérimentaux obtenus ont montré l'intérêt d'une indexation sémantique par les concepts de WordNet. Nous n'avons cependant pas expérimenté le modèle dans sa totalité. En particulier, la technique de découverte des règles d'association sémantiques n'a pas été expérimentée.

Publications dans le cadre de la thèse

1. Dans le cadre de notre proposition d'un modèle de RI flexible basé sur les CP-Nets

1. [Fatiha Boubekeur](#), [Lynda Tamine](#). Recherche d'Information flexible basée CP-Nets. Dans : Conférence francophone en Recherche d'Information et Applications (CORIA 2006), Lyon, 15/03/06-17/03/06,

Association Francophone de Recherche d'Information et Applications (ARIA), p. 161-166, mars 2006.

2. [Fatiha Boubekur](#), [Mohand Boughanem](#), [Lynda Tamine](#). Towards Flexible Information Retrieval Based on CP-Nets. Dans : Flexible Query Answering (FQAS 2006), Milan, Italie, 07/01/06-10/06/06, Henrik Legind Larsen, Gabriella Pasi, Daniel Ortiz-Arroyo (Eds.), [World Scientific Publishing](#), Advances in Artificial Intelligence, p. 222-231, juin 2006.

[Lynda Tamine](#), [Fatiha Boubekur](#), [Mohand Boughanem](#). On Using Graphical Models for Supporting Context-Aware Information Retrieval. Dans : International Conference on the Theory of Information Retrieval (ICTIR 2007), Budapest (Hungary), 18/10/07-20/10/07, [Foundation for Information Society](#), p. 213-222, octobre 2007.

2. Dans le cadre de notre proposition de modèle de RI sémantique basé sur les CP-Nets

[Fatiha Boubekur](#), [Mohand Boughanem](#), [Lynda Tamine](#). *Semantic Information Retrieval Based on CP-Nets*. Dans : IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2007), London, 23/07/07-26/07/07, [IEEE](#), (support électronique), juillet 2007.

[Fatiha Boubekur](#), [Mohand Boughanem](#), [Lynda Tamine](#). *Une approche d'indexation conceptuelle de documents basée sur les graphes CP_Nets*. Dans: cinquième édition du colloque sur l'optimisation et les systèmes d'information COSI'08, 8-10 juin 08, Tizi-Ouzou, Algérie.

6. [Fatiha Boubekur](#), [Mohand Boughanem](#), [Lynda Tamine](#). *Exploiting association rules and ontology for semantic document indexing*. Dans: 12th International conference IPMU08, Information Processing and Management of Uncertainty in knowledge-Based Systems, Malaga, 22-27, june 08, Spain.

Organisation du mémoire

Ce mémoire est organisé en deux parties principales. La première partie, composée de deux chapitres, est dédiée à la présentation de la RI classique (chapitre 1) et de l'indexation sémantique (chapitre 2). La seconde partie présente nos contributions. Elle est divisée en deux

INTRODUCTION GENERALE

chapitres 3 et 4, dédiés respectivement à la présentation de notre modèle de RI flexible et de notre modèle de RI sémantique basés sur les CP-Nets. Le détail de cette organisation est donné comme suit :

Dans le chapitre 1, nous présentons les différents aspects liés à la RI et aux SRI. Nous nous attacherons en particulier à définir les modèles de recherche tant classiques (booléen, vectoriel et probabiliste) que des modèles plus flexibles (extensions floues du modèle booléen). L'indexation automatique est aussi explicitée ainsi que les mécanismes de raffinement des requêtes. Enfin, nous rappellerons les mesures d'évaluation courantes d'un SRI et présenterons les techniques d'évaluation des SRI mises en œuvre dans le cadre des campagnes d'évaluation.

Le chapitre 2 sera dédié à la présentation des approches d'indexation sémantique des documents. L'objectif de telles approches est d'indexer les documents par les sens des mots ou par les concepts, plutôt que par les mots eux-mêmes. Le but est de pallier les problèmes d'appariement lexical des SRI classiques et pouvoir ainsi traiter avec l'ambiguïté naturelle de la langue. L'approche d'indexation basée sur les concepts (indexation conceptuelle) est présentée en section 2.2. Les approches d'indexation basées sur les sens des mots (indexation sémantique) sont présentées en section 2.3. Ces dernières s'appuient sur des techniques linguistiques de désambiguïsation des sens des mots. Nous dédions alors la section 2.3.1 à la présentation des travaux en désambiguïsation linguistique avant de présenter l'état de l'art sur l'indexation sémantique en section 2.3.2.

En chapitre 3, nous présentons notre première contribution à la définition d'un modèle de RI flexible basé sur les CP-Nets. Le chapitre s'articule autour de trois sections. La section 3.1 présente nos motivations. En section 3.2, nous présentons le formalisme CP-Net sur lequel se basent nos modèles. La section 3.3 présente notre modèle de RI basé sur les CP-Nets. En particulier, nous y définissons (1) notre approche de pondération automatique de requêtes qualitatives, (2) la technique de représentation CP-Net des documents et (3) notre méthode d'évaluation des requêtes CP-Nets.

Le chapitre 4, présente notre seconde contribution à travers un nouveau modèle de RI sémantique basé sur les CP-Nets. Le chapitre s'articule autour de 4 sections. En section 4.2, nous présentons les motivations qui ont été à l'origine de nos propositions. En section 4.3, nous présentons les outils sur lesquels est basée notre approche d'indexation sémantique, à savoir WordNet et les règles d'association. Les fondements théoriques de notre approche d'indexation sémantique, un exemple illustratif ainsi que quelques résultats expérimentaux sont donnés en section 4.4. La section 4.5 présente notre approche d'évaluation des requêtes CP-Nets. Quelques résultats expérimentaux sont donnés en section 4.6.

Enfin, en conclusion générale, nous présentons les perspectives de nos présentes propositions.

PARTIE 1

De la RI classique à la RI sémantique

Chapitre 1

Recherche d'information

1.1 Introduction

La recherche d'information (RI) traite de la représentation, du stockage, de l'organisation et de l'accès à l'information. Le but d'un système SRI est de retrouver, parmi une collection de documents préalablement stockés, les documents qui répondent au besoin utilisateur exprimé sous forme de requête. Pour cela, un SRI met en oeuvre un ensemble de processus de sélection des documents pertinents pour la requête.

Le but de ce chapitre est de présenter les concepts de base de la RI. Dans une première partie, nous nous intéressons aux approches de RI classique, puis nous présentons les approches de RI basées sur la logique floue.

Ce chapitre est organisé comme suit : en section 1.2, nous présentons les concepts de base de la RI classique. Nous y décrivons notamment le processus d'indexation en paragraphe 1.2.1, puis la taxonomie de modèles en paragraphe 1.2.2. Les techniques de reformulation des requêtes sont présentées en paragraphe 1.2.3. Le paragraphe 1.2.4 présente les outils et méthodes d'évaluation d'un SRI. En section 1.3, nous présentons les approches de RI flexible.

1.2 La RI classique

De manière générale, la recherche dans un SRI consiste à comparer la représentation interne de la requête aux représentations internes des documents de la collection. La requête est formulée, par l'utilisateur, dans un langage de requêtes qui peut être le langage naturel, un langage à base de mots clés ou le langage booléen. Elle sera transformée en une représentation interne équivalente, lors d'un processus d'interprétation. Un processus similaire, dit indexation, permet de construire la représentation interne des documents de la base documentaire. Le processus de recherche consiste alors à mettre

en correspondance et à calculer le degré d'appariement des représentations internes des documents et de la requête. Les documents qui correspondent au mieux à la requête, ou documents dits pertinents, sont alors retournés à l'utilisateur, dans une liste ordonnée par ordre décroissant de degré de pertinence lorsque le système le permet. Afin d'améliorer les résultats de la recherche, le système peut être doté d'un mécanisme d'amélioration et de raffinement de la requête par reformulation.

Le fonctionnement général d'un SRI est donnée au travers du processus de recherche communément appelé processus en U [Belkin et al., 92], présenté en figure 1.1.

Ce processus fait ressortir trois mécanismes de base : le processus d'indexation (quelques fois dit processus d'interprétation pour les requêtes), le processus de recherche et le processus de reformulation des requêtes. Nous les détaillons dans les paragraphes suivants.

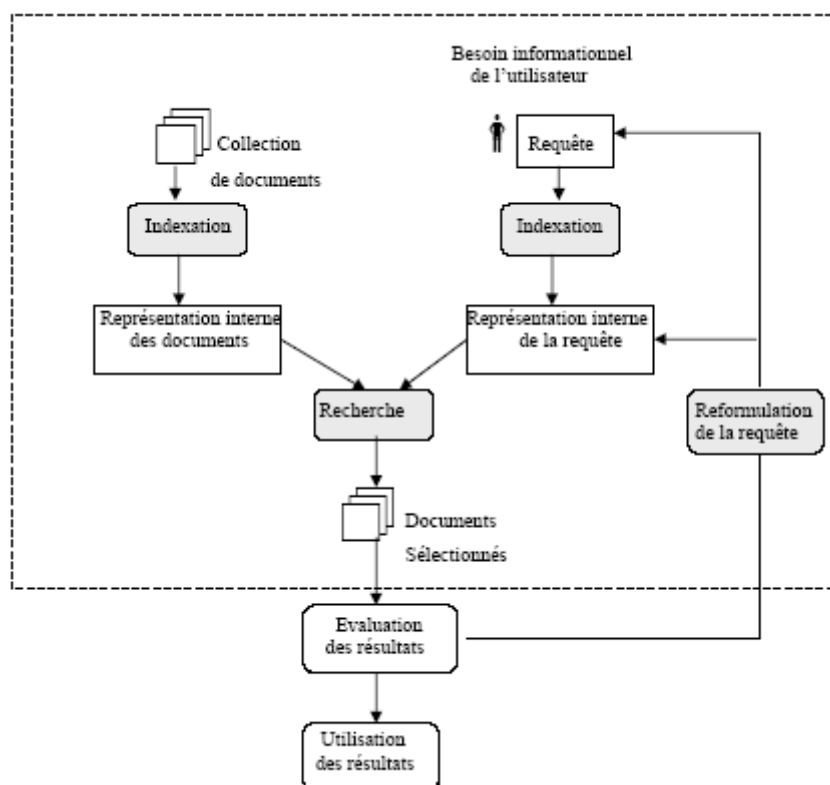


FIGURE 1.1 : Processus en U de la RI

1.2.1 L'indexation

L'indexation est une phase très importante pour un SRI car de sa qualité dépend la qualité des réponses du système et donc les performances de ce dernier. Une bonne indexation doit permettre de retrouver tous les documents pertinents au besoin de l'utilisateur et pas (ou peu) de documents non pertinents pour celui-ci.

CHAPITRE 1. RECHERCHE D'INFORMATION

En phase d'indexation, le document (ou la requête) est analysé(e) et les mots clés caractérisant son contenu informationnel, sont extraits. Un mot clé peut-être soit un mot simple ou un groupe de mots. Les mots-clés descriptifs du contenu sémantique d'un document sont dits termes d'indexation. L'ensemble de tous les termes d'indexation constitue le langage d'indexation. Ce langage peut être libre ou contrôlé. Un langage d'indexation libre est construit à partir des termes extraits du document analysé. Un langage d'indexation contrôlé est construit à partir d'un ensemble de termes préalablement définis et organisés généralement dans un thésaurus. Lorsqu'un document est analysé, on ne garde que les mots clés qui appartiennent à ce thésaurus.

1.2.1.1 Approches d'indexation

Techniquement, l'indexation peut-être manuelle, automatique ou semi-automatique [Salton, 88; Salton et al., 88].

En indexation manuelle, c'est un opérateur humain, généralement expert du domaine, qui se charge de caractériser, selon ses connaissances propres, le contenu sémantique d'un document. Cette approche présente deux inconvénients :

1. elle est subjective, puisque le choix des termes d'indexation dépend de l'indexeur et de ses connaissances du domaine,
2. elle est pratiquement inapplicable aux corpus de textes volumineux.

Néanmoins, tel que rapporté dans [Savoy, 05], elle est plus performante que l'indexation automatique en termes de précision *moyenne* des documents retrouvés en réponse à une requête utilisateur donnée.

En indexation automatique [Luhn, 57; Maron, 60; Salton, 68], c'est un processus complètement automatisé qui se charge d'extraire les termes caractéristiques du document. L'intérêt d'une telle approche réside dans sa capacité à traiter les textes nettement plus rapidement que l'approche précédente, et de ce fait, elle est particulièrement adaptée aux corpus volumineux. L'indexation automatique est l'approche la plus étudiée en RI, nous la détaillons en section suivante.

L'indexation semi-automatique [Maniez et al., 91; Balpe et al., 95; Jacquemin et al., 02], appelée aussi indexation supervisée, est une combinaison des deux approches d'indexation précédentes. Dans ce cas, les indexeurs utilisent un vocabulaire contrôlé sous forme de thésaurus ou de base terminologique. Le choix final des termes d'indexation à partir du vocabulaire fourni, est laissé ainsi à l'indexeur humain (généralement spécialiste du domaine).

Dans la section suivante, nous nous intéressons particulièrement à l'approche d'indexation automatique, plus répandue, puisque c'est celle qui nous intéresse dans le cadre de notre travail.

1.2.1.2 L'indexation automatique

L'indexation automatique classique est fondée sur l'analyse des documents en vue de l'extraction des termes (mots-clés simples ou composés) représentatifs de leur contenu informationnel. Elle repose sur les étapes suivantes : l'extraction des termes d'indexation, la réduction du langage d'indexation et la pondération des termes d'indexation.

1. *L'extraction des termes d'indexation* repose sur une analyse linguistique du texte du document. Plusieurs niveaux d'analyse peuvent être distingués : le niveau lexical, syntaxique et sémantique.
 - En analyse lexicale, les mots composant le texte sont extraits et les mots vides (prépositions, pronoms personnels,...) éliminés. Une étape supplémentaire peut être nécessaire en vue d'éliminer les variantes morphologiques (genre, nombre, dérivations, ...) des mots. Le traitement associé repose sur deux procédures : la lemmatisation et la troncature (ou racinisation). La racinisation consiste à supprimer le suffixe (et plus rarement le préfixe) des mots significatifs du texte indexé. La lemmatisation (*stemming* en anglais) a pour objectif de prendre la forme canonique du mot. Des expériences ont montré que la racinisation et la lemmatisation améliorent significativement les performances pour les langues riches morphologiquement (ex. le français, l'italien, etc.) [Gaussier et al., 1997; Gaussier et al., 2000].
 - En analyse syntaxique, il s'agit de repérer les groupes de mots ou des mots composés [Fagan, 87; Salton, 88]. L'utilisation des termes composés doit permettre d'augmenter la précision de réponse dans la mesure où le critère possède une signification plus précise et un usage plus restreint (plus spécifique) que les mots qui le composent. Les approches d'analyse syntaxique se basent en général sur l'utilisation de patrons (*templates*) syntaxiques [Bourigault, 96; Aussenac-Gilles et al., 00; Jacquemin, 01; Jones et al., 02] pour détecter les termes composés.
 - L'analyse sémantique s'intéresse à reconnaître les sens des mots, les mots synonymes, les concepts représentatifs de ces mots, et plus généralement les relations sémantiques entre les mots. Le chapitre 2 sera dédié à cette dernière approche (i.e. l'indexation sémantique) puisque c'est celle qui nous intéresse dans le cadre de notre travail de thèse.

CHAPITRE 1. RECHERCHE D'INFORMATION

La réduction du langage d'indexation vise à réduire le nombre de termes d'indexation en éliminant tous les mots non importants (mots rares ou mots trop fréquents) du langage d'indexation. Pour mesurer l'importance d'un mot dans un document, l'indexation s'appuie sur la fréquence d'occurrence de ce mot dans le document. Les mots de fréquences quasi nulles et les mots à fréquences trop élevées peuvent être éliminés de l'index. Cette hypothèse tire ses origines de la conjecture de Luhn [Luhn, 58] qui, pratiquement, définit un seuil de fréquence minimal S_{min} et un seuil de fréquence maximal S_{max} tels que, tout terme d'indexation t de fréquence intermédiaire ($S_{min} \leq freq(t) \leq S_{max}$), est considéré comme significatif et appartient donc au langage d'indexation.

La pondération des termes d'indexation consiste à associer un poids d'importance (ou valeur de représentativité) w_{ij} à chaque terme t_j d'un document d_i . De manière générale, les formules de pondération utilisées sont basées sur la combinaison d'un facteur de pondération local quantifiant la représentativité locale du terme dans le document, et d'un facteur de pondération global quantifiant la représentativité globale du terme vis-à-vis de la collection de documents. Plusieurs formules existent, dont :

$$w_{ij} = \frac{tf_{ij}}{df_j} = tf_{ij} \times \frac{1}{df_j} = tf_{ij} \times idf_j \quad [\text{Salton et al., 73}]$$

Où :

tf_{ij} est la fréquence d'occurrences du terme t_j dans le document d_i .

df_j est la fréquence documentaire du terme t_j (i.e. la proportion de documents de la collection qui contiennent t_j) et idf_j sa fréquence documentaire inverse.

La mesure $tf * idf$ est une bonne approximation de l'importance d'un terme dans un document, particulièrement dans des corpus de documents de tailles intermédiaires. Pour des documents plus longs des normalisations ont été proposées, dont :

- La normalisation pivotée de Singhal [Singhal et al., 96]

$$w_{ij} = \frac{tf_{ij} * idf_j}{1 + \frac{slope}{(1 - slope) * pivot} * \sqrt{\sum_j (tf_{ij} * idf_j)^2}}$$

Où :

tf_{ij} est le nombre d'occurrences du terme t_j dans l'unité documentaire d_i

idf_j est la fréquence documentaire inverse définie classiquement par : $\log(n/N_j)$ tel que n est le nombre de documents de la collection et N_j le nombre de documents indexés par le terme t_j .

pivot est une constante qui représente l'écart nul entre la probabilité de pertinence et la probabilité de sélection des documents.

slope est un facteur de normalisation fixé empiriquement, de sorte à minimiser l'écart entre la pertinence et la sélection.

- La formule de Robertson [Robertson et al., 97]

$$w_{ij} = \frac{tf_{ij} * (K_1 + 1)}{K_1 \left((1-b) + b * \frac{dl_i}{\Delta l} \right) + tf_{ij}}$$

Où :

w_{ij} est le poids du terme t_j dans le document d_i .

K_1 constante qui permet de contrôler l'influence de la fréquence du terme t_j dans le document d_i . Sa valeur dépend de la longueur des documents dans la collection. Le plus souvent, sa valeur est fixée à 1,2.

b constante qui permet de contrôler l'effet de la longueur du document. Sa valeur la plus souvent utilisée est : 0,75.

dl_i est la longueur du document d_i .

Δl est la longueur moyenne des documents dans la collection entière.

1.2.2 Taxonomie des modèles de RI

Si c'est l'indexation qui permet de déterminer les termes représentatifs des documents et requêtes, c'est le modèle qui assure leur interprétation dans un formalisme de représentation propre et qui offre le mécanisme de leur appariement en vue de calculer les degrés de pertinence des documents pour les requêtes.

La figure 1.2 présente une classification des différents modèles de RI [Baeza-Yates et al., 99]. Les modèles de RI se déclinent en trois grandes catégories qui sont les modèles booléens, les modèles vectoriels et les modèles probabilistes.

Les modèles vectoriels sont des modèles algébriques. Les documents et requêtes sont représentés par des vecteurs de poids dans un espace vectoriel composé de tous les termes d'indexation. La pertinence d'un document vis à vis d'une requête est définie par des mesures de distances entre vecteurs. Plusieurs modèles proposés en RI se basent sur le modèle vectoriel, dont : le modèle connexionniste et le modèle LSI (*Latent Semantic Indexing*). Les modèles probabilistes s'appuient sur la théorie des probabilités. La pertinence d'un document vis à vis d'une requête est vue comme une probabilité de pertinence document/requête. Les modèles de RI basés sur le modèle probabiliste sont le modèle BIR (*Binary Independance retrieval*), le modèle inférentiel Bayésien et le

modèle de langue. Dans ce qui suit, nous décrivons pour chacune de ces classes, le modèle de base et quelques modèles associés.

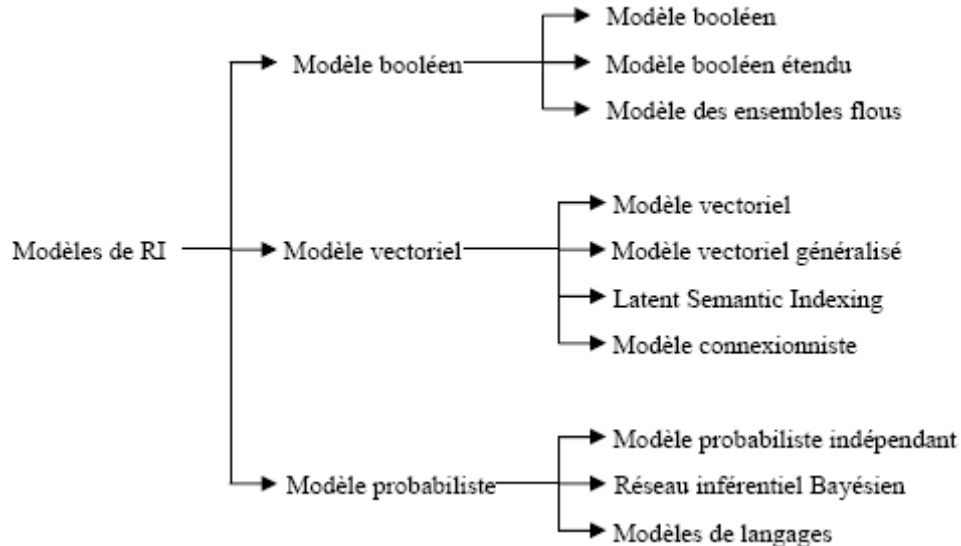


FIGURE 1.2 : Taxonomie des modèles en RI

1.2.2.1 Les modèles booléens

1.2.2.1.1 Le modèle booléen de base

Dans ce modèle, basé sur la théorie des ensembles, le document est représenté par un ensemble de termes. La requête est représentée par un ensemble de mots clés reliés par des opérateurs booléens (AND, OR et NOT). L'appariement requête-document est strict et se base sur des opérations ensemblistes selon les règles suivantes :

$$\begin{aligned}
 RSV(d, t_i) &= 1 \text{ si } t_i \in d, \quad 0 \text{ sinon} \\
 RSV(d, t_i \text{ AND } t_j) &= 1 \text{ si } (t_i \in d) \wedge (t_j \in d), \quad 0 \text{ sinon} \\
 RSV(d, t_i \text{ OR } t_j) &= 1 \text{ si } (t_i \in d) \vee (t_j \in d), \quad 0 \text{ sinon} \\
 RSV(d, \text{NOT } t_i) &= 1 \text{ si } t_i \notin d, \quad 0 \text{ sinon.}
 \end{aligned}$$

Bien que ce modèle soit simple à mettre en oeuvre, il présente néanmoins trois inconvénients majeurs :

- l'appariement est strict et ne permet de classer les documents que dans deux catégories, l'ensemble des documents pertinents et l'ensemble des documents non pertinents, dont les éléments ne sont pas ordonnables, tous les termes d'un document ou d'une requête sont d'égales importances (pondérés à 0 ou 1), ce qui n'est pas le cas en réalité,

- les expressions booléennes ne sont pas accessibles à un large public et des confusions existent du fait de la différence de «sens» des opérateurs logiques AND et OR et de leurs connotations respectives en langage naturel.

Le modèle booléen étendu et modèle basé sur les ensembles flous dérivent du modèle booléen.

1.2.2.1.2 Modèle booléen étendu

Le modèle booléen étendu a été introduit par Salton [Salton et al., 1983]. C'est une extension du modèle précédent qui vise à tenir compte d'une pondération des termes dans le corpus. Cela permet de pallier les problèmes du modèle de base en ordonnant les documents retrouvés par le SRI. La requête demeure une expression booléenne classique. Tandis que les termes d'un document sont maintenant pondérés. En général le poids d'un terme dans un document est fonction du nombre d'occurrences de ce terme dans le document. L'appariement requête_document est le plus souvent déterminé par les relations introduites dans le modèle p -norm basées sur les p -distances, avec $1 \leq p \leq \infty$. La valeur de p est indiquée au moment de la requête. Si m est le nombre de termes dans la requête, les fonctions de similarité se calculent comme suit :

$$RSV(d, Q_{ou}) = \left(\frac{x_1^p + x_2^p + \dots + x_m^p}{m} \right)^{\frac{1}{p}}$$

$$RSV(d, Q_{et}) = 1 - \left(\frac{(1-x_1)^p + (1-x_2)^p + \dots + (1-x_m)^p}{m} \right)^{\frac{1}{p}}$$

Si $p = 1$, on se ramène au modèle booléen.

1.2.2.2 Les modèles vectoriels

1.2.2.2.1 Le modèle vectoriel de base

Dans ce modèle, un document est représenté sous forme d'un vecteur dans l'espace vectoriel composé de tous les termes d'indexation. Les coordonnées d'un vecteur document représentent les poids des termes correspondants. Formellement, un document d_i est représenté par un vecteur de dimension n ,

$$d_i = (w_{i1}, w_{i2}, \dots, w_{in}) \quad \text{pour } i = 1, 2, \dots, m.$$

Où w_{ij} est le poids du terme t_j dans le document d_i ,
 m est le nombre de documents dans la collection,
 n est le nombre de termes d'indexation.

CHAPITRE 1. RECHERCHE D'INFORMATION

Une requête Q est aussi représentée par un vecteur de mots-clés défini dans le même espace vectoriel que le document.

$$Q = (w_{Q1}, w_{Q2}, \dots, w_{Qn})$$

Où w_{Qj} est le poids de terme t_j dans la requête Q . Ce poids peut être soit une forme de $tf*idf$, soit un poids attribué manuellement par l'utilisateur.

La pertinence du document d_i pour la requête Q est mesurée comme le degré de corrélation des vecteurs correspondants. Cette corrélation peut être exprimée par l'une des mesures suivantes :

Le produit scalaire :

$$Sim(d_i, Q) = \sum_{j=1}^n w_{Qj} * w_{ij}$$

La mesure du cosinus:

$$Sim(d_i, Q) = \frac{\sum_{j=1}^n w_{Qj} * w_{ij}}{\left(\sum_{j=1}^n w_{Qj}^2\right)^{1/2} * \left(\sum_{j=1}^n w_{ij}^2\right)^{1/2}}$$

La mesure de Dice :

$$Sim(d_i, Q) = \frac{2 * \sum_{j=1}^n w_{ij} * w_{Qj}}{\sum_{j=1}^n w_{Qj}^2 + \sum_{j=1}^n w_{ij}^2}$$

La mesure de Jaccard :

$$Sim(d_i, Q) = \frac{\sum_{j=1}^n w_{ij} * w_{Qj}}{\sum_{j=1}^n w_{Qj}^2 + \sum_{j=1}^n w_{ij}^2 - \sum_{j=1}^n w_{ij} * w_{Qj}}$$

Le coefficient de superposition :

$$Sim(d_i, Q) = \frac{\sum_{j=1}^n w_{ij} * w_{Qj}}{\min_i \left(\sum_{j=1}^n w_{Qj}^2, \sum_{j=1}^n w_{ij}^2 \right)}$$

L'un des avantages du modèle vectoriel réside dans sa simplicité conceptuelle et de mise en oeuvre. En outre, il permet de trier les résultats d'une recherche à travers une mesure de similarité document/requête, en plaçant en tête les documents jugés les plus similaires à la requête. Cependant, ce modèle ne permet pas de modéliser les associations entre les termes

d'indexation. Chacun des termes est considéré comme indépendant des autres. Le modèle vectoriel généralisé (*Generalized Vector Space Model*) [Wong et al, 1985] permet cependant de résoudre le problème d'indépendance des termes.

1.2.2.2.2 Le modèle connexionniste

Les SRI basés sur l'approche connexionniste utilisent le fondement des réseaux de neurones, tant pour la modélisation des unités textuelles que pour la mise en œuvre du processus de RI. L'idée de base est que la RI est un processus associatif qui peut être représenté par les mécanismes de propagation d'activation des réseaux de neurones. De plus, les capacités d'apprentissage de ces modèles peuvent permettre d'obtenir des SRI adaptatifs.

Deux modèles théoriques ont été utilisés : les modèles à auto-organisation et les modèles à couches.

Les modèles à auto-organisation [Lin et al., 91] permettent à partir de la description des documents, d'en réaliser une classification par l'apprentissage du réseau de neurones. Ces modèles sont basés sur les cartes auto-organisatrices de Kohonen [Kohonen, 89].

Les modèles à couches : Les SRI basés sur un modèle connexionniste à couches [Kwok, 89; Belew, 89; Boughanem, 92a-b; Mothe, 94] sont représentés par un minimum de trois couches de neurones interconnectées : la couche requête (Q), la couche termes (T) et la couche documents (D). Le mécanisme de recherche est basé sur une activation initiale des neurones termes induite par une requête, et qui se propage vers les documents à travers les connexions du réseau. Dans le modèle MERCURE [Boughanem, 92], une requête Q est représentée par un vecteur de poids sous forme :

$$Q_u^{(t)} = (q_{u1}^{(t)}, q_{u2}^{(t)}, \dots, q_{uT}^{(t)})$$

Les poids des termes dans la requête sont affectés aux liens requête-termes. L'activité initiale du réseau correspond à l'activation d'un nœud requête en envoyant un signal de valeur 1 à travers les liens requête-termes. Chaque neurone terme t_j affecté par la requête, reçoit une entrée $In(t_j)$ et fournit une sortie $Out(t_j)$ respectivement définies par :

$$In(t_j) = q_{uj}^{(t)} \quad Out(t_j) = g(In(t_j))$$

Un document d_i qui a des termes t_j en commun avec la requête recevra une entrée $In(di)$ et calculera sa sortie $Out(di)$ telles que :

$$In(d_i) = \sum_{j=1}^T Out(t_j) * w_{ij} \quad Out(d_i) = g(In(d_i))$$

Où w_{ij} est le poids du terme t_j dans le document d_i .

Les valeurs de sortie des différents documents correspondent à leurs degrés de pertinence pour la requête donnée.

1.2.2.2.3 Modèle d'indexation sémantique latente (LSI)

L'objectif du modèle LSI est de construire des index conceptuels portant sur la sémantique des mots dans les documents. Ces index sont tirés à partir de la structure sémantique latente des textes des documents. Pour ce faire, partant de l'espace vectoriel de tous les termes d'indexation, le modèle LSI construit un espace d'indexation de taille réduite k , par application de la décomposition en valeurs singulières (SVD) de la matrice termes-documents [Deerwester et al., 90]. Ces k dimensions capturent une partie importante de la structure sémantique des documents [Berry et al., 94] portée par les associations des termes et documents, et éliminent le bruit dû à la variabilité dans l'usage des mots.

Chaque vecteur document est au final représenté dans l'espace k -dimensionnel réduit des termes non bruités. Les documents qui partagent des termes co-occurents ont des représentations proches. La requête utilisateur est aussi représentée par un vecteur dans l'espace k -dimensionnel. Une mesure de similarité est ensuite calculée entre le k -vecteur requête et chacun des k -vecteurs documents de la collection. A l'issue de la recherche, le système sélectionne les documents pertinents même s'ils ne contiennent aucun mot de la requête.

1.2.2.3 Les modèles probabilistes

1.2.2.3.1 Le modèle probabiliste de base

Le premier modèle probabiliste a été proposé par Maron et Kuhns [Maron et al., 60] au début des années 60. Le principe de base consiste à présenter les résultats d'un SRI dans un ordre basé sur la probabilité de pertinence d'un document vis-à-vis d'une requête. Robertson [Robertson, 77] définit son modèle PRP (*Probability Ranking Principle*), sur ce même principe. Etant donné une requête utilisateur notée Q et un document d , le modèle probabiliste tente d'estimer la probabilité que le document d appartienne à la classe des documents pertinents (non pertinents). Un document est alors sélectionné si la probabilité qu'il soit pertinent à Q , notée $P(R/d)$, est supérieure à la probabilité qu'il soit non pertinent à Q , notée $P(NR/d)$. Le score d'appariement entre le document d et la requête Q , noté $RSV(d,Q)$, est donné par [Robertson, 94b]:

$$RSV(d, Q) = \frac{P(R/d)}{P(NR/d)}$$

Ce qui donne, d'après le théorème de Bayes et après simplification :

$$RSV(d, Q) = \frac{P(R/d)}{P(NR/d)} \approx \frac{P(d/R)}{P(d/NR)}$$

tel que $P(d/R)$ (respectivement $P(d/NR)$) est la probabilité que le document appartienne à l'ensemble R des documents pertinents (respectivement à l'ensemble NR des documents non pertinents).

Différentes méthodes sont utilisées pour estimer ces différentes probabilités. Nous décrivons particulièrement le modèle d'indépendance binaire, connu sous le modèle BIR (Binary Independence Retrieval). On considère dans ce modèle que la variable document $d(t_1 = x_1, t_2 = x_2, \dots, t_n = x_n)$ est représenté par un ensemble d'événements qui dénotent la présence ($x_i = 1$) ou l'absence ($x_i = 0$) d'un terme dans un document. En supposant que ces événements soient indépendants, d'où l'appellation BIR, les probabilités de pertinence (resp. de non pertinence) d'un document, notées $P(d/R)$ (resp. $P(d/NR)$), sont données par :

$$P(d/R) = P(t_1 = x_1, t_2 = x_2, t_3 = x_3, \dots / R) = \prod_i P(t_i = x_i / R)$$

$$P(d/NR) = P(t_1 = x_1, t_2 = x_2, t_3 = x_3, \dots / NR) = \prod_i P(t_i = x_i / NR)$$

t_i est le $i^{\text{ème}}$ terme utilisé pour décrire le document d , et x_i est sa valeur 0 si le terme est absent, 1 si le terme est présent dans le document. La distribution des termes suit une loi de Bernoulli ; $P(d/R)$ peut s'écrire :

$$P(d/R) = \prod_i P(t_i = x_i / R) = \prod_i P(t_i = 1/R)^{x_i} * P(t_i = 0/R)^{1-x_i}$$

On fait le même développement pour $P(d/NR)$. Notons $P(t_i = 1/R)$, par p_i , et $P(t_i = 0/R)$ par q_i , $RSV(d, Q)$ peut s'écrire, après transformation, comme suit :

$$RSV(d, Q) = \frac{p_i^{x_i} * (1-p_i)^{(1-x_i)}}{q_i^{x_i} (1-q_i)^{(1-x_i)}}$$

En se ramenant à la fonction log et après un petit développement, la fonction RSV s'écrit alors :

$$RSV(d, Q) = \sum_{i, x_i=1} \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

Si en outre, on suppose connus l'ensemble R des documents pertinents et l'ensemble NR des documents non pertinents, alors on peut aisément estimer les probabilités p_i et q_i , en utilisant les proportions définies en Tableau 1.1, comme suit :

$$p_i = \frac{r_i}{n} \quad \text{et} \quad q_i = \frac{R_i - r_i}{N - n}$$

#doc. pert. contenant t_i r_i	#doc. pert. ne contenant pas t_i $n - r_i$	#doc. pert. n
#doc.non-pert. contenant t_i $R_i - r_i$	#doc. non-pert. ne contenant pas t_i $N - R_i - n + r_i$	#doc.non-pert. $N - n$
#doc. contenant t_i R_i	#doc. ne contenant pas t_i $N - R_i$	#échantillons N

TABLEAU 1.1 : Distribution de probabilités de pertinence des termes d'un corpus d'apprentissage

Ainsi la RSV se réduit à :

$$RSV(d, Q) = \sum_{i: x_i=1} \log \frac{r_i(N - R_i - n + r_i)}{(n - r_i)(R_i - r_i)}$$

Un des inconvénients de ce modèle est l'impossibilité d'estimer ses paramètres si des collections d'entraînement ne sont pas disponibles. Pour pallier cet inconvénient, Robertson a proposé le modèle 2-poisson basé notamment sur la notion de termes élités [Robertson 94a ; Robertson, 94b]. Le résultat de ces travaux est la formule BM25, largement utilisée dans les travaux actuels de RI.

1.2.2.3.2 Le modèle de langue

Le principe des approches utilisant un modèle de langue est différent des approches classiques en RI. En effet, plutôt que d'évaluer le degré de similarité des documents et requêtes, le modèle de langue considère que la pertinence d'un document pour une requête est en rapport avec la probabilité que la requête puisse être générée par le document [Ponte et al., 98 ; Boughanem et al., 04].

Formellement, soit M_d , le modèle de langue du document d ; la pertinence de d vis-à-vis d'une requête Q revient à estimer $P(Q/M_d)$, c'est-à-dire, la probabilité que la requête Q soit générée par M_d . Etant donné une requête Q , cette pertinence est mesurée par :

$$RSV(d, Q) = P(Q = (t_1, t_2, \dots, t_n) / M_d) = \prod_i P(t_i / d)$$

$P(t_i/d)$ peut être estimé en se basant sur l'estimation maximale de vraisemblance (*maximum likelihood estimation*). Elle est donnée par :

$$P(t_i / d) = \frac{tf(t_i / d)}{\sum_t tf(t / d)}$$

où $tf(t_i/d)$ est la fréquence du terme t_i dans le document d .

Dans ce type d'estimation que lorsqu'un terme de la requête est absent du document, on a systématiquement $RSV(d,Q) = 0$. Afin de pallier cet inconvénient, des techniques de lissage (*smoothing parameter*) peuvent être utilisées. Le lissage consiste à assigner des probabilités non nulles aux termes, qui n'apparaissent pas dans les documents. Différentes techniques de lissage existent dont le lissage de Laplace, le lissage de Good-turing ou le lissage de Backoff. Une description de ces techniques est donnée dans [Boughanem et al.,04].

1.2.3 Reformulation de requêtes

La reformulation de requête consiste, à partir d'une requête initiale formulée par l'utilisateur, à construire une requête qui répond mieux à son besoin informationnel. Les techniques de reformulation de requête se classifient en méthodes locales et méthodes globales. Les méthodes locales ajustent une requête relativement aux documents qui sont retournés comme documents pertinents pour la requête initiale. Elles se basent sur la technique dite de réinjection de pertinence (*relevance feedback*). Les méthodes globales se basent sur l'expansion de requête en s'appuyant sur des ressources linguistiques (thésaurus ou ontologies), ou sur des techniques d'associations de termes telles que les règles d'association. Dans ce qui suit, nous donnons un aperçu de ces approches.

1.2.3.1 Méthodes locales

Les méthodes locales s'appuient sur la technique de réinjection de pertinence [Buckley et al., 94; Harman,92; Robertson et al., 97; Rocchio, 71]. L'idée de la réinjection de pertinence est de faire participer l'utilisateur dans le processus de recherche de sorte à améliorer l'ensemble final de résultats. Le procédé de base est le suivant :

- l'utilisateur formule sa requête,
- le système renvoie un premier ensemble de résultats de recherche,
- l'utilisateur marque quelques documents retournés comme pertinents ou non pertinents,
- le système calcule une meilleure représentation du besoin en l'information sur la base de la rétroaction utilisateur,
- le système visualise un ensemble révisé de résultats de la recherche.

La réinjection de pertinence peut passer par une ou plusieurs itérations de ce type. Le système utilise l'information sur la pertinence utilisateur pour reconstruire la requête. La nouvelle requête Q_m est obtenue à partir de la requête initiale Q_0 en appliquant un

CHAPITRE 1. RECHERCHE D'INFORMATION

algorithme spécifique de réinjection de pertinence, dont par exemple, l'algorithme de Rocchio [Salton et al., 1983; Salton, 1989]. Formellement :

$$Q_m = \alpha Q_0 + \beta \frac{1}{|R|} \sum_{d_p \in R} d_p - \gamma \frac{1}{|NR|} \sum_{d_{np} \in NR} d_{np}$$

Où :

d_p (respectivement d_{np}) est le vecteur associé à un document pertinent (respectivement non pertinent),

R est l'ensemble des documents pertinents,

NR est l'ensemble des documents non pertinents,

α, β, γ étant des constantes telles que $\alpha + \beta + \gamma = 1$.

Les paramètres α, β et γ sont choisis en fonction de l'importance que l'on souhaite donner à la requête initiale (respectivement aux jugements de pertinence).

1.2.3.2 Méthodes globales

Les méthodes globales se basent sur l'expansion de requête. La forme la plus commune d'expansion de requête est l'analyse globale, en utilisant un thesaurus [Qiu, 93] ou une ontologie [Mandala et al., 91; Voorhees, 94; Navigli et al., 03; Moldovan et al., 00; Baziz et al., 03a; Baziz et al., 03b]. Pour chaque terme t , la requête peut être automatiquement étendue avec des mots du thesaurus synonymes ou liés au terme t . Le système peut ainsi appairer la requête à des documents pertinents qui ne contiennent aucun des mots de la requête originale. Outre les relations sémantiques, les termes de la requête peuvent être étendus par des termes qui leur sont autrement liés par des relations de co-occurrence [Schutze et al., 97] ou des relations contextuelles [Mitra et al., 98; Xu et al., 96] qu'un thesaurus ne peut exhiber. Parmi les techniques d'extraction des relations contextuelles entre termes, les règles d'association ont été largement utilisées en RI pour l'expansion de requêtes [Wei et al., 00; Haddad, 02; Song et al., 07]. Nous les présenterons en chapitre 4.

1.2.4 Conclusion

Nous avons présenté dans cette section les concepts fondateurs de la RI. Nous y avons en particulier exposé les techniques d'indexation automatique, les principaux modèles de recherche et les mécanismes de reformulation des requêtes. Les premiers modèles mis en place sont des modèles booléens, simples et intuitifs, basés sur la théorie des ensembles. L'appariement utilisé est strict et ne permet de classer les documents que dans deux catégories : l'ensemble des documents pertinents et l'ensemble des documents non pertinents. Les modèles vectoriels, algébriques, offrent la possibilité d'ordonner les documents retrouvés selon leurs degrés de similarité avec la requête. La robustesse du modèle et ses bonnes performances dans les tests l'ont propulsé au rang des modèles les plus

populaires de RI. Les modèles probabilistes, basés sur la théorie des probabilités, sont plus efficaces que les modèles booléens. Les documents y sont ordonnés selon leurs probabilités de pertinence pour la requête. Ces modèles ont une base théorique saine [Croft et al., 92] et se sont montrés particulièrement performants dans TREC (à l'exemple du système OKAPI [Robertson et al., 92]). L'inconvénient, cependant, avec ces modèles est de trouver des méthodes efficaces pour estimer les probabilités utilisées pour l'évaluation de la pertinence [Crestani et al., 98].

Les modèles de RI présentés dans cette section représentent la grande majorité des systèmes existants. Ces modèles stricts pour les uns (booléens) et flexibles pour les autres (vectoriels et probabilistes), ne prennent cependant pas en compte les préférences utilisateur sur les critères de recherche. En outre, l'agrégation utilisée est exclusivement de type conjonctif (AND) et/ou disjonctif (OR). La RI flexible permet de pallier à ces deux inconvénients en introduisant une flexibilité au niveau des requêtes. Nous présentons cette approche de la RI dans la section suivante.

1.3 La RI flexible

La RI flexible se réfère à l'utilisation de mécanismes de formulation des requêtes, d'indexation et d'évaluation non strictes (floues).

1. Des mécanismes de formulation de requêtes flexibles ont été introduits permettant d'une part d'exprimer les préférences utilisateur sur les critères de recherche, et d'autre part de définir une agrégation plus souple (flexible) entre les critères de recherche.

En indexation, des poids ont été associés aux termes d'indexation permettant ainsi de les différencier selon leur degré de représentativité dans le document. Les poids sont ensuite utilisés pour une évaluation flexible de la pertinence du document pour la requête.

L'évaluation flexible permet de définir pour un document donné, son degré de pertinence pour la requête et de classer les documents par ordre de pertinence. Dans ce contexte, des opérateurs d'agrégation flexibles ont été définies permettant de relaxer les évaluations classiques par le minimum et le maximum. Et des méthodes d'ordonnement flexible des documents, basées sur des méthodes d'analyse multicritères ont été proposées.

Nous présentons dans ce qui suit les extensions flexibles du modèle booléen.

1.3.1 Indexation floue des documents

Dans ce modèle, la requête est exprimée par un ensemble de mots clés non pondérés reliés par des opérateurs booléens (*AND*, *OR* et *NOT*). Le document est représenté comme un ensemble flou de termes [Radecki, 79].

$$R(d) = \{(t, \mu_d(t)) \mid t \in d\}$$

tel que $\mu_d(t)$ définit le degré d'appartenance du terme t au document d . Concrètement, cette valeur équivaut au poids $w_d(t)$ du terme t dans le document d . L'évaluation de la pertinence d'un document d pour une requête booléenne est alors donnée par :

$$\begin{aligned} RSV(d, t_i) &= w_d(t_i) \\ RSV(d, t_i \text{ AND } t_j) &= \min(w_d(t_i), w_d(t_j)) \\ RSV(d, t_i \text{ OR } t_j) &= \max(w_d(t_i), w_d(t_j)) \\ RSV(d, \text{NOT } t_i) &= 1 - w_d(t_i) \end{aligned}$$

En associant des poids aux termes d'indexation, le modèle booléen basé sur les ensembles flous est capable d'ordonner les documents par ordre décroissant de leur pertinence vis-à-vis de la requête.

1.3.2 Formulation de requêtes flexibles

Dans les requêtes, la flexibilité a été introduite à deux niveaux :

Au niveau des critères de recherche : Pour permettre l'expression des préférences utilisateur sur les critères de recherche, les termes de la requête ont été pondérés [Buell et al., 1981; Bordogna et al., 1991; Pasi, 1999]. Des poids numériques ont d'abord été utilisés. Puis, des poids qualitatifs, plus simples et plus intuitifs, ont été formulés à partir de termes linguistiques tels que : *important, très important...* [Bordogna et al., 1993].

Au niveau des opérateurs liant les critères de recherche: des opérateurs flous, intermédiaires entre le *AND* et le *OR* ont été proposés, et des quantificateurs linguistiques tels que : *tous (all), au moins k (at least k),...* ont été introduits dans le langage de requête [Bordogna et al., 95] comme opérateurs d'agrégation flous qualitatifs.

Nous présentons ci-après les mécanismes mis en œuvre dans chaque cas.

1.3.2.1 Prise en compte des préférences utilisateur dans le langage de requête

En associant des poids aux termes de la requête, l'utilisateur peut ainsi fournir une description plus précise de son besoin informationnel [Herrera-Viedma, 99]. Une requête est alors définie comme une expression booléenne dont les composants élémentaires sont des couples $\langle t, w \rangle$ où t est un critère de recherche et w est le poids qui lui est associé [Herrera-Viedma, 00]. Les poids de requête permettent à l'utilisateur de spécifier des restrictions qui doivent être satisfaites par la représentation floue des documents retrouvés par le SRI. Les poids de requête ont d'abord été formalisés comme des valeurs numériques [Bookstein, 80], [Bordogna et al., 91a], [Buell et al., 81b], [Kantor, 81], [Salton et al., 83a], [Waller et al., 79], puis des poids linguistiques plus intuitifs ont été définis [Bordogna et al., 91b].

Les poids numériques de requête indiquent une contrainte qui doit être satisfaite par la représentation des documents de la collection indexée. La nature de la contrainte imposée par le critère de sélection pondéré dépend de la sémantique associée au poids [Bordogna et al., 91a ; Kraft et al., 95]. Dans la littérature, différentes sémantiques des poids de requête ont été proposées. Le poids peut être interprété comme poids d'importance, comme seuil, ou comme description du document idéal.

La sémantique d'importance [Bookstein, 80; Waller et al., 79; Bookstein, 80; Radecki, 79; Crestani et al., 99; Yager, 87] définit les poids de requête comme des mesures de l'importance relative de chaque terme de la requête par rapport aux autres termes (de la requête). En associant des poids d'importance relative aux termes dans une requête, l'utilisateur spécifie qu'il recherche plus les documents contenant le critère le plus important (poids le plus élevé) que ceux contenant des critères moins importants (poids moins élevés).

La sémantique du seuil [Buell et al., 81a; Buell et al., 81b; Kraft et al., 83] définit les poids des requêtes comme des conditions à satisfaire pour chaque terme de la requête considéré dans l'appariement document-requête. Autrement dit, le seuil indique le niveau d'acceptation du degré de signification d'un terme dans un document pour qu'il soit sélectionné.

La sémantique de la perfection [Bordogna et al., 91a; Cater et al., 89; Bordogna et al., 91a; Kraft et al., 95] consiste à considérer la requête pondérée comme une description du document idéal souhaité par l'utilisateur. En associant des poids aux termes de la requête, l'utilisateur souhaite rechercher tous les documents dont le contenu satisfait ou est plus ou moins proche du besoin informationnel idéal représenté par la requête pondérée.

CHAPITRE 1. RECHERCHE D'INFORMATION

La limitation principale des poids numériques de requête est de forcer l'utilisateur à quantifier le concept qualitatif et flou d'importance alors qu'il est plus naturel d'utiliser des quantificateurs linguistique (tels que *important*, *très important*, *assez important* ...). Bordogna et Pasi [Bordogna et al., 91b] ont défini un modèle flou de recherche dans lequel les descripteurs linguistiques sont formalisés dans le cadre de la théorie des ensembles flous [Zadeh, 75] par des variables linguistiques. Un critère élémentaire de recherche est un couple $\langle t, w \rangle$ où t est un terme et w est une valeur qualitative appartenant à l'ensemble des termes de la variable linguistique *Important*. Par exemple, l'ensemble des termes de la variable linguistique pourrait être l'ensemble défini par : $T(\text{Important}) = \{\text{important}, \text{très important}, \text{assez important}, \text{peu important}, \dots\}$. Dans ce cas, le terme linguistique est dit primaire, alors que les termes modulés par les modificateurs linguistiques *très*, *peu*, *assez* (soit *très important*, *assez important*, ...) sont dits termes non primaires. Les significations des termes non primaires dans $T(\text{Important})$ sont obtenus en définissant d'abord la fonction de compatibilité associée au terme primaire *important*, $\mu_{\text{important}}$, et puis en modifiant $\mu_{\text{important}}$ selon la sémantique du modificateur linguistique utilisé [Crestani et al., 99].

1.3.2.2 Agrégation linguistique des critères de recherche

Des opérateurs d'agrégation linguistiques flexibles (tels que *au moins n*, *la plupart de*, *tous*, ...), plus simples et plus intuitifs que les opérateurs booléens classiques, ont été définis [Bordogna et al., 91b]. Les conditions d'une requête booléenne complexe sont plus facilement et intuitivement formulées. Les opérateurs de moyenne pondérée ordonnée (OWA) [Yager, 88] ont été utilisés pour définir les quantificateurs linguistiques.

Exemple

Si l'on souhaite qu'au moins 3 des quatre termes climat, satellite, météorologie et image soient satisfaits, la requête booléenne devra être formulée comme suit :

(climat AND satellite AND météorologie) OR (climat AND satellite AND image) OR (climat AND météorologie AND image) OR (météorologie AND image AND satellite).

En utilisant des quantificateurs linguistiques, la même requête est plus simplement exprimée par :

au moins 3 (climat, satellite, météorologie, image)

Outre le quantificateur *au moins k* qui est défini comme un seuil strict, d'autres quantificateurs avec une signification vague peuvent être définis. Le quantificateur *presque k* est interprété comme seuil flou sur le nombre de critères à satisfaire. L'utilisateur obtient une certaine satisfaction égale quand moins de k critères sont

satisfaits. Le quantificateur *plus de k* spécifie que plus le nombre de critères satisfaits est supérieur à k , plus la valeur globale de satisfaction est élevée. La valeur de pertinence d'un document d pour une requête $q = \text{quantificateur}(q_1, \dots, q_n)$ est calculée comme suit :

$$RSV(d, q) = OWA_{\text{quantifieur}}(e(d, q_1), \dots, e(d, q_n))$$

dans laquelle $OWA_{\text{quantifieur}}$ est l'opérateur OWA lié au quantificateur quantifier. Les q_i sont les critères élémentaires de recherche.

L'opérateur *AND Possible* (*possibly and*) [Bordogna et al., 91b], permet de spécifier des critères de sélection optionnels par rapport à des critères essentiels.

Exemple

Pour exprimer l'intérêt pour les documents traitant des « systèmes experts » (critères essentiels), tandis qu'on déclare un moins grand intérêt pour les documents traitant également de « fuzzy » ou « ANN » (critères facultatifs), la requête suivante peut être formulée :

tous (expert, systems) AND possible au moins 1 (fuzzy, ANN)

L'opérateur *AND Possible* fournit un autre niveau de flexibilité du mécanisme de recherche, permettant de ne pas ignorer les documents qui satisfont seulement les critères essentiels.

1.3.3 Evaluation flexible des requêtes

L'évaluation des requêtes a pour objet de calculer la pertinence des documents pour une requête donnée puis de classer les documents retournés par ordre de pertinence décroissant.

Les méthodes d'agrégation de RI classique utilisent des opérateurs de conjonction (ou disjonction) pondérés. Les requêtes conjonctives (respectivement disjonctions) sont évaluées par des opérateurs conjonctifs (respectivement disjonctifs). Ces opérateurs peuvent être le minimum (respectivement le maximum) pondéré ou la moyenne pondérée. Cependant, ce type d'agrégation est trop restrictif. En particulier, dans le cas de requêtes conjonctives par exemple, la non satisfaction d'un seul critère par un document donné, implique que le document n'est pas sélectionné. Pour relaxer la conjonction, des opérateurs plus flexibles tels que la moyenne pondérée ordonnée (OWA [Yager, 88]) ou le minimum pondéré ordonné (OWmin [Dubois et al., 96]) ont été introduits. L'idée derrière ce type d'agrégation est de donner une faible importance aux poids les plus faibles dans le vecteur d'évaluation, minimisant ainsi l'impact des faibles termes pour éviter de pénaliser un document contenant de faibles termes.

CHAPITRE 1. RECHERCHE D'INFORMATION

Cependant, le problème avec les opérateurs d'agrégation, qu'ils soient stricts ou flexibles, est qu'ils ne permettent pas de distinguer entre des documents ayant une même pertinence globale. Une conséquence est qu'il est impossible de distinguer des documents ayant la même valeur de pertinence globale. Comme exemple, considérons une requête à trois termes, agrégés par la moyenne (les mêmes remarques s'appliquent à d'autres opérateurs d'agrégation) :

$$\begin{aligned} rsv(d_1, q) &= (w_{d_1}(t_1) + w_{d_1}(t_2) + w_{d_1}(t_3)) / 3 = (0.1 + 0.7 + 0.7) / 3 = 0.5 \\ rsv(d_2, q) &= (w_{d_2}(t_1) + w_{d_2}(t_2) + w_{d_2}(t_3)) / 3 = (0.5 + 0.5 + 0.5) / 3 = 0.5 \end{aligned}$$

Le problème est alors de savoir si l'on préfère un document avec une pertinence moyenne pour tous les critères ou ceux ayant une forte pertinence pour la plupart d'entre eux.

Dans les approches proposées dans [Boughanem et al., 05; 07], il ne s'agit plus d'agrégier les poids en une valeur unique, mais plutôt d'ordonner les documents directement sur la base des vecteurs de poids des termes présents dans la requête. Dans [Boughanem et al., 07], deux fonctions de tri avancé sont considérées. Ces méthodes, basées sur l'ordre lexicographique, sont le *discrimin* et le *leximin*. Elles raffinent le minimum classique et permettent ainsi de départager des vecteurs dont le minimum serait égal.

Discrimin : Concrètement, on compare les vecteurs deux à deux par la valeur de l'agrégation de leurs éléments distincts. Ainsi, les valeurs communes pour un même rang dans les deux vecteurs sont éliminées avant de cumuler les valeurs restantes par un opérateur d'agrégation conjonctif (*min* ou OW_{min} pondéré). Ceci permet de n'effectuer le tri que sur les valeurs réellement déterminantes. Appliqué à la recherche d'information, cet opérateur permet donc de ne pas tenir compte des termes de la requête qui ont le même poids dans deux documents afin de déterminer l'ordre de ceux-ci. Par exemple, soient les deux vecteurs à comparer suivants :

$$\begin{aligned} rsv(d_1, q) &= (1; 0.5; 0.1; 0.3) \\ rsv(d_2, q) &= (0.2; 0.7; 0.1; 1) \end{aligned}$$

Ces vecteurs représentent les degrés des termes de la requête q pour les documents d_1 et d_2 . En utilisant le *min* comme opérateur d'agrégation des composantes du vecteur, ce qui revient à considérer la requête comme une conjonction, les deux vecteurs sont à 0.1, et ne peuvent donc pas être distingués. L'utilisation du *discrimin* permet d'éliminer le troisième terme qui est commun aux deux vecteurs et n'est donc pas discriminant dans leur comparaison relative. Ainsi, la valeur pour $rsv(d_1, q)$ devient 0.3 contre 0.2 pour $rsv(d_2, q)$, ce qui permet de les trier. Le *discrimin* permet donc un tri plus précis que le simple *min*.

Leximin : Il revient à appliquer le discrimin sur des vecteurs préalablement réordonnés. La considération des valeurs communes est ainsi indépendante de leur place dans le vecteur. Dans le cas de l'utilisation du minimum comme fonction d'agrégation, ce tri est équivalent à un tri lexicographique des vecteurs triés.

Par ailleurs l'opérateur somme a aussi été raffiné en gardant la trace des informations sur les poids individuels des termes de la requête dans le document, dans une certaine mesure. L'opérateur utilisé est une somme tronquée (dite *LexiSum*) qui élimine progressivement les poids selon leurs valeurs. Étant donné le vecteur ordonné de poids des termes $W = (w_1, \dots, w_n)$, il s'apparente à l'ordre lexicographique des vecteurs de la forme $(w_1 + w_2 + \dots + w_n, w_1 + w_2 + \dots + w_{n-1}, \dots, w_1 + w_2, w_1)$. Ainsi, on considère d'abord la somme de tous les poids comme dans des systèmes classiques, la somme de $n - 1$ poids si les deux sommes sont égales, etc.

Plusieurs variantes ont aussi été introduites, dont *LexiSumR* qui utilise le vecteur $(w_1 + w_2 + \dots + w_n, w_2 + \dots + w_n, \dots, w_{n-1} + w_n, w_n)$ qui supprime les petits poids d'abord. En comparant les résultats obtenus d'une part avec l'agrégation par la somme pondérée à la base de l'approche classique (utilisé dans Mercure [Boughanem, 92]) et d'autre part avec les différents raffinements vectoriels de la somme, et la méthode de classement basée *leximin* qui raffine le minimum (éventuellement appliqué avec un *OWmin*), il a été rapporté que plupart de procédures d'ordre raffinées n'apportent pas d'améliorations significatives en termes de fonctionnement en comparaison avec la somme classique, pour des requêtes tant courtes que longues, à part *LexiSumR* dans le cas de longues requêtes où une amélioration de 4.8 % sur la précision à 10 a été constatée. Ces résultats préliminaires ont montré que l'ordonnancement des documents peut tirer avantage des vecteurs des poids des termes complets, plutôt que de les agréger en une valeur unique.

1.4 Evaluation d'un SRI

L'évaluation constitue une étape importante dans la mise en oeuvre d'un SRI. Cette étape permet de mesurer les caractéristiques du système en termes de qualité de service et de facilité d'utilisation. Cleverdon [Cleverdon, 70] définit six principales mesures de la qualité d'un SRI : l'univers du discours de la collection, le temps de réponse, la présentation des résultats, l'effort requis de l'utilisateur pour retrouver, parmi les documents retournés, ceux qui répondent à son besoin, le taux de rappel du système, la précision du système.

CHAPITRE 1. RECHERCHE D'INFORMATION

Le premier point se réfère au degré auquel le document inclut l'information pertinente. Le temps de réponse, la prestation de sortie et l'effort requis de l'utilisateur sont autant de mesures de la qualité du service rendu à l'utilisateur. Le but d'un SRI est de retrouver l'information recherchée par l'utilisateur et de la lui retourner dans un délai acceptable, en la lui présentant sous une forme aisément exploitable. Ceci implique notamment la manière de présenter les résultats, l'accès aux documents complets, et l'interfaçage en général.

Les mesures de rappel et de précision sont intrinsèques au modèle de recherche du système et couvrent une pertinence dite algorithmique ou système. On retrouve dans [Borlund, 03] une définition plus large de la notion de pertinence, dépendant de nombreux critères examinés liés au contexte de la recherche, tels que : le degré de correspondance (*aboutness*), l'utilité (*usefulness/ utility*), rentabilité (*usability*), l'importance, ... sur les résultats retournés par rapport aux objectifs, aux intérêts, à la situation intrinsèque du moment. Ces différents critères ont amené à la catégorisation de la pertinence utilisateur en 4 classes de pertinence : la pertinence thématique, la pertinence cognitive, la pertinence situationnelle et la pertinence motivationnelle (ou affective [Saracevic, 96]).

1. la pertinence thématique traduit le degré d'adéquation de l'information retrouvée au thème (et non au contenu) de la requête; c'est la pertinence classique telle que définie dans le paradigme de Cleverdon [Cleverdon, 70],
2. la pertinence cognitive représente la relation intellectuelle entre le besoin informationnel intrinsèque de l'utilisateur et l'information portée par les documents telle qu'interprétée par l'utilisateur,
3. la pertinence situationnelle est vue comme l'utilité de l'information retrouvée par rapport au but de la recherche tel que par l'utilisateur,
4. la pertinence motivationnelle ou affective décrit la relation entre les intentions, les buts et les motivations de la recherche tels que fixés par l'utilisateur d'une part et les informations retrouvées d'autre part.

La mesure communément utilisée dans les campagnes d'évaluation classique en RI est sans doute la pertinence thématique. On adopte pour cela, une approche quantitative des SRI qui s'attache à mesurer le degré d'adéquation du document à la requête. Pour mesurer cette adéquation, le SRI procède à la comparaison de la représentation interne de la requête et de la représentation interne des documents. Le degré de similitude du document et de la requête mesure la pertinence du document pour cette requête. Il s'agit là de la pertinence système, ou pertinence algorithmique [Saracevic, 96] (ou pertinence logique [Cooper, 71]). Pour évaluer cette pertinence, nous devons connaître a priori l'ensemble des documents qui sont pertinents pour une

requête donnée. C'est à cette fin que des collections de tests ont été élaborées. Une collection de tests comprend :

1. un ensemble de documents (ou collection de documents) à indexer, sur lesquels le système sera évalué,
2. une liste de requêtes prédéfinies,
3. les jugements de pertinence, manuellement établis, pour chaque requête. Il s'agit, pour chaque requête, de la liste des documents pertinents pour cette requête.

Les collections de tests sont, le plus généralement, mises en place dans le cadre de campagnes d'évaluation des SRI, dont les campagnes TREC¹ (*Text Retrieval Conference*) [Harman 1992] qui constituent la référence en ce qui concerne l'évaluation des SRI, les campagnes CLEF (*Cross-Language Evaluation Forum*) qui se rattachent plus particulièrement aux systèmes multilingues, les campagnes NTCIR sur les langues asiatiques et les campagnes *Amaryllis* (1996-1999) spécialisées dans les systèmes français.

Nous présentons dans ce qui suit la campagne d'évaluation TREC, et expliquons le protocole d'évaluation d'un SRI utilisé dans le cadre de cette campagne. Puis, nous introduisons quelques mesures d'évaluation de la pertinence algorithmique d'un système.

1.4.1 La campagne d'évaluation TREC

TREC est un projet international, initié au début des années 90 par le NIST (*National Institute of Standards and Technology*) aux Etats-Unis, dans le but de proposer des moyens homogènes d'évaluation de systèmes documentaires sur des bases de documents conséquentes. Il est aujourd'hui co-sponsorisé par le NIST et l'ARPA (ex-DARPA/ITO, pour *Defense Advanced Research Projects Agency - Information Technology Office*, qui mène plusieurs actions dans le domaine des technologies de l'informatique et de la communication, et qui dépend du ministère de la défense).

Le projet TREC consiste en une série d'évaluations annuelles des technologies pour la RI, dont l'objectif est :

1. d'une part, d'offrir aux chercheurs le moyen de mesurer sur des procédures d'évaluation uniformes, l'efficacité de leurs systèmes,
2. d'autre part, de leur permettre de comparer les résultats de leurs systèmes.

Les pistes explorées par TREC sont entre autres, la recherche (ou tâche ad-hoc), le filtrage, la question-réponse, la vidéo, le web ...

La tâche ad-hoc est la tâche principale dans TREC. Elle vise à évaluer les performances d'un SRI sur des ensembles statiques de documents, seules les requêtes

¹ <http://trec.nist.gov>

changent. Pour cette tâche, les participants du TREC disposent d'une collection d'environ 02 gigaoctets de texte, sur un CD-ROM fourni par le NIST. Avec ces documents, le NIST procure également aux participants un ensemble de 50 requêtes en langage naturel.

Les participants testent leurs systèmes sur les documents fournis, recherchant les réponses aux requêtes données, puis classent les documents de la collection par ordre de pertinence, pour chaque requête. Les 1000 premiers documents retrouvés pour chaque requête sont soumis au NIST, chargé de l'évaluation. Le protocole d'évaluation utilisé se base sur deux principales métriques qui sont les taux de rappel et de précision. Nous les définissons en section suivante.

1.4.1.1 Les mesures d'évaluation d'un SRI

Etant donnée une requête Q , les documents de la collection peuvent être globalement classifiés en fonction de leur rapport à la requête (pertinents/non pertinents) comme le montre la figure 1.3.

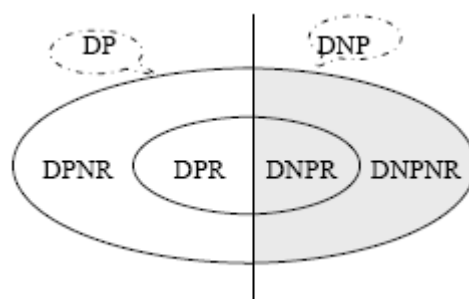


FIGURE 1.3 : Distribution des documents dans une collection face à une requête

Où DP est l'ensemble des documents pertinents pour la requête Q , DPR l'ensemble des documents pertinents retrouvés, $DPNR$ l'ensemble des documents pertinents non retrouvés, DNP l'ensemble des documents non pertinents pour Q , $DNPR$ l'ensemble des documents non pertinents retrouvés et $DNPNR$ l'ensemble des documents non pertinents non retrouvés

On définit les mesures de rappel et de précision d'un SRI par les proportions suivantes :

$$Précision : P = \frac{|DPR|}{|DPR \cup DNPR|} \quad Rappel : R = \frac{|DPR|}{|DP|}$$

La précision est la proportion de documents retrouvés qui sont pertinents. Une précision égale à 1 signifie que le système n'a retrouvé que des documents pertinents. Le rappel est la proportion de documents pertinents qui sont retrouvés. Un rappel égal à 1 signifie que tous les documents pertinents ont été retrouvés. L'idéal serait d'avoir une précision et un

rappel égaux à 1, signifiant que tous les documents pertinents sont retrouvés et qu'aucun document non pertinent n'a été retrouvé. En pratique, cet idéal n'est jamais atteint puisque ces deux quantités évoluent en sens inverse. Intuitivement, si on augmente le rappel en retrouvant plus de documents pertinents, on diminue la précision en retrouvant aussi plus de documents non pertinents. Inversement, une plus grande précision risque de rejeter des documents pertinents diminuant ainsi le rappel.

1.4.1.2 Protocole d'évaluation TREC

Pour chaque requête, les 1000 premiers documents restitués par le système sont examinés et des précisions sont calculées à différents points (à 5, 10, 15, 30, 100 et 1000 premiers documents restitués). La précision exacte découle de ces précisions. C'est la précision à x , x étant le nombre total de documents pertinents dans la collection, pour la requête examinée. Puis, une précision moyenne *MAP* est calculée pour chaque requête. Il s'agit de la moyenne des précisions de chaque document pertinent pour cette requête. La précision d'un document est la précision à x , tel que x est le rang de ce document dans l'ensemble des documents pertinents retrouvés. Finalement, les précisions moyennes pour l'ensemble des requêtes sont calculées permettant d'obtenir une mesure de la performance globale du système.

1.4.2 Autres mesures d'évaluation d'un SRI

D'autres mesures d'évaluation d'un SRI existent. Ainsi, des mesures complémentaires au rappel et à la précision, respectivement le bruit et le silence ont été définies comme suit :

$$\text{Bruit} : B = \frac{|DNPR|}{|DPR \cup DNPR|} \quad \text{Silence} : S = \frac{|DPNR|}{|DP|}$$

L'indice de *Fallout* (ou *Hallucination* d'après C. Berrut²) peut être utilisé à la place du rappel [Ishioka, 03]. Il définit le pourcentage de documents non pertinents qui ont été retrouvés ([Kraft et al., 78] cité dans [Ishioka, 03]). Il exprime l'erreur du système. Formellement :

$$\text{Fallout} : a = \frac{|DNPR|}{|DNP|}$$

L'élimination est la mesure complémentaire du *Fallout*. Elle définit le pourcentage de documents non pertinents non retrouvés. Elle est définie par :

² <http://isdn.enssib.fr/archives/transversal/JDN/indexation/BerrutJDN20 mars supports.pdf>

CHAPITRE 1. RECHERCHE D'INFORMATION

$$\text{Elimination : } E = \frac{|DNPNR|}{|DNP|}$$

Par ailleurs, Van Rijsbergen [Van Rijsbergen, 79] a introduit la *F-mesure* comme combinaison du rappel et de la précision. La F-mesure est définie à travers la formule suivante :

$$F_{\beta} = \frac{(\beta^2 + 1) * P * R}{\beta^2 * P + R}$$

Où β traduit l'importance relative du rappel et de la précision.

Par exemple, $\beta=2$ représente une précision deux fois plus importante que le rappel. Dans le cas particulier où $\beta=1$, la F-mesure définit la moyenne harmonique du rappel et de la précision:

$$F_1 = \frac{2 * P * R}{P + R}$$

En pratique, plus grande est la valeur de la *F-mesure*, meilleure est la recherche [Ishioka, 03].

1.5 Conclusion

Nous avons présenté dans ce chapitre les concepts fondamentaux de la RI. Le but d'un SRI est de rechercher l'information pertinente pour une requête utilisateur. Son efficacité est mesurée par des paramètres qui reflètent sa capacité à accomplir un tel but. Ce but est par nature non déterministe. Les SRI classiques incapables de prendre en compte une telle imprécision. Pour pallier à ce manque, de nouveaux modèles flexibles ont été proposés. Néanmoins, l'imprécision ne caractérise pas uniquement le processus de recherche ou le langage de requête. En effet, l'imprécision est aussi portée par les mots même de la langue du fait de leur ambiguïté naturelle. L'indexation par les mots clés est de ce fait imprécise. De nouvelles techniques d'indexation sont nécessaires pour pallier l'ambiguïté de la langue et pouvoir traiter avec la sémantique des documents et requêtes. C'est l'objet de l'indexation sémantique que nous présentons dans le chapitre suivant.

Chapitre 2

Indexation sémantique en RI

2.1 Introduction

L'indexation sémantique s'intéresse principalement à la représentation des documents et requêtes par les sens des mots qu'ils contiennent plutôt que par les mots eux mêmes. L'objectif sous-jacent est d'améliorer la représentation des entités indexées et de pallier aux problèmes de l'indexation classique basée mots.

L'objectif du présent chapitre est de présenter les principales approches d'indexation sémantique. En section 2.2, nous présentons la problématique de l'indexation classique basée mots-clés. Le reste du chapitre est dédié à la présentation des approches d'indexation sémantique. Ainsi, l'approche d'indexation conceptuelle est décrite en section 2.3. La section 2.4 est dédiée à la présentation des approches d'indexation sémantique basées sur la désambiguïsation. Tout d'abord, un aperçu des méthodes de désambiguïsation est présenté en paragraphe 2.4.1, puis les approches d'indexation sémantique en paragraphe 2.4.2. Ces approches sont basées soit sur la désambiguïsation basée corpus, ou sur la désambiguïsation basée sur les ressources externes. Les premières sont présentées en paragraphe 2.4.2.1, les secondes en paragraphe 2.4.2.2.

2.2 Problématique

En indexation classique, les entités textuelles (documents et requêtes) sont représentées par des mots clés issus de leurs contenus. L'utilisation des mots pour représenter le contenu des documents et requêtes pose deux problèmes, l'ambiguïté des mots et leur disparité.

L'ambiguïté des mots, dite ambiguïté lexicale, se rapporte à des mots lexicalement identiques et portant des sens différents. Elle est généralement divisée en deux types [Krovetz, 97; Krovetz et al, 92] : l'ambiguïté syntaxique et l'ambiguïté sémantique.

L'ambiguïté syntaxique se rapporte à des différences dans la catégorie syntaxique. Par exemple, « *play* » peut apparaître en tant que nom ou verbe. L'ambiguïté sémantique se rapporte à des différences dans la signification, et est décomposée en homonymie et polysémie selon que les sens sont liés ou non [Krovetz, 97].

Le problème d'ambiguïté implique que des documents non pertinents, contenant les mêmes mots que la requête sont retrouvés. Par exemple, dans une recherche à l'aide du mot clé *AIDS* (SIDA en français), Krovetz et al. [Krovetz et al., 92] rapportent que 34 références contenant le mot *AIDS* ont été retrouvées mais toutes ne traitaient pas de la maladie.

La disparité des mots (word mismatch) se réfère à des mots lexicalement différents mais portant un même sens. Ceci implique que des documents, pourtant pertinents, ne partagent pas de mots avec la requête, ne sont pas retrouvés. Dans le même contexte de recherche sur *AIDS*, des documents portant sur le *VIH*, pourtant pertinents ne seront pas retrouvés.

Les travaux du domaine ont d'abord adressé ces problèmes séparément en apportant des solutions spécifiques à chacun d'eux, puis une solution globale s'est dégagée.

(1) Solutions spécifiques

- une réponse au premier problème, en l'occurrence l'ambiguïté des mots, est d'utiliser les expressions ou termes composés, pour réduire l'ambiguïté. Cependant, il n'est pas toujours possible de fournir une expression dans laquelle le mot apparaît seulement avec le sens désiré, et la formulation des expressions exige un effort cognitif de la part de l'utilisateur.
- une réponse au second problème, en l'occurrence la disparité des mots, consiste à étendre la requête à l'aide de mots synonymes d'un thésaurus [Salton et al., 83]. Cette extension n'est pas aléatoire. Pour enrichir un mot dans la requête par ses synonymes, on doit non seulement connaître le sens du mot dans la requête, mais aussi le sens du mot qui est utilisé pour l'étendre [Krovetz et al., 92].

(2) Solution globale

La solution globale permettant de répondre à ces deux problèmes consiste en l'indexation sémantique. L'indexation sémantique tente d'apporter des solutions au niveau de la représentation des documents et des requêtes. L'objectif est d'indexer par les sens des mots plutôt que par les mots. Dans un contexte où l'ambiguïté est présente, l'indexation sémantique est sensée améliorer les performances du SRI.

CHAPITRE 2. INDEXATION SEMANTIQUE EN RI

L'indexation sémantique s'intéresse à deux principaux points : d'abord retrouver le sens correct de chaque mot dans le document (respectivement de la requête), ensuite représenter ce document (respectivement cette requête).

En réponse au premier point portant sur l'identification du sens des mots, l'indexation sémantique s'appuie sur des techniques dites de désambiguïsation des mots ou WSD (*Word Sense Disambiguation*). Deux principales approches de désambiguïsation existent : les approches endogènes et les approches exogènes [Audibert, 03]. Les premières se basent sur des corpus d'entraînement pour calculer le sens correct d'un mot [Weiss, 73; Schütze, 92;98], alors que les secondes s'appuient sur l'exploitation du contexte local et des définitions issues de ressources linguistiques externes telles que les dictionnaires informatisés ou MRD (*Machine Readable Dictionary*), [Lesk, 86 ; Veronis et al., 90 ; Ide et al., 90 ; Wilks et al., 90 ; Guthrie et al., 91], les thésaurus [Yarowsky, 92], les ontologies [Sussna, 93 ; Resnik, 93a ; 93b ; 95] ou une combinaison d'entre elles [Agirre et al., 01]. A ces approches de désambiguïsation sont donc associées deux approches d'indexation sémantique que l'on nommera respectivement les approches basées sur le corpus et approches basées sur les ressources externes en rapport avec la technique de désambiguïsation utilisée.

En réponse au second point portant sur la représentation sémantique des documents et requêtes, l'indexation sémantique s'intéresse à leur représentation en se basant sur les sens des mots qu'ils contiennent. Dans ce contexte, deux principales approches de représentation existent: la représentation basée sur les sens et la représentation combinée mots-clés/sens. Dans la première, les mots des documents et requêtes sont désambiguïsés et ce sont les sens correspondants calculés, qui sont finalement utilisés comme termes d'indexation. Dans la seconde approche, les sens sont utilisés conjointement avec les mots clés qu'ils représentent. Un terme d'indexation est alors représenté par le couple (mot-clé, sens associé).

Notons enfin, qu'il existe une approche d'indexation sémantique dite indexation conceptuelle³, qui s'affranchit des problèmes de désambiguïsation et qui tente plutôt d'indexer les documents et requêtes par des entités conceptuelles qui sont extraites des textes correspondants.

³ Notons que souvent l'indexation conceptuelle est définie comme une indexation sémantique puisque les concepts véhiculent la sémantique. Bien que nous adhérons à ce point de vue, nous avons choisi de suivre la classification donnée dans [Mihalcea et al., 00] selon laquelle l'indexation conceptuelle réfère principalement à l'approche de Woods, tandis que toute indexation basée sur les sens des mots relève de l'indexation sémantique.

2.3 L'indexation conceptuelle

L'indexation conceptuelle se réfère à la construction de taxonomies conceptuelles à partir des textes. Cette approche est due à Woods [Woods, 97]. Le système conceptuel d'indexation et de recherche proposé extrait automatiquement des mots et des expressions de textes et les organise en un réseau sémantique (taxonomie conceptuelle) qui intègre des relations syntaxiques, sémantiques et morphologiques. La construction d'une taxonomie de concepts à partir des textes est le plus souvent réalisée en parsant automatiquement chaque expression en une ou plusieurs structures conceptuelles qui représentent comment les éléments de l'expression sont réunis pour construire son sens(s). Ceci permet à un système de déterminer automatiquement quand le sens d'une expression est plus général que celui d'une autre étant donnée sa connaissance des rapports de généralité entre les différents éléments qui composent l'expression.

Exemple

Etant donnée l'information que *voiture* est un **genre-de** automobile et que *lavage* est un **genre-de** Nettoyage (figure 2.1), un système peut automatiquement déterminer que *lavage de voiture* est un **genre-de** nettoyage d'automobile.

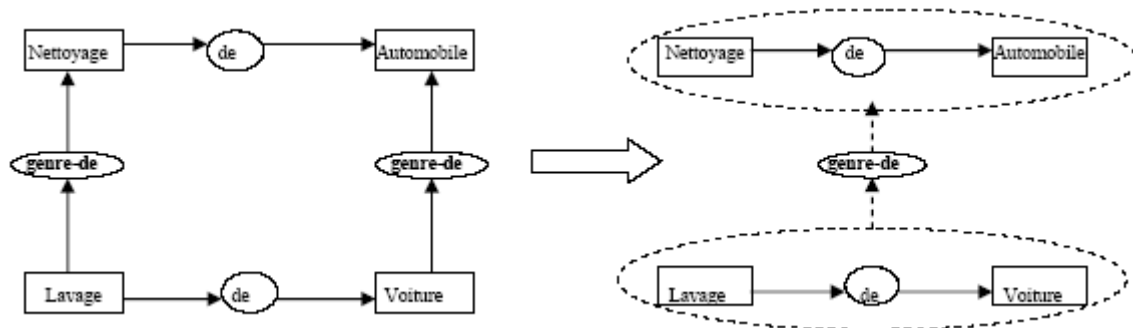


FIGURE 2.1 : Un exemple de taxonomie conceptuelle.

Ainsi, un système peut automatiquement déterminer des rapports de généralité entre les concepts structurés s'il a une connaissance de base sur les rapports de subsumption (rapports de généralité/spécificité) entre des concepts de base.

L'approche d'indexation conceptuelle de Woods a été testée sur de petites collections de texte (dont les pages du manuel UNIX composé de 1819 fichiers et occupant une taille d'environ 10MB). La comparaison des résultats de ce système avec ceux obtenus en utilisant des techniques classiques de recherche, a montré une amélioration du rappel de l'ordre de 0.3% par rapport aux SRI classiques.

2.4 L'indexation sémantique basée sur la désambiguïsation

Même si les mots de la langue sont par nature ambigus, il n'en demeure pas moins qu'il a été nécessaire d'étudier l'impact de l'ambiguïté sur la RI et l'opportunité d'introduire les techniques de désambiguïsation en indexation des documents. Les travaux de Krovetz et Croft [Krovetz et al., 92; Krovetz, 93] ont été les premiers à étudier l'impact de l'ambiguïté sur les performances du processus de recherche d'information et à initier l'idée que la désambiguïsation pouvait aider à améliorer la RI. Les études ont été menées principalement à deux niveaux: d'abord déterminer le degré d'ambiguïté lexicale dans les collections de test en RI, ensuite déterminer l'utilité des sens des mots dans la séparation des documents pertinents et non pertinents.

- Pour déterminer le degré d'ambiguïté dans les collections de test, des statistiques sur les sens des mots qu'elles contiennent ont été établies. Le nombre moyen de sens dans les documents et requêtes est déterminé par un processus de consultation d'un dictionnaire. Les statistiques rapportées par Krovetz et Croft sur les collections CACM⁴ et TIME⁵ indiquent que ces deux collections ont un fort taux d'ambiguïté (le nombre moyen de sens pour la collection CACM est de 4.7 et de 3.7 pour la collection TIME) et par conséquent présentent un fort potentiel pour bénéficier de la désambiguïsation. Par ailleurs, les résultats rapportés indiquent que les mots dans les requêtes sont bien plus ambigus que ceux dans les documents.
- Pour étudier l'impact de l'indexation par les sens des mots sur l'efficacité de la recherche, des statistiques sur le nombre de disparités de sens dans les documents pertinents ont été établies. Les résultats rapportés ont montré que la disparité des sens est faible dans les documents pertinents. Les sens permettent bien de séparer les documents pertinents des documents non pertinents.

Les travaux qui s'en sont suivis [Sanderson, 94 ; Krovetz, 97; Gonzalo et al., 98], ont montré que l'impact de l'ambiguïté des sens sur l'efficacité de la recherche n'était pas dramatique, mais qu'une désambiguïsation précise (précision de plus de 90% selon [Sanderson, 94], de 60% selon [Gonzalo et al., 99]) des mots améliorerait probablement l'efficacité de la recherche lorsque peu de mots de la requête apparaissent dans le document. De là, l'indexation par les sens des mots (ou indexation sémantique) a été

⁴ [http : //www.dcs.gla.ac.uk/idom/ir_resources/test_collections/cacm/](http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/cacm/)

⁵ [https : //isserver11.princeton.edu/](https://isserver11.princeton.edu/)

pressentie comme un moyen qui permettrait d'améliorer les performances de la recherche. Pour retrouver les sens corrects des mots dans un document, l'indexation sémantique a recours aux techniques de désambiguïsation des sens des mots.

Avant de décrire les approches d'indexation par les sens des mots, nous présentons d'abord les principes fondateurs des approches de désambiguïsation puis les travaux les plus significatifs dans le domaine.

2.4.1 Les approches de désambiguïsation des sens des mots (WSD)

Un certain nombre d'approches de désambiguïsation des sens des mots existent, qui sont principalement divisées en approches exogènes et approches basées sur les corpus. Les approches exogènes utilisent des bases de connaissances externes (dictionnaires, thesaurus, lexiques, ontologies,...) pour désambiguïser des mots ambigus. Les approches basées sur le corpus sont généralement de type statistique et utilisent de gros corpus de textes pour construire la connaissance nécessaire à la désambiguïsation.

Nous présentons dans ce qui suit les travaux les plus représentatifs de chacune d'elles. Notons toutefois, qu'il existe aussi des approches mixtes qui combinent plusieurs techniques.

2.4.1.1 Les approches exogènes

Partant de l'hypothèse que « quand plusieurs mots co-occurrent dans un contexte, le sens le plus probable pour chacun de ces mots est celui qui maximise ses relations avec les sens des mots co-occurents » [Audibert, 03; Ide et al., 98], les approches exogènes se basent sur l'exploitation du contexte et des définitions issues de ressources linguistiques externes telles que les dictionnaires informatisés ou MRD (*Machine Readable Dictionary*), [Lesk, 86; Veronis et al., 90; Ide et al., 90; Wilks et al., 90; Guthrie et al., 91], les thésaurus [Yarowsky, 92], les ontologies [Sussna, 93; Resnik, 93a; Resnik, 93b; Resnik, 95] ou une combinaison d'entre elles [Agirre et al., 01].

2.4.1.1.1 Les approches basées sur les dictionnaires informatisés

En s'appuyant sur un dictionnaire informatisé, Lesk [Lesk, 86] a construit l'un des premiers systèmes de désambiguïsation basés sur un MRD⁶. Le principe de désambiguïsation de Lesk peut être défini comme suit :

⁶ Lesk a testé son approche sur trois MRD : le Webster's 7th Collegiate, le Collins English Dictionary et le Oxford Advanced Learner's Dictionary of Current English.

CHAPITRE 2. INDEXATION SEMANTIQUE EN RI

1. Pour chaque occurrence de mot ambigu, retrouver tous les sens du mot dans un dictionnaire.
2. Pour chaque sens S du mot à désambiguïser :
 - Consulter sa définition,
 - $Score(S)$ = le nombre de mots en commun entre la définition du mot à désambiguïser et les définitions des mots cooccurrents dans son contexte}.
 - Retenir le sens S qui maximise $Score(S)$.

Cette méthode permet de désambiguïser correctement dans 50% à 70% des cas. Cependant, elle présente l'inconvénient (cité dans [Sanderson, 97]) d'être très sensible aux mots qui se trouvent dans chaque définition. En effet, le choix des sens basés sur un nombre restreint de mots communs peut être source d'erreurs. Ainsi par exemple, bien que sémantiquement liés, les mots *sandwich* et *Breakfast* n'ont pas de mots en commun dans leurs définitions respectives suivantes :

- *Two (or more) slices of bread a filling between them,*
- *The first meal of the day (usually in the morning).*

L'algorithme de Lesk les considérant de ce fait comme totalement sémantiquement indépendants.

Par ailleurs, la présence ou l'absence d'un mot donné peut radicalement changer le résultat. En effet, dans les cas où aucun mot ne co-occure entre le contexte et les définitions ambiguës de l'occurrence à désambiguïser, l'approche de Lesk ne permet pas de désambiguïser.

La méthode de Lesk sert tout de même de base pour la plupart des travaux postérieurs en désambiguïsation basée sur les dictionnaires informatisés.

La méthode de Lesk a été étendue par Véronis et Ide [Veronis et al., 90; Ide et al., 90] en générant un réseau de neurones à partir des définitions du dictionnaire anglais Collins (*Collins English Dictionary* ou CED). Dans ce réseau, chaque entrée lexicale est représentée par un regroupement complexe de noeuds se composant de :

- un noeud central (ou *nœud mot*) qui représente l'entrée lexicale (mot) elle-même.
- des nœuds sens qui représentent les différents sens de ce mot dans le CED
- des nœuds mots

Le noeud central est relié à un certain nombre de nœuds sens. Chacun de ces nœuds sens est relié aux nœuds mots représentant les mots qui apparaissent dans sa définition. Ces mots sont à leur tour reliés aux nœuds sens selon leurs définitions dans le CED ... etc. La structure ainsi établie est un réseau hautement complexe, dans lequel les mots sémantiquement liés sont connectés via un ou plusieurs chemins dans

le réseau. Les expériences menées sur 23 mots polysémiques ont montré des résultats prometteurs [Ide et al., 90; Véronis et al., 90].

Pour résoudre le problème des définitions courtes posées par Lesk, Wilks et al. [Wilks et al., 90] ont utilisé l'approche de désambiguïsation de Lesk avec le dictionnaire LDOCE (*Longman Dictionary of Contemporary English*) [Longman, 88], dont les définitions ont été manuellement étendues. La technique d'expansion utilisée consiste à enrichir toutes les définitions du dictionnaire LDOCE avec les mots qui co-occurrent généralement avec le texte de ces définitions. Cette information de co-occurrence a été dérivée de toutes les définitions du dictionnaire.

Wilks a examiné la précision de son désambiguïseur sur le mot 'bank' qui apparaît dans environ 200 phrases du dictionnaire LDOCE. Pour évaluer son désambiguïseur, Wilks a d'abord désambiguïsé manuellement ces phrases. Puis les résultats de la désambiguïsation automatique sont comparés à ceux de la désambiguïsation manuelle. Wilks a rapporté que son système retrouvait le sens correct du mot 'bank' dans plus de 50% des cas.

Dans une approche identique à celle de Wilks, Guthrie et al. [Guthrie et al., 91] ont exploité un ensemble de catégories de sujets⁷ existantes, assignées à plusieurs définitions de sens dans le LDOCE, pendant le processus d'expansion de définitions. A la différence de Wilks, dans l'approche de Guthrie et al., une définition assignée à une certaine catégorie est étendue avec les seuls mots co-occurents dans les autres définitions assignées à la même catégorie. Aucun test n'a cependant été rapporté pour ce désambiguïseur.

2.4.1.1.2 Les approches basées sur un thésaurus

Yarowsky [Yarowsky, 92], se basant sur l'encyclopédie Grolier multimédia [Grolier] et sur les 1042 catégories sémantiques⁸ dans lesquelles tous les mots du thésaurus Roget [Kirkpatrick, 88] sont placés, propose une approche de désambiguïsation en deux étapes: la première consiste à assigner une catégorie (parmi les 1024 citées ci-dessus) au mot à désambiguïser, la seconde consiste à assigner le sens correct à l'occurrence de ce mot dans la catégorie ainsi déterminée.

Pour décider à quelle catégorie sémantique une occurrence de mot ambigu doit être assignée, un ensemble de mots indices (ou mots déterminants selon la terminologie de [Ricart, 06]), est construit pour chaque catégorie sémantique, en utilisant l'encyclopédie Grolier. Pour dériver l'ensemble des mots déterminants d'une catégorie Ω donnée :

⁷ Le LDOCE comporte 124 catégories de sujets majeures dont certaines contiennent des sous-catégories (par exemple, economics, engineering, ...).

⁸ Il s'agit de larges catégories couvrant des domaines comme, les machines/outils ou les insectes/animaux

CHAPITRE 2. INDEXATION SEMANTIQUE EN RI

- Rechercher dans l'encyclopédie, toutes les occurrences m_j^i de chacun des mots m_i de Ω ,
- Pour chaque occurrence m_j^i ainsi trouvée, recueillir son contexte ζ_j^i . Le contexte est défini par l'ensemble des 100 mots entourant l'occurrence, soit 50 mots à droite et 50 mots à gauche. Soit donc C l'ensemble de tous les contextes liés à la catégorie Ω , défini par :

$$C = \bigcup_i \left(\bigcup_j \zeta_j^i \right)$$

- Pour chaque mot m_i dans C , calculer $Score(m_i)$ sur la base du résultat de la comparaison de sa fréquence d'occurrence dans C à sa fréquence d'occurrence dans toute l'encyclopédie.
- Les mots de scores les plus élevés sont utilisés comme mots déterminants pour leur catégorie sémantique.

Pour désambiguïser un mot dans une catégorie donnée, on examine son contexte. Si un mot déterminant apparaît dans ce contexte, le mot ambigu appartient probablement à la catégorie du mot déterminant.

Dans ses tests, Yarowsky a appliqué son désambiguïseur sur 12 mots ambigus. Plusieurs centaines d'occurrences de ces mots ont été manuellement désambiguïsés servant de base d'évaluation de la désambiguïsation automatique. La précision moyenne du désambiguïseur est de 92%. Cependant, comme l'a rapporté Sanderson [Sanderson, 97], aucune comparaison n'est possible entre ce travail et d'autres travaux antérieurs de désambiguïsation en particulier car aucun n'avait utilisé le thésaurus Roget.

2.4.1.1.3 Les approches basées sur un lexique

L'approche de désambiguïsation de Sussna [Sussna, 93] s'appuie sur WordNet et sur le contexte local du mot à désambiguïser. Le principe de l'approche est simple : pour désambiguïser un mot ambigu apparaissant dans un certain contexte, on recherche tous les synsets de WordNet contenant ce mot. Chaque synset est affecté d'un score égal à la somme des distances sémantiques entre les mots du contexte et ce synset. Le synset qui maximise le score est retenu comme sens de l'occurrence ambiguë du mot.

Pour calculer la distance sémantique entre deux mots quelconques dans le réseau WordNet, Sussna a assigné un poids à toutes les relations entre synsets de WordNet. Le plus fort poids assigné à une relation reflète la proximité sémantique exprimée par cette relation. Par exemple, il a assigné le poids le plus élevé aux relations de synonymie dans un synset, tandis que des relations d'antonymie ont eu le poids le

plus faible. La distance sémantique entre deux synsets est alors calculée comme la somme des poids des relations sur le chemin le plus court entre ces deux synsets.

Sussna a testé sa technique de désambiguïsation sur dix documents extraits de la collection TIME, à partir desquels 319 occurrences de mots ambigus ont été examinées. Ces occurrences ont d'abord été manuellement désambiguïsées servant de référence à l'évaluation du désambiguïseur. Le désambiguïseur a résolu ces occurrences avec une précision de 56%.

Resnik dans [Resnik, 93a ; 93b; 95] explore une mesure de similarité sémantique construite à partir de la taxinomie *is-a* des noms de WordNet. Le principe qui sous-tend cette mesure est que plus deux mots sont sémantiquement proches, plus le concept qui les subsume est spécifique. La méthode de Resnik approche les performances de désambiguïsation humaine.

2.4.1.2 Approches basées sur le corpus (approches endogènes)

Le principe de base de l'acquisition de connaissances à partir des corpus pour la désambiguïsation lexicale est simple. En étudiant un grand nombre de contextes de chacune des occurrences d'un mot polysémique, il est possible d'identifier statistiquement des indices récurrents se démarquant (des indices saillants selon la terminologie de [Audibert, 03]) pour chacune d'elles. Cette phase d'identification automatique des connaissances est appelée apprentissage. À l'issue de cette phase, l'algorithme de désambiguïsation est capable d'assigner le terme adéquat aux mots apparaissant dans une nouvelle phrase en se basant sur les connaissances acquises durant la phase d'apprentissage. Les approches basées sur le corpus se divisent en approches supervisées [Weiss, 73; Kelly et al., 75] et approches non supervisées [Small et al., 82; Schütze, 92;98]. Les premières s'appuient sur des corpus manuellement étiquetés tandis que les secondes s'affranchissent de cette limitation.

2.4.1.2.1 Approches basées sur les corpus étiquetés

En examinant 20 occurrences d'un mot ambigu, Weiss [Weiss, 73] a manuellement construit deux types de règles permettant la désambiguïsation. Il s'agit de règles générales de contexte et de règles de modèle.

1. Une règle générale de contexte établit qu'une occurrence de mot ambigu a un certain sens si un mot particulier apparaît *près de* cette occurrence du mot ambigu. Par exemple, si le mot '*print*' apparaît près du mot '*type*' alors son sens est probablement lié à l'impression.
2. Une règle de modèle établit qu'une occurrence d'un mot ambigu a un certain sens si un mot particulier apparaît *à un endroit spécifique* relatif à cette occurrence. Par

CHAPITRE 2. INDEXATION SEMANTIQUE EN RI

exemple, si le mot 'of' apparaît juste après le mot 'type', alors le sens de cette occurrence est probablement « *variety of* ».

L'ensemble des règles ainsi construites a été testé sur 30 occurrences de plus du mot ambigu. La précision du désambiguïseur résultant est de l'ordre de 90%.

Utilisant une approche similaire à celle de Weiss, Kelly et Stone [Kelly et Stone, 75] ont manuellement créé un ensemble de règles pour 6.000 mots. Ces règles sont composées de :

1. règles contextuelles semblables à celles créées par Weiss,
2. règles de vérification des aspects grammaticaux d'une occurrence de mot. La catégorie grammaticale d'un mot est en effet, parfois, un indicateur fort de son sens comme dans l'exemple 'the train' et 'to train'.

Les règles de grammaire et de contexte ont été regroupées en deux ensembles de sorte que seules certaines règles sont appliquées dans certaines situations. Des hypothèses conditionnelles contrôlent l'application des ensembles de règle. À la différence du système de Weiss, ce désambiguïseur a été conçu pour traiter une phrase entière en même temps. Le système n'a cependant pas eu de succès, et Kelly et Stone ont bien rapporté que cette technique ne peut pas réussir à échelle réelle.

Une autre approche de désambiguïsation a été tentée par Small et Rieger [Small et al., 82] employant des programmes appelés 'experts de mots'. L'idée était de construire un expert pour chaque mot ambigu. En désambiguïsant des mots dans une phrase (qui constitue alors le contexte du mot à désambiguïser), l'expert de chacun de ses mots sera appelé. Un expert examine le contexte de son mot, prend des décisions au sujet des sens possibles de ce mot et communique ces décisions aux autres experts. Si, en traitant sa connaissance, un expert ne peut rien faire de plus, il devient 'dormant' et attend que d'autres experts de mots lui communiquent leurs décisions. Cette connaissance additionnelle fournira d'autres indices à l'expert dormant pour lui permettre de 'se réveiller' et finir de désambiguïser son mot. Aucun test ni résultat n'ont été rapportés pour ce désambiguïseur [Sanderson, 97].

Le principal inconvénient des approches décrites jusqu'ici est qu'elles sont basées sur des règles manuellement créées pour déterminer les sens des mots. Quand ces approches étaient testées sur de plus grands vocabulaires, les résultats obtenus étaient peu concluants [Kelly et al., 75; Small et al., 82].

2.4.1.2.2 Approches basées sur les corpus non étiquetés

Dans ce type d'approche, la notion de sens est généralement directement induite du corpus.

Schütze [Schütze, 92; 98] propose une méthode basée sur le modèle vectoriel utilisé en RI [Salton et al., 75]. Dans cette approche, chaque mot m du corpus d'apprentissage est représenté par un vecteur dans un espace de grande dimension. Un vecteur pour un mot m est dérivé à partir des mots qui co-occurrent dans le contexte de m . Le contexte d'une occurrence est défini par une fenêtre de cinquante mots autour de l'occurrence en question. Une entrée d'un mot m_i dans le vecteur associé à m correspond au nombre de cooccurrences de m_i dans le contexte de m . On définit alors le vecteur de contexte d'une occurrence de mot donné comme la moyenne des vecteurs associés aux mots de son contexte.

Le processus de désambiguïsation consiste d'abord à identifier, pour chaque mot m du corpus, tous les vecteurs de contexte associés à toutes les occurrences de m . Ces vecteurs de contexte sont ensuite regroupés en clusters en fonction de leur degré de similitude. Chaque cluster définissant un sens possible du mot m . Pour chaque cluster obtenu, on calcule son barycentre (centre de gravité) et on lui associe le sens représenté par le cluster. Pour désambiguïser une nouvelle occurrence du mot m , on calcule la distance de son vecteur de contexte à chacun des barycentres des clusters associés à m . Le sens correspondant au barycentre le plus proche est retenu.

Les méthodes basés sur les corpus non étiquetés possèdent l'avantage de la disponibilité des corpus, mais véhiculent un inconvénient majeur : les sens ne correspondent à aucun ensemble de sens bien défini. Les distinctions de sens peuvent parfois s'avérer déroutantes et sont, de plus, souvent difficilement utilisables par d'autres applications que celle pour laquelle ils ont été définis [Wilks et al., 97].

2.4.2 Les approches d'indexation sémantique

L'indexation sémantique s'intéresse à la représentation des documents et requêtes par les sens des mots qu'ils contiennent. Les sens des mots sont retrouvés par application d'une méthode de désambiguïsation. Dans ce qui suit, nous distinguons les approches d'indexation basées sur la désambiguïsation endogène (basée sur le corpus), des approches d'indexation basées sur la désambiguïsation exogène (basée sur les ressources externes). Les premières sont présentées en section 2.3.2.1, tandis que les secondes sont définies en section 2.3.2.2.

2.4.2.1 Indexation sémantique basée sur la désambiguïsation endogène

Dans ce cas, des corpus d'apprentissage sont d'abord utilisés pour construire la connaissance nécessaire à la désambiguïsation. Les mots d'index sont ensuite identifiés dans la collection à indexer, puis désambiguïsés. Finalement, les textes de la collection sont indexés en utilisant les sens ainsi retrouvés.

Pour construire la connaissance nécessaire à la désambiguïsation, un grand nombre de contextes de chacune des occurrences d'un mot ambigu est examiné à partir d'un corpus d'entraînement, à l'issue de quoi une connaissance sur les règles d'agencement et de fonctionnement des mots [Weiss, 73], ou sur les usages des mots [Schütze et al., 95] est extraite. Cette connaissance est ensuite utilisée pour assigner le sens adéquat aux mots apparaissant dans un nouveau contexte.

Les systèmes de désambiguïsation de Weiss [Weiss, 73] et de Schütze et Pedersen [Schütze et al., 95] sont basés sur ce principe. Le désambiguïseur de Weiss se base sur des règles de désambiguïsation manuellement construites par apprentissage à partir des contextes associées à différents mots d'un corpus. En désambiguïsant tous les mots ambigus dans une collection de documents et en indexant la collection par les sens adéquats, les résultats rapportés par le système SMART [Salton, 83], était une amélioration de seulement 1% sur la précision de la recherche.

Le désambiguïseur de Schütze et Pedersen [Schütze et al., 95] se base sur le degré de recouvrement du contexte de l'occurrence du mot à désambiguïser, et des usages possibles de ce mot dans le corpus examiné. Les usages sont obtenus en regroupant des contextes similaires. Un usage de mot (*word usage*) définit alors un sens individuel pour ce mot. En indexant la collection TREC-1 catégorie B, avec seulement 25 requêtes, Schütze et Pedersen ont rapporté que l'indexation basée sur la combinaison des mots-clés et de leur trois meilleurs usages du mot apportait un gain en précision de 14%.

2.4.2.2 Indexation sémantique basée sur la désambiguïsation exogène

Le principe de base des approches d'indexation basées sur la désambiguïsation exogène, diffère des approches d'indexation précédentes (section 2.3.2.1), principalement dans la méthode utilisée pour la désambiguïsation. Ici, la connaissance nécessaire à la désambiguïsation n'est plus apprise à partir d'un corpus, mais est extraite de la ressource linguistique externe utilisée. Formellement, cette connaissance se traduit par des scores associés aux différents sens d'un mot, sur la base de :

- la distance sémantique de ce sens aux différents sens associés aux autres termes dans le document (contexte global) [Mihalcea et al., 00 ; Khan et al., 04 ; Baziz et al., 04 ; 05a,b,c],
- degré de recouvrement entre d'une part, le contexte local de ce mot et d'autre part le voisinage [Voorhees, 93] de ce sens ou la définition de ce sens (ensemble de synonymes) [Katz et al., 98] dans la ressource linguistique utilisée.

La plupart des approches d'indexation sémantique basées sur la désambiguïsation exogène, s'appuient en général sur des ontologies pour déterminer les différents sens du mot mais aussi pour désambiguïser les sens des mots. Le principe de base de l'indexation consiste alors à extraire dans un premier temps, l'ensemble des termes descripteurs du document. Il s'agit ici d'une indexation classique. Ces termes sont ensuite désambiguïsés. Pour ce faire, les sens de chaque terme d'indexation sont d'abord retrouvés à partir de la ressource externe. Puis, des scores sont associés aux différents sens ainsi retrouvés. Le sens qui maximise le score est alors retenu comme sens adéquat du terme d'indexation correspondant. Une fois les termes d'indexation désambiguïsés, la représentation des textes indexés se fait soit à partir des seuls sens (ou concepts) identifiés lors de l'étape de désambiguïsation, soit à partir d'une combinaison des mots-clés et sens corrects associés. Les approches d'indexation de [Baziz et al., 04; 05 ; Khan et al., 04 ; Mihalcea et al., 00 ; Voorhees, 93; Katz et al., 98] sont basées sur ce principe.

Dans l'approche d'indexation de Voorhees, les textes à indexer sont analysés phrase par phrase. La phrase définit alors le contexte local de chacun des mots qui y apparaissent. A chaque mot non vide rencontré dans la phrase, on recherche dans WordNet le (ou les) concepts (synsets ou ensembles de sens dans WordNet) qui lui correspondent. Un mot ambigu correspond à plusieurs synsets dans Wordnet. Pour déterminer le synset (sens) adéquat pour un mot ambigu dans une phrase, chaque synset de ce mot est classé en se basant sur le nombre de mots co-occurents entre un voisinage (Voorhees l'a appelé *hood*) de ce synset et le contexte local du mot ambigu correspondant. Le synset le mieux classé est alors considéré comme sens adéquat de l'occurrence analysée du mot ambigu.

En considérant l'ensemble des synsets et les relations d'hyponymie et hyperonymie dans WordNet comme les sommets et les arcs orientés d'un graphe, Voorhees définit le voisinage d'un synset *s* comme :

"... le plus large sous graphe connexe qui contient s, contient seulement les descendants d'un ancêtre de s et ne contient aucun synset ayant un descendant qui inclut une autre instance d'un membre (c. à d. un mot) de s."

CHAPITRE 2. INDEXATION SEMANTIQUE EN RI

A titre d'exemple, à partir du fragment de la structure de WordNet donnée par la figure 2.2, le voisinage du premier sens de "house" inclurait les termes *housing*, *lodging*, *apartment*, *flat*, *cabin*, *gatehouse*, *bungalow*, *cottage*. Les termes *structure* et *construction* (situé en haut de la hiérarchie), ne seraient pas inclus puisque un des descendants de leur synset contient un autre sens du terme *house*.

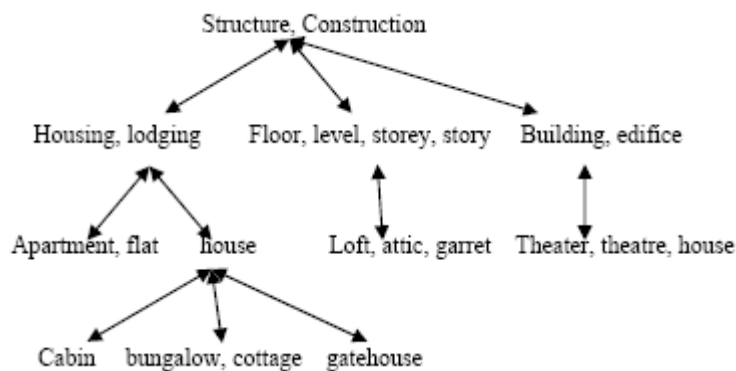


FIGURE 2.2 : Exemple de voisinage du mot house.

En utilisant une version modifiée du système SMART [Salton, 83], Voorhees a expérimenté cette approche sur une collection de test désambiguïsée (les requêtes de la collection de test sont aussi désambiguïsées manuellement) par rapport aux performances du même processus sur la même collection dans son état d'origine (ambigu). Les tests ont été effectués sur les collections CACM, CISI, CRANFIELD 1400, MEDLINE, et TIME. Les résultats de ses expérimentations ont montré pour chacune de ces collections, une nette diminution des performances du SRI dans le cas d'utilisation des collections désambiguïsées. Une raison possible est que le taux de désambiguïsation n'est pas assez élevé. La technique de désambiguïsation pourrait aussi être en cause.

Dans une approche similaire, Katz et al [Katz et al., 98] analysent les textes à indexer mot par mot. Chaque mot non vide rencontré est projeté sur WordNet dans l'objectif d'identifier le (ou les) synset(s) correspondant(s). Si un mot apparie plusieurs synsets, il est ambigu. Pour désambiguïser, Katz et al proposent aussi une approche basée sur le contexte local. Le contexte local d'un mot est défini comme étant la liste ordonnée des mots démarrant du mot utile le plus proche du voisinage gauche ou droit jusqu'au mot cible.

Exemple

Dans le texte "... the jury had been **charged** to investigate reports of irregularities in the primary...", le contexte local droit de *charged* est "*X to investigate*". Son contexte local gauche est "*the jury has been X*".

CHAPITRE 2. INDEXATION SEMANTIQUE EN RI

L'hypothèse de Katz et al., est que des mots utilisés dans le même contexte local (appelés *sélecteurs*), ont souvent des sens proches. Les sélecteurs des mots d'entrée sont extraits des contextes locaux gauche et droit, puis l'ensemble S de tous les sélecteurs obtenus est comparé avec les synsets de WordNet. Le synset qui a le plus de mots en commun avec S est sélectionné comme sens adéquat du mot cible.

Katz et al. ont testé leur désambiguïseur sur le corpus Semcor. La précision rapportée est de 60%. En incorporant ce désambiguïseur au système SMART, Katz et al. ont rapporté que leur algorithme n'améliorait pas les performances du système. Ceci pourrait être du aux erreurs de désambiguïsation.

Dans l'approche d'indexation de Khan, les termes d'indexation sont d'abord extraits par une approche classique d'indexation (tokenisation, élimination des mots vides, puis lemmatisation). Les termes d'indexation identifiés sont ensuite projetés sur une ontologie de domaine (du sport), en utilisant une liste de synonymes. L'objectif est de sélectionner les concepts de l'ontologie correspondants à ces termes. Un terme d'indexation qui s'apparie à plus d'un concept de l'ontologie est ambigu. Pour désambiguïser, on détermine le degré de corrélation des concepts sélectionnés, sur la base de leur proximité sémantique. La proximité sémantique de deux concepts est calculée par un score basé sur leur distance minimale mutuelle dans l'ontologie. Les concepts ambigus qui ont les plus hauts scores sont alors retenus. La requête est alors étendue avec les concepts ainsi sélectionnés.

En indexant ainsi des paragraphes annotant des passages audio dans le domaine du sport, Khan et al. ont rapporté que leur modèle, comparé à un modèle vectoriel classique basé mots-clés, assurait un haut degré de précision et de rappel (de l'ordre de 90% chacun).

Dans une approche similaire, Baziz et al. [Baziz et al., 04 ; 05a,b,c] proposent une technique d'indexation sémantique des documents à base de concepts et de relations entre concepts. Les termes d'indexation sont d'abord extraits du document par une approche classique d'indexation. Les termes d'indexation sont ensuite projetés sur l'ontologie linguistique WordNet afin d'identifier les concepts (ou sens) correspondants dans l'ontologie. Lorsqu'un terme d'indexation apparie plus d'un concept dans WordNet, il est ambigu. Il faut le désambiguïser. L'approche de désambiguïsation proposée est basée sur le principe que, parmi les différents sens possibles (dits concepts candidats) d'un terme donné, le plus adéquat est celui qui a le plus de liens avec les autres concepts du même document. Formellement, l'approche consiste à affecter un score à chaque concept candidat d'un terme d'indexation donné. Le score d'un concept candidat est obtenu en sommant les valeurs de similarité qu'il a avec les autres concepts candidats (correspondant aux différents sens des autres termes du document). Le concept candidat ayant le plus haut score est alors retenu comme sens adéquat du terme d'indexation associé.

Finalement, le document est représenté comme un réseau de concepts et de liens entre concepts. Les liens (arcs) entre les différents concepts sont pondérés par les valeurs de similarité sémantique (ou proximité sémantique [Leacock et al., 94; Lin, 98; Resnik, 99; Lesk, 86]) entre concepts liés.

L'approche d'indexation sémantique ainsi proposée (dite approche DocCore) a été évaluée d'une part dans le cadre de la collection de test MuchMore⁹ [Buitelaar et al., 04], d'autre part dans le cadre de la campagne CLEF 2004. Dans les deux cas, un SRI basé sur le modèle connexionniste est utilisé [Boughanem et al, 92]. Les résultats rapportés montrent que l'utilisation des sens (concepts de WordNet) seuls pour représenter les documents ne permet pas d'améliorer les résultats comparativement à la méthode classique basée sur les mots clés. Cependant, la combinaison de l'indexation classique et de l'indexation sémantique apporte une nette amélioration de la précision.

Dans l'approche DocTree [Baziz et al., 05] complétant l'approche DocCore, une fois les réseaux sémantiques de la requête et des documents construits (à partir de la méthode DocCore), la requête et le document sont ensuite projetés sur le sous réseau conceptuel de l'ontologie WordNet, constitué uniquement de la relation de subsumption (*IS-A*). La requête et le document sont donc représentés par des sous hiérarchies formées par les concepts qu'ils contiennent et qui appartiennent ceux de l'ontologie. Les deux représentations sont comparées en utilisant des opérateurs flous et une valeur de pertinence est alors calculée. Cette valeur exprime jusqu'à quel point le document contient les thèmes (*features*) exprimés dans la requête.

2.5 Conclusion

Nous avons consacré ce chapitre à l'état de l'art sur l'indexation sémantique en RI. L'indexation conceptuelle se base sur la représentation des textes par des taxonomies conceptuelles, tandis que l'indexation sémantique indexe par les sens des mots et se fonde sur des techniques de désambiguïsation des sens. Nous avons passé en revue les différentes approches d'indexation sémantique qui ont apporté la preuve que la désambiguïsation par les sens des mots était bénéfique à la RI [Schütze et al., 95]. En particulier, dans [Mihalcea et al., 00 ; Baziz et al., 04 ; 05] il a été montré que l'indexation par des synsets de WordNet, en plus de l'indexation basée mots-clés classique, peut réellement améliorer l'efficacité de la RI.

Les sens des mots, les synonymes ne sont pas les seuls éléments susceptibles de porter la sémantique d'un texte. En effet, déjà comme le montrait Deerwester à travers sa technique LSI, la sémantique du texte peut être latente, cachée dans le texte, et pas seulement

⁹ <http://muchmore.dfki.de/>

CHAPITRE 2. INDEXATION SEMANTIQUE EN RI

explicite, donnée par un dictionnaire ou autre ressource linguistique. En rejoignant presque l'idée de cette dimension sémantique latente, les techniques de fouille de textes visent à explorer la connaissance enfouie dans le texte. Cette connaissance, exprimée sous forme de liens de co-occurrence conditionnelle entre les différents composants du texte d'un document ou entre différents documents d'un corpus, est un bon indicateur d'une sémantique que ni un dictionnaire ni thesaurus ne peut exhiber. C'est cette idée qui a été à la base de nos recherches sur l'indexation sémantique qui sera détaillée dans le chapitre 4. La partie suivante de la thèse est dédiée à la définition de notre contribution à la définition de modèles de RI flexibles basés sur les CP-Nets.

PARTIE 2

Modèles de RI flexibles basés sur les CP-Nets

Chapitre 3

Modèle de RI flexible basé sur les CP-Nets

3.1 Introduction

Dans les modèles de RI flexibles, les termes de la requête ont été pondérés et des quantificateurs linguistiques tels que : *tous*, *au moins k*, ... ont été introduits dans le langage de requête comme opérateurs d'agrégation flous qualitatifs, offrant par là même un langage de requête plus souple que la simple utilisation mots-clés connectés par les opérateurs *AND* et *OR*. La pondération des termes de la requête a permis la formulation de préférences utilisateur sur les critères de recherche. Des poids numériques ont d'abord été utilisés. Puis, des poids qualitatifs, plus simples et plus intuitifs, ont été formulés à partir de termes linguistiques tels que: important, très important.... Notre travail s'inscrit dans cette optique, et nous proposons dans ce contexte une approche de RI flexible basée sur l'utilisation des graphes CP-Nets (*Conditional Preferences Networks*). Plus particulièrement, nous proposons outre un langage de requête graphique basé sur les CP-Nets, et permettant la formulation des préférences utilisateur de manière simple et intuitive, une méthode d'agrégation flexible basée sur les CP-Nets.

Ce chapitre est organisé comme suit : En section 3.2, nous décrivons le formalisme CP-Net sur lequel se base notre approche. La section se décline en trois sous sections. Dans la sous- section 3.2.1, nous définissons les notations et les concepts utilisés dans la définition des CP-Nets. La sous-section 3.2.2 est dédiée à la présentation du modèle CP-Net. En section 3.3 nous présentons le modèle UCP-Net, extension du modèle CP-Net avec des valeurs d'utilité. La section 3.3 présente notre modèle de pondération des requêtes basé CP-Nets. En intégrant l'idée de la définition d'un modèle entièrement basé sur les CP-Nets, nous définissons en section 3.4 une approche d'interprétation CP-Nets des documents indexés, puis nous présentons en section 3.5 un modèle d'évaluation flexible des requêtes CP-Nets.

3.2 Problématique et motivations

L'introduction des poids dans les termes de la requête [Buell et al., 81; Bordogna et al., 91; Pasi, 99] a permis d'exprimer les préférences utilisateur sur les critères de recherche. L'utilisateur peut ainsi fournir une description plus précise de son besoin informationnel. Cependant, les approches classiques de pondération des requêtes posent les problèmes suivants :

1. La pondération force l'utilisateur à quantifier le concept qualitatif et vague d'importance. Cette tâche n'est pas évidente en particulier lorsque le nombre de critères de recherche est élevé et que la requête est complexe, d'une part car il n'existe pas de bonne méthode pour pondérer correctement les termes de la requête, d'autre part, lorsque le nombre de valeurs sur lesquelles portent les préférences est élevé, il est quasiment impossible d'énumérer un poids valide pour tous les termes de la requête. Ceci est d'autant plus vrai pour les requêtes portant sur les préférences conditionnelles.
2. Les préférences conditionnelles ne sont pas spécifiquement prises en charge dans les langages de requêtes classiques. De tels énoncés préférentiels peuvent certes être traduits dans le langage booléen, cependant leur pondération n'est pas une tâche évidente et peut conduire à des incohérences.

Nous illustrons le problème posé pour les préférences conditionnelles à travers l'exemple qui suit.

Etant donné le besoin utilisateur exprimé à travers l'énoncé suivant :

"I am looking for housing in Paris or Lyon of studios or university room type. Knowing that I prefer to be in Paris rather than to be in Lyon, if I should go to Paris, I will prefer being into residence hall (RH), whereas if I should go to Lyon, a studio is more preferable to me than a room in residence hall. Moreover the Center town of Paris is more preferable to me than its suburbs; whereas if I must go to Lyon, I will rather prefer to reside in suburbs than in the center".

Une telle requête fait ressortir des préférences conditionnelles. En traduisant les préférences qui y sont exprimées en valeurs numériques, une requête correspondante possible serait :

$$(Paris\ 0.9 \wedge (RH\ 0.6 \vee Studio\ 0.3) \wedge (Center\ 0.5 \vee Suburbs\ 0.4)) \vee (Lyon\ 0.8 \wedge (RH\ 0.5 \vee Studio\ 0.8) \wedge (Center\ 0.7 \vee Suburbs\ 0.8)).$$

CHAPITRE 3. MODELE DE RI FLEXIBLE BASE SUR LES CP-NETS

Dans cette représentation, les poids des termes *R.H* et *Studio*, *Center* et *Suburbs*, sont différents lorsqu'ils sont associés avec *Paris* ou *Lyon* respectivement. Ceci traduit exactement les préférences conditionnelles exprimées dans l'énoncé du besoin utilisateur. La forme normale disjonctive de cette requête est donnée par :

$$\begin{aligned} & (Paris\ 0.9 \wedge RH\ 0.6 \wedge Center\ 0.5) \vee (Paris\ 0.9 \wedge Studio\ 0.3 \wedge Center \\ & 0.5) \vee (Paris\ 0.9 \wedge RH\ 0.6 \wedge Suburbs\ 0.4) \vee (Paris\ 0.9 \wedge Studio\ 0.3 \wedge \\ & Suburbs\ 0.4) \vee (Lyon\ 0.8 \wedge RH\ 0.5 \wedge Center\ 0.7) \vee (Lyon\ 0.8 \wedge Studio \\ & 0.8 \wedge Center\ 0.7) \vee (Lyon\ 0.8 \wedge RH\ 0.5 \wedge Suburbs\ 0.8) \vee (Lyon\ 0.8 \wedge \\ & Studio\ 0.8 \wedge Suburbs\ 0.8) \end{aligned} \quad (3.1)$$

Même si cette représentation supporte naturellement les préférences conditionnelles, elle reste problématique si quelques précautions ne sont pas prises au préalable. En effet, en supposant que chaque sous requête conjonctive de la requête globale possède un poids d'importance total, calculé par agrégation des poids individuels de ses propres termes (en utilisant l'opérateur *min* ou l'opérateur OWA [Dubois et al., 86; Yager, 87] ou simplement en moyennant les poids par exemple), on obtient un poids d'importance de $(Paris \wedge Studio \wedge Center)$ égal à 0.56 tandis que le poids d'importance de $(Lyon \wedge Studio \wedge Center)$ est de 0.76 impliquant que la dernière alternative est préférée à la première. Ce résultat est contradictoire avec les préférences formulées par l'utilisateur. La pondération que nous avons donnée, de façon tout à fait aléatoire et intuitive, est incorrecte.

Cet exemple fait ressortir l'impact d'une pondération aléatoire ou intuitive d'une requête qualitative, sur la précision et l'exactitude de la sémantique qu'elle tente d'exprimer. Ceci illustre la tâche difficile de la pondération des requêtes qualitatives.

De ce fait, des travaux se sont orientés vers l'utilisation de préférences qualitatives plus simples et plus intuitives, formulées à partir de termes linguistiques tels : important, très important... [Bordogna et al., 93 ; Bordogna et al., 95]. Cependant, le problème de la définition des poids numériques des termes est reporté sur la définition de la sémantique du concept flou important et des modulateurs linguistiques très, peu, moyennement.

Pour pallier ces inconvénients, nous proposons, au travers de cette contribution [Boubekeur et al., 06a ;b], une approche mixte d'expression des préférences utilisateur combinant l'expressivité et la simplicité du formalisme qualitatif à la puissance calculatoire du formalisme quantitatif. Nous nous intéressons particulièrement aux préférences conditionnelles. Une représentation qualitative, naturelle, simple et compacte de telles formes de préférences est supportée par les CP-Nets [Boutilier et al., 99]. Nous proposons un modèle de RI basé sur les CP-Nets. Plus particulièrement, nous proposons :

1. une approche de formulation des requêtes utilisateur flexibles (portant sur les préférences conditionnelles) à base de CP-Nets,
2. une méthode de pondération automatique de la requête. Cette pondération correspond à la quantification du CP-Net par des valeurs de préférence (ou

valeurs d'utilité). L'extension des CP-Nets par association de valeurs d'utilités, conduit à un UCP-Net [Boutilier et al., 01], correspondant à une requête pondérée correcte. La requête CP-Net ainsi pondérée doit être évaluée.

3. une approche d'évaluation flexible des requêtes basée sur la sémantique des CP-Nets.

3.3 Les CP-Nets

Les CP-Nets (*Conditional Preference Networks*) ont été introduits en 1999 par Boutilier et al. [Boutilier et al., 99], comme outil de représentation compacte des relations de préférences qualitatives. Ce modèle graphique exploite l'indépendance préférentielle conditionnelle dans la structuration des préférences utilisateur sous l'hypothèse *ceteris-paribus*¹⁰. Nous définissons ces notions en paragraphe suivant, avant d'introduire le modèle CP-Net.

3.3.1 Notations et définitions préliminaires

Soit $V = \{X_1, X_2, \dots, X_n\}$, un ensemble de variables (caractéristiques ou attributs) sur lesquelles les préférences utilisateur sont définies, étant donné un problème décisionnel fixé, et soit V' un sous ensemble de V .

Selon la terminologie de [Boutilier et al., 99; Boutilier et al., 01a; Brafman et al., 02b], on note :

- $Dom(X_i) = \{x_{i1}, x_{i2}, \dots, x_{im}\}$, le domaine de valeurs de la variable X_i ,
- $Asst(V')$, l'ensemble de toutes les instanciations possibles de V' . Une instanciation de V' résulte de l'affectation d'une valeur à chaque variable dans V' .
- $Asst(V)$, l'espace de toutes les instanciations possibles sur les variables de V .
- Chaque élément dans $Asst(V)$ définit une alternative.
- $O = Asst(V) = Dom(X_1) \times Dom(X_2) \times \dots \times Dom(X_n)$ l'ensemble de toutes les alternatives possibles,
- Une assignation de valeur d'un sous-ensemble X de V est notée x .
- La concaténation de deux assignations partielles disjointes respectivement de X et de Y est notée xy .

¹⁰ Toutes choses égales par ailleurs (*all else being equal*)

CHAPITRE 3. MODELE DE RI FLEXIBLE BASE SUR LES CP-NETS

- Si $X \cup Y = V$, alors xy est un résultat complet (ou alternative). Il est dit « complétion » de l'assignation partielle x . $comp(x)$ est l'ensemble des complétions de x .

Définition d'un préordre complet : Une relation R sur un ensemble Ω donné définit un préordre complet (total) sur Ω , si et seulement si :

1. \succeq est réflexive ($\forall x \in \Omega, x R x$)
2. \succeq est transitive ($\forall x, y, z \in \Omega, (x R y) \wedge (y R z) \Rightarrow (x R z)$)
3. \succeq est complète ($\forall x, y \in \Omega \mid x \neq y, \neg(x R y) \Rightarrow (y R x)$)

Définition de la relation de préférence : Une relation de préférence, notée \succeq , définie sur l'ensemble des alternatives O , est un préordre complet sur O .

D'où,

$$\forall o, o' \in O, o \succeq o' \text{ ou } o' \succeq o$$

tel que $o \succeq o'$ signifie que l'alternative o est au moins aussi préférée que l'alternative o' .

Définition de l'indépendance préférentielle : Un ensemble de caractéristiques X est préférentiellement indépendant de son complément $Y=V-X$ si et seulement si :

$$\forall x_1, x_2 \in Asst(X), \forall y_1, y_2 \in Asst(Y), x_1 y_1 \succeq x_2 y_1 \Leftrightarrow x_1 y_2 \succeq x_2 y_2$$

Si X est préférentiellement indépendant de son complément $Y= V- X$, on notera $PI(X, Y)$. Ceci équivaut à dire que l'ordre de préférence sur les éléments x_1 et x_2 de X reste inchangé quelques soient les valeurs des éléments y_i de Y . On dit que x_1 est préférable à x_2 *ceteris paribus* (ie. toutes choses égales par ailleurs).

Remarque : Une variable X est préférentiellement dépendante d'une variable Y lorsque les préférences sur les valeurs de X dépendent des valeurs de Y . Y est dit parent de X , et on note $Y= Pa(X)$. Le couple $(X, Pa(X))$ définit une famille de V .

Définition de l'indépendance préférentielle conditionnelle : Soient X, Y et Z des ensembles non vides qui partitionnent V . X est conditionnellement préférentiellement indépendant de Y étant donné $z \in Z$ (et on note $CPI(X, z, Y)$) si et seulement si :

$$\forall x_1, x_2 \in Asst(X), y_1, y_2 \in Asst(Y), x_1 y_1 z \succeq x_2 y_1 z \Leftrightarrow x_1 y_2 z \succeq x_2 y_2 z$$

En d'autres termes, l'indépendance préférentielle de X et de Y ne se produit que lorsque Z prend la valeur z . Si de plus, $\forall z \in Z, CPI(X, z, Y)$, alors X est

conditionnellement préférentiellement indépendant de Y étant donné Z (on note $CPI(X, Z, Y)$).

Définition d'une fonction d'utilité : Une fonction d'utilité u pour l'ordre de préférence \succeq défini sur O , est une fonction à valeur réelle sur O ,

$$\begin{aligned} u : O &\rightarrow \mathfrak{R} \\ o_i &\rightarrow u(o_i) \end{aligned}$$

telle que : $\forall o_1, o_2 \in O, o_1 \succeq o_2 \Leftrightarrow u(o_1) \geq u(o_2)$

Définition de l'indépendance généralisée additive (Generalized Additive Independance) [Bacchus et al., 95] : Soient X_1, X_2, \dots, X_k des ensembles de variables non nécessairement disjoints, et $V = \bigcup_{i=1..k} X_i$. X_1, X_2, \dots, X_k sont indépendants généralisés additifs (ou *GAI*) pour la fonction d'utilité u si et seulement si u peut être décomposée en une somme de facteurs d'utilité f_i définis sur chacun des ensembles X_i ($i= 1.. k$). Formellement,

$$\exists f_{i(i=1..k)} / u(V) = \sum_{i=1..k} f_i(X_i)$$

3.3.2 Le modèle CP-Net

Etant donné un problème décisionnel défini sur un ensemble de N variables (ou attributs) X_1, X_2, \dots, X_n sur lesquelles l'utilisateur exprime ses préférences, chaque variable X_i est définie sur son propre domaine de valeurs $Dom(X_i) = \{x_{i1}, x_{i2}, \dots, x_{im}\}$. Durant la formulation des préférences, pour chaque variable X , l'utilisateur doit spécifier une liste de variables parentes de X (noté $Pa(X)$), qui vont affecter ses préférences sur les valeurs de X . Ainsi, pour chaque valeur de $Pa(X)$, l'utilisateur doit spécifier un ordre de préférence total sur les valeurs de X *ceteris paribus*. Cette information est utilisée pour créer un graphe dans lequel chaque noeud X possède $Pa(X)$ comme prédécesseur immédiat.

Chaque noeud X dans le graphe est annoté par une table de préférences conditionnelles (*Conditional Preference Table*) $CPT(X)$, décrivant les préférences utilisateur sur les valeurs x_i de X , étant donnée chaque assignation de ses parents.

La structure des énoncés d'indépendance préférentielle conditionnelle ainsi obtenus constitue le graphe CP-Net.

Exemple

Etant donné un ensemble de 3 variables (ou attributs) $V= \{A, B, C\}$ binaires définies par $Dom(A)= \{a_1, a_2\}$, $Dom(B)= \{b_1, b_2\}$ et $Dom(C)= \{c_1, c_2\}$. Mes préférences sur les valeurs de ces trois attributs sont définies comme suit :

1. je préfère inconditionnellement a_1 à a_2 (i.e $a_1 \succ a_2$),
2. mes préférences sur les valeurs de B dépendent des valeurs prises par A . Ainsi, si A prend la valeur a_1 , je préfère b_1 à b_2 , sinon je préfère b_2 à b_1 .

Ces préférences conditionnelles se notent comme suit

$$\left\{ \begin{array}{l} a_1 : b_1 \succ b_2 \\ a_2 : b_2 \succ b_1, \end{array} \right.$$

et serviront à annoter le noeud B dans le graphe CP-Net. La variable A qui détermine mes préférences sur les valeurs de B , est le parent de B dans le graphe CP-Net.

De manière similaire, mes préférences sur les valeurs de C dépendent de celles de A , et s'écrivent comme suit :

$$\left\{ \begin{array}{l} a_1 : c_1 \succ c_2 \\ a_2 : c_2 \succ c_1 \end{array} \right.$$

Le CP-Net qui encode mes préférences sur les variables A, B et C est ainsi défini à travers le graphe illustré en figure 3.1.

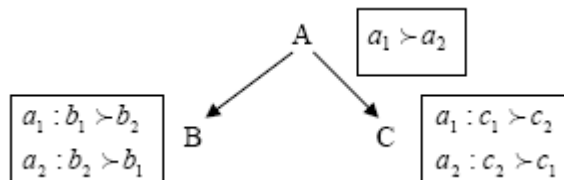


FIGURE 3.1: Un exemple de CP-Net

La relation de préférence capturée par un CP-Net induit un ordre de préférence partiel sur l'ensemble des assignations aux variables du CP-Net. Cet ordre partiel peut être représenté par un graphe de préférences orienté et acyclique. Les nœuds du graphe sont des alternatives (i.e. des assignations à toutes les variables) du CP-Net. Une relation du nœud o_i vers le nœud o_j signifie que o_j est l'alternative immédiatement plus préférable à o_i .

Par convention, le graphe des préférences induit est ainsi ordonné par ordre de préférence qualitatif décroissant :

1. le sommet du graphe de préférences représente l'alternative la moins préférable,
2. la feuille du graphe représente l'alternative la plus préférable.

A titre d'exemple, le graphe des préférences induit par le CP-Net de la figure 3.1 est donné en figure 3.2. Dans ce cas, le meilleur choix de l'utilisateur (i.e. l'alternative correspondant à sa plus haute préférence) est $a_1 b_1 c_1$ alors que le résultat correspondant à l'alternative la moins préférable est $a_2 b_1 c_1$. Les liens internes sont construits de proche en proche en en *flippant*¹¹ une variable à la fois (en commençant par les nœuds les plus internes du CP-Net) de sa valeur actuelle à sa valeur immédiatement plus préférable étant donnée la valeur de ses parents. Ainsi de $a_2 b_1 c_1$, en flippant la variable B de sa valeur actuelle b_1 à sa valeur immédiatement plus préférable étant donnée a_2 , soit b_2 , on obtient $a_2 b_2 c_1$. En flippant la valeur C de sa valeur c_1 à sa valeur immédiatement plus préférable étant donnée a_2 , soit donc c_2 , on obtient $a_2 b_1 c_2$. L'alternative immédiatement plus préférable à $a_2 b_1 c_2$ s'obtient alors en flippant la valeur de C , de c_1 à c_2 , ce qui donne l'alternative $a_2 b_2 c_2$. Cette même sortie est immédiatement plus préférable à $a_2 b_1 c_2$. A partir de $a_2 b_2 c_2$, on construit l'alternative immédiatement plus préférable en flippant A , de sa valeur courante a_2 à sa valeur immédiatement plus préférable a_1 , ce qui donne l'alternative $a_1 b_2 c_2$, à partir de laquelle, en procédant comme précédemment, on construit de proche en proche les alternatives immédiatement plus préférables en flippant une à une les variables à leurs plus préférables valeurs étant donnée la valeur de leur parent A , ce qui nous mène à l'alternative la plus préférable du CP-Net., soit $a_1 b_1 c_1$.

La sémantique des CP-Nets est simple, définie en termes d'ensembles d'ordres de préférence qui sont consistants avec l'ensemble des contraintes imposées par les tables CPT.

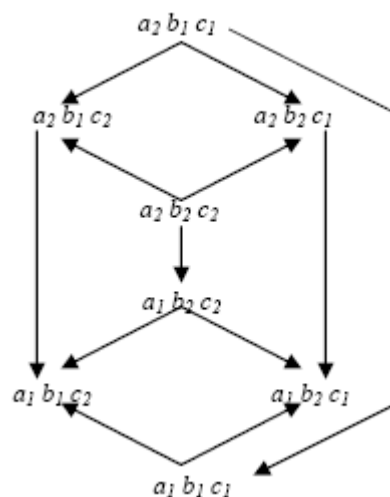


FIGURE 3.2 : Graphe de préférences induit.

¹¹ Le terme est utilisé dans [Boutilier et al., 99] pour désigner une transition de la valeur actuelle d'une variable à sa valeur immédiatement plus préférable (improving flipping) ou immédiatement moins préférable (worsening flipping)

3.3.3 Les UCP-Nets

Le formalisme UCP-Net (*Utility CP-Net*) [Boutilier et al., 01b] est une extension du modèle CP-Net qui permet de représenter l'information sur l'utilité quantitative plutôt que de simples ordres qualitatifs. Le formalisme est basé sur la notion d'indépendance généralisée additive.

Définition formelle d'un UCP-Net Soient $V = \{X_1, X_2, \dots, X_k\}$ un ensemble de variables donné, $f_{X_i}(X_i, U_i)$ (tel que $U_i = Pa(X_i)$) une quantification définie pour chaque famille (X_i, U_i) , et u une fonction d'utilité sur un ordre de préférences \succeq . Un UCP-Net est un graphe orienté acyclique (ou DAG) G sur V qui vérifie les conditions suivantes :

1. $u(X_1, X_2, \dots, X_n) = \sum_{i=1..k} f_{X_i}(X_i, U_i)$
2. Le DAG G est un CP-Net valide pour \succeq (i.e. \succeq satisfait le CP-Net).

Exemple

Le CP-Net de la figure 3.1 peut être étendu en incluant un facteur pour chaque famille du graphe : $f_A(A), f_B(B, A), f_C(C, A)$ tels que $f_B(B, A)$ (respectivement $f_C(C, A)$) s'interprète comme l'utilité de B (respectivement de C) étant donné A . En particulier, nous avons :

$$\begin{aligned} f_A(a_1) &= 0.97 & ; & & f_A(a_2) &= 0.56 \\ f_B(b_2, a_1) &= 0.13 & ; & & f_B(b_2, a_2) &= 0.56 \\ f_C(c_1, a_1) &= 0.76 & ; & & f_C(c_1, a_2) &= 0.20 \end{aligned}$$

Sémantiquement, ces différents facteurs sont GAI, d'où :

$$u(A, B, C) = f_A(A) + f_B(B, A) + f_C(C, A)$$

Plus particulièrement :

$$u(a_i, b_j, c_k) = f_A(a_i) + f_B(b_j, a_i) + f_C(c_k, a_i), \forall a_i \in Dom(A), b_j \in Dom(B), c_k \in Dom(C)$$

d'où :

$$\begin{aligned} u(a_1, b_2, c_1) &= f_A(a_1) + f_B(b_2, a_1) + f_C(c_1, a_1) = 0.97 + 0.13 + 0.76 = 1.86 \\ u(a_2, b_2, c_1) &= f_A(a_2) + f_B(b_2, a_2) + f_C(c_1, a_2) = 0.56 + 0.56 + 0.20 = 1.32 \end{aligned}$$

Chacun des facteurs sert à quantifier la table *CPT* dans le graphe. Le graphe UCP-Net obtenu est donné en figure 3.3.

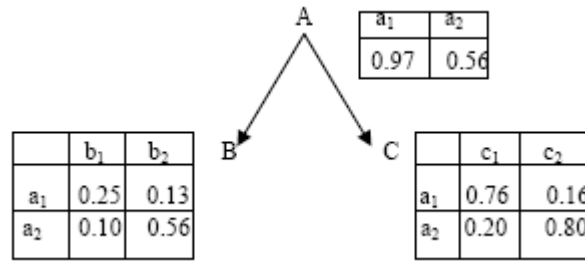


FIGURE 3.3 : Un exemple de UCP-Net.

Une condition nécessaire et suffisante pour qu'un DAG quantifié soit un UCP-Net valide est donnée par la proposition suivante [Boutilier et al., 01b] :

Proposition : Soit G un DAG sur X_1, X_2, \dots, X_n dont les facteurs reflètent la structure GAI d'une fonction d'utilité u . Alors G est un UCP-Net valide ssi chaque variable X_i domine ses descendants.

La relation de dominance est formellement définie à travers la définition suivante : étant donnée une variable X dans un DAG quantifié, tel que pour la famille $(X, U) / U = Pa(X)$ est l'ensemble des parents de X , on définit la quantification $f_X(X, U)$. Et soient $Y = \{Y_1, Y_2, \dots, Y_m\}$ l'ensemble des descendants de X , $Z_i = Pa(Y_i)$ l'ensemble des parents de Y_i excluant X et tout élément dans U , $Z = \bigcup_i Z_i$, et U_i un sous ensemble de variables dans U qui sont parentes de Y_i (la relation entre ces variables est montrée en figure 3.4).

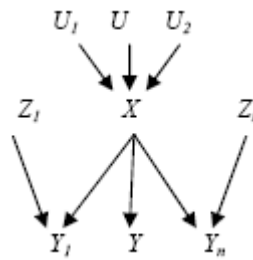


FIGURE 3.4 : Famille étendue de X

Définition de la dominance X domine ses descendants étant donné $u \in Dom(U)$ si:

$$\forall x_1, x_2 \in Dom(X) / f_X(x_1, u) \geq f_X(x_2, u), \quad \forall z \in Dom(Z), \forall y_{i(i=1..n)} \in Dom(Y),$$

$$f_X(x_1, u) - f_X(x_2, u) \geq \sum_i f_{Y_i}(y_i, x_2, u_i, z_i) - f_{Y_i}(y_i, x_1, u_i, z_i) \quad (3.2)$$

X domine ses descendants si la relation (3.2) est vraie $\forall u \in Dom(U)$.

Déterminer si un DAG quantifié est un UCP-Net implique d'examiner chaque famille étendue du CP-Net (Une famille étendue désigne une variable du CP-Net, l'ensemble de ses parents et l'ensemble de ses descendants). Le nombre de tests impliqués est exponentiel en taille des familles étendues du CP-Net. Plusieurs conditions suffisantes fortes existent cependant pour démontrer plus simplement qu'un DAG est un UCP-Net valide. Une de ces conditions suffisantes est donnée dans [Boutilier et al., 01b] comme suit :

Proposition : Soit G un DAG quantifié sur un ensemble de variables $V = \{X_1, X_2, \dots, X_n\}$. Pour chaque variable X , soit U l'ensemble de ses parents : Pour x_1, x_2 appartenant à $Dom(X)$, on définit :

$$\begin{aligned} Minspan(x_1, x_2) &= \min_{u \in Dom(U)} (|f_X(x_1, u) - f_X(x_2, u)|) \\ Minspan(X) &= \min_{x_1, x_2 \in Dom(X)} Minspan(x_1, x_2) \\ Maxspan(x_1, x_2) &= \max_{u \in Dom(U)} (|f_X(x_1, u) - f_X(x_2, u)|) \\ Maxspan(X) &= \max_{x_1, x_2 \in Dom(X)} Maxspan(x_1, x_2) \end{aligned} \quad (3.3)$$

Alors G est un UCP-Net si :

$$\forall X \in V, Minspan(X) \geq \sum_i Maxspan(Y_i), Y_i \text{ étant les descendants de } X \quad (3.4)$$

Intuitivement, il s'agit de montrer que toute variable domine ses descendants. La dominance est ici exprimée par le fait que la plus petite différence entre deux valeurs quelconques de X , étant donnée n'importe quelle valeur de ses parents, est supérieure ou égale à la somme des plus grandes différences entre deux valeurs quelconques de chacun de ses descendants.

Nous avons présenté dans cette section les fondements théoriques du modèle CP-Net, et explicité sa sémantique. Puis, nous avons défini son extension à l'utilisation de valeurs d'utilité, conduisant au formalisme UCP-Net. Les CP-Nets ont été utilisés avec succès dans divers problèmes décisionnels (voir annexe B). Nous nous proposons en section suivante, de leur définir un cadre d'utilisation dans le contexte de la RI.

3.4 Modèle de RI basé CP-Nets

Les préférences conditionnelles constituent la forme la plus usuelle et la plus intuitive des préférences humaines. Ces préférences ne sont pas spécifiquement prises en charge dans les SRI. Il est certes possible de les traduire dans une formulation booléenne dans laquelle les critères de recherche sont pondérés pour traduire l'ordre

de préférences sous-jacent. Cependant, comme nous l'avons montré en section 3.1, il n'existe pas de méthode pour pondérer correctement de telles préférences, et une pondération aléatoire peut conduire à des contradictions. C'est dans l'objectif de résoudre ce problème, que nous proposons à travers la présente contribution, un modèle de RI flexible basé sur les CP-Net. En particulier, nous proposons :

1. une approche de représentation CP-Net de requêtes préférentielles exprimant des préférences qualitatives de l'utilisateur,
2. une approche de pondération automatique des requêtes CP-Nets,
3. une approche d'évaluation des requêtes CP-Nets.

Nous présentons ces approches respectivement en paragraphes 3.3.1, 3.3.2 et 3.3.3 suivants.

3.4.1 Représentation CP-Net des requêtes préférentielles

Pour formuler sa requête, l'utilisateur doit préalablement spécifier un ensemble de caractéristiques (ou variables) sur lesquelles vont porter ses préférences. Chaque caractéristique est définie sur son propre domaine de valeurs (une valeur est un terme de la requête). Pour chaque variable donnée X , l'utilisateur doit spécifier toutes ses dépendances préférentielles, ainsi que l'ordre de préférences correspondant sur $\text{Dom}(X)$.

Cette description est utilisée pour construire le CP-Net requête : les nœuds du CP-Net sont les variables sur lesquelles portent les préférences utilisateur, les liens entre les nœuds définissent les dépendances préférentielles spécifiées par l'utilisateur (On supposera dans ce qui suit que le graphe résultant est un DAG). L'ordre de préférences sur un domaine de valeurs est traduit en table *CPT*.

La figure 3.5 illustre le CP-Net correspondant à la requête (3.1) (section 3.1). Les variables concernées sont *City*, *Housing* et *Place* telles que :

$$\text{Dom}(\textit{City}) = \{\textit{Paris}, \textit{Lyon}\},$$

$$\text{Dom}(\textit{Housing}) = \{\textit{RH}, \textit{Studio}\}$$

$$\text{Dom}(\textit{Place}) = \{\textit{Center}, \textit{Suburbs}\}.$$

En outre, $CPT(\textit{City})$ spécifie que *Paris* est inconditionnellement préférable à *Lyon* ($\textit{Paris} \succ \textit{Lyon}$), tandis que $CPT(\textit{Housing})$ par exemple, spécifie un ordre de préférence sur les valeurs de *Housing*, sous la condition des valeurs prises par la variable *City* (ainsi par exemple, si *Paris* alors $\textit{RH} \succ \textit{Studio}$).

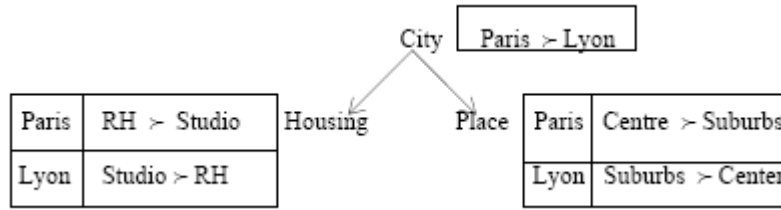


FIGURE 3.5 : Représentation CP-Net d'une requête booléenne

La requête CP-Net est ensuite pondérée par des facteurs d'utilité (poids de préférence). Notre processus de pondération automatique de la requête CP-Net correspond à la génération du UCP-Net correspondant et est basé sur la propriété de dominance (formule (3.4)) (énoncée en section 3.3.3), nous le présentons ci-après.

3.4.2 Pondération automatique de la requête

Pondérer la requête CP-Net, revient à traduire les ordres de préférences qualitatives portées par les tables CPT du CP-Net requête, en valeurs de préférences quantitatives. Cela revient donc à quantifier le CP-Net requête. L'extension naturelle des CP-Nets aux valeurs numériques de préférences est donnée par le modèle UCP-Net. Pondérer la requête CP-Net revient donc à générer le UCP-Net correspondant.

Notre approche de génération automatique du UCP-Net requête est basée sur les propriétés suivantes :

1. Toute variable X doit vérifier la propriété de dominance (formule (4)).
2. Un ordre de préférences sur $Dom(X)$, étant donnée une valeur $u \in Dom(Pa(X))$, est traduit par une distribution uniforme des valeurs d'utilités (ou degrés de préférence) sur $Dom(X)$ étant donnée u . Intuitivement, il s'agit de distribuer uniformément des degrés de préférences sur les valeurs x_i de X , de sorte à traduire numériquement les ordres de préférence qualitatifs introduits sur les x_i dans le CP-Net considéré. Ainsi, si par exemple, une variable X , apparaît dans le CP-Net avec deux valeurs x_1 et x_2 telles que $x_1 \succ x_2$, ceci se traduit dans notre approche par : $f_X(x_2)=0$ et $f_X(x_1)=1$. Pour une variable X à trois valeurs telle que $x_1 \succ x_2 \succ x_3$, on obtient: $f_X(x_3)=0$, $f_X(x_2)=1/2$ et $f_X(x_1)=2/2$. Pour respecter la propriété de dominance à la base de tout UCP-Net, on impose en outre une condition supplémentaire sur les degrés de préférences associés aux variables représentant les nœuds internes du CP-Net.

L'approche est formellement définie dans ce qui suit.

Soit X un nœud de la requête CP-Net, tel que $|Dom(X)|= k$, et soit $u(i)$ le degré de préférence d'ordre i (en supposant un degré de préférence croissant lorsque i croît) sur les valeurs de X .

1. Pour tout nœud feuille X , nous générons les utilités sur $Dom(X)$, suivant la propriété 1, comme suit :

$$u(1) = 0 \text{ et } u(i) = u(i - 1) + (1 / k - 1), \quad \forall 1 < i \leq k \quad (3.5)$$

2. Tout nœud interne X , possède des descendants, et doit donc respecter la propriété de dominance (propriété 2 énoncée plus haut). Pour tout nœud interne X du CP-Net, on calcule alors la quantité :

$$S = \sum_i Maxspan(B_i)$$

où les B_i sont les descendants de X .

Comme X doit dominer ses descendants on impose que :

$$Minspan(X) \geq S$$

Plusieurs valeurs répondent à la condition, nous choisissons la plus petite, S , et posons $Minspan(X) = S$.

Nous générons alors les utilités du nœud interne X comme suit :

$$u(1) = 0 \text{ et } u(i) = u(i - 1) + S, \quad \forall 1 < i \leq k \quad (3.6)$$

On calcule alors $Minspan(X)$ et $Maxspan(X)$ de manière triviale comme suit:

$$Minspan(X) = |u(i+1) - u(i)| \text{ et } Maxspan(X) = |u(k) - u(1)| \quad (3.7)$$

Les valeurs d'utilité obtenues pouvant être supérieures à 1 (cas des nœuds internes), nous proposons une normalisation des facteurs d'utilité individuels du CP-Net et des utilités globales de chacune de ses alternatives en divisant chaque valeur d'utilité du CP-Net par l'utilité globale de l'alternative la plus préférée.

Illustration

Nous nous proposons de construire l'UCP-Net correspondant au CP-Net de la Figure 3.5, en utilisant notre algorithme de pondération.

Chacun des nœuds feuilles, *Housing* et *Place*, est défini sur deux valeurs, auxquelles on associe les degrés de préférences $u(1)=0$ pour la valeur la moins préférée, et $u(2) = 1$ pour la valeur la plus préférée. Ce qui donne respectivement :

$$\begin{aligned} f_{Housing}(RH, Paris) = 1 & \quad ; \quad f_{Housing}(Studio, Paris) = 0 \\ f_{Housing}(RH, Lyon) = 0 & \quad ; \quad f_{Housing}(Studio, Lyon) = 1 \end{aligned}$$

Et

CHAPITRE 3. MODELE DE RI FLEXIBLE BASE SUR LES CP-NETS

$$f_{Place}(Center, Paris) = 1 \quad ; \quad f_{Place}(Suburbs, Paris) = 0$$

$$f_{Place}(Center, Lyon) = 0 \quad ; \quad f_{Place}(Suburbs, Lyon) = 1$$

Concernant le nœud racine *City*, qui possède deux descendants, *Housing* et *Place*, on calcule d'abord : $S = Maxspan(Housing) + Maxspan(Place)$. On a :

$$Maxspan(Housing) = 1 - 0 = 1$$

$$Maxspan(Place) = 1 - 0 = 1$$

D'où $S=2$, ce qui implique les degrés de préférences suivants sur les deux valeurs de la variable *City* : $u(1)=0$ pour la valeur la moins préférable, et $u(2) = 0+S=2$ pour la valeur la plus préférable. Ce qui donne :

$$f_{City}(Paris) = 2 \quad ; \quad f_{City}(Lyon) = 0$$

En normalisant l'ensemble des quantifications ainsi obtenues par la quantification de la meilleure alternative (soit donc par $u(Paris, RH, Center) = 2 + 1 + 1 = 4$), on obtient les valeurs suivantes :

$$f_{Housing}(RH, Paris) = 0.25 \quad ; \quad f_{Housing}(Studio, Paris) = 0$$

$$f_{Housing}(RH, Lyon) = 0 \quad ; \quad f_{Housing}(Studio, Lyon) = 0.25$$

$$f_{Place}(Center, Paris) = 0.25 \quad ; \quad f_{Place}(Suburbs, Paris) = 0$$

$$f_{Place}(Center, Lyon) = 0 \quad ; \quad f_{Place}(Suburbs, Lyon) = 0.25$$

$$f_{City}(Paris) = 0.5 \quad ; \quad f_{City}(Lyon) = 0$$

Ces valeurs sont utilisées pour construire l'UCP-Net présenté en figure 3.6.

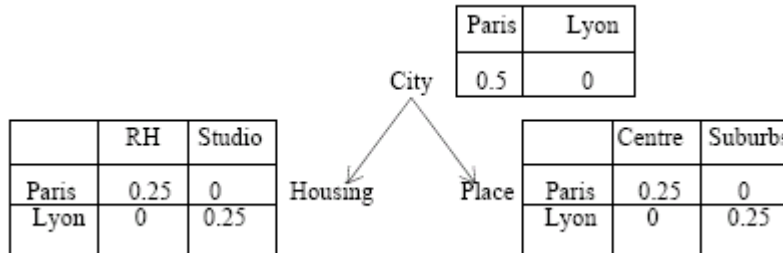


FIGURE 3.6 : L'UCP-Net requête

Les différents facteurs d'utilité étant GAI (selon les fondements des UCP-Nets), nous avons alors à titre d'exemple :

$$u(Paris, Studio, Center) = 0.5 \text{ et } u(Lyon, studio, Center) = 0.25$$

traduisant le fait que la première alternative est préférée à la seconde (alors que la pondération aléatoire telle que présentée dans l'exemple en section 3.1 produisait un

résultat contradictoire). L'UCP-Net ainsi obtenu peut alors être traduit en la requête booléenne pondérée correcte suivante :

$$(Paris\ 0.5 \wedge (RH\ 0.25 \vee Studio\ 0)) \wedge (Center\ 0.25 \vee Suburbs\ 0)) \vee (Lyon\ 0 \wedge (RH\ 0 \vee Studio\ 0.25) \wedge (Center\ 0 \vee Suburbs\ 0.25)).$$

La requête pondérée ainsi obtenue est alors évaluée. Nous présentons notre approche d'évaluation des requêtes en paragraphe suivant.

3.4.3 Evaluation de la requête CP-Net

Le but de l'évaluation est de calculer le degré de pertinence des documents pour la requête. L'objectif est de classer les documents par ordre de pertinence et de retourner à l'utilisateur, les documents les plus pertinents pour sa requête. Notre approche d'évaluation est basée sur les étapes suivantes :

- le processus de recherche est lancé dans un premier temps sur l'ensemble des termes de la requête CP-Net sans tenir compte de la pondération au préalable. Le résultat est une liste de documents pertinents probables pour la requête,
- les documents retrouvés sont ensuite représentés par des CP-Nets, puis documents et requêtes sont reformulés en expressions booléennes.
- Un processus d'évaluation calcule la valeur de pertinence de tels documents pour la requête UCP-Net, et ordonne les documents par degré de pertinence.

La première étape est une recherche classique. Les documents qui appartiennent des termes de la requête sont alors retournés par le système. Les étapes 2 et 3 sont propres à notre approche, nous les décrivons dans ce qui suit.

3.4.3.1 Le document vu comme un CP-Net

Partant de la constatation que seuls les termes du document qui s'apparient avec les termes de la requête participent à l'évaluation de la pertinence de ce document pour la requête, chaque document supposé pertinent pour une requête $Q = (V, E)$ est représenté par un CP-Net $D = (V', E')$ dans le même espace de termes que la requête. On réalise ainsi une projection du document sur l'espace de la requête.

La topologie correspondante est semblable à celle du CP-Net requête $Q = (V, E)$ mais les tables CPT sont différentes. En effet, les CPT dans le CP-Net document $D = (V', E')$ quantifient numériquement l'importance des termes d'indexation dans D . Cette importance se traduit par les poids des termes correspondants dans D . Les poids sont généralement exprimés par une variante de $tf*idf$.

Le document (respectivement la requête) est alors traduit en expression booléenne, comme une disjonction de conjonctions. Chaque conjonction étant construite sur l'ensemble des

CHAPITRE 3. MODELE DE RI FLEXIBLE BASE SUR LES CP-NETS

éléments du produit cartésien $Dom(X_1) \times Dom(X_2) \times \dots \times Dom(X_n)$ où les X_i ($1 \leq i \leq n$) sont les noeuds du CP-Net document (respectivement CP-Net requête).

Formellement, on a :

$$D = \vee_{j_i} (\wedge_i (t_{i,j_i}, p_{i,j_i})) \quad (3.8)$$

$$Q = \vee_{j_i} (\wedge_i (t_{i,j_i}, f_{i,j_i})) \quad (3.9)$$

Où

$t_{i,j_i} \in Dom(X_i)$, $1 \leq i \leq n$, $1 \leq j_i \leq |Dom(X_i)|$,

p_{i,j_i} est le poids de t_{i,j_i} dans D (basé sur sa fréquence d'occurrence),

f_{i,j_i} est le poids du terme t_{i,j_i} (son utilité) dans Q étant donnée une valeur de ses parents.

Soit $m = |Dom(X_1)| \times |Dom(X_2)| \times \dots \times |Dom(X_n)|$, en posant $\wedge_i t_{i,j_i} = T_k$, $1 \leq k \leq m$, les représentations (2) et (3) sont respectivement réduites à :

$$D = \vee_k (T_k, S_k) = \vee (T_k, S_k) \quad (3.10)$$

$$Q = \vee_k (T_k, U_k) = \vee (T_k, U_k) \quad (3.11)$$

Où S_k et U_k sont les poids agrégés des valeurs p_{i,j_i} respectivement introduits dans (3.8) et (3.9). S_k et U_k sont calculés comme suit :

Calcul des U_k

Puisque les facteurs f_{i,j_i} dans le UCP-Net requête sont GAI, leur agrégation est additive et donnée par :

$$U_k = \sum_i f_{i,j_i} \quad (3.12)$$

Calcul des S_k

Une valeur S_k définit le poids d'une conjonction de termes t_{i,j_i} dans le document. Ce poids est mesuré comme agrégation des poids individuels des termes correspondants. Le poids d'un terme dans un document est classiquement défini, sur la base de ses statistiques d'occurrences, comme mesure de $tf*idf$. Cependant, la projection du document dans l'espace de la requête et sa représentation dans un espace topologique similaire, introduit une nouvelle dimension d'importance pour les termes du document, selon la sémantique du CP-Net. Les nœuds parents dans un graphe CP-Net sont plus importants que leurs descendants. Pour tenir compte de cette importance

(que l'on appellera importance de position), nous proposons de calculer les poids agrégés S_k , comme moyenne pondérée des poids p_{i,j_i} de termes t_{i,j_i} dans le document, comme suit :

On attribue d'abord une valeur d'importance de position G_X à chacun des noeuds X du CP-Net document selon leurs niveaux respectifs dans le graphe :

1. pour tout nœud feuille X , $G_X = 1$,
2. pour tout nœud interne X , notons B_l les descendants de X et G_{B_l} leurs ordres d'importance respectifs, on a :

$$G_X = \max_l (G_{B_l}) + 1 \quad (3.13)$$

Le poids agrégé S_k introduit dans (3.10) est alors donné par :

$$S_k = \frac{\sum_i (p_{i,j_i} * G_{X_i})}{\sum_i G_{X_i}} \quad (3.14)$$

Où X_i est le noeud contenant le terme t_{i,j_i} de D .

Illustration

En supposant par exemple, que le processus de recherche lancé initialement sur la requête donnée en figure 3.6, retourne le document D_1 présenté en tableau 3.1, où chaque paire (t, p) représente le terme et le poids associé dans le document respectivement. L'UCP-Net associé au document D_1 est obtenu en représentant le document dans le même espace de termes que la requête. La topologie du graphe obtenu est identique à celle du CP-Net requête.

$D_1 ((Paris, 0.7), (Lyon, 0.5), (RH, 0.2))$
--

TABLEAU 3.1 : Document retourné

Bien que la notion de préférences entre les termes du document soit ici absente, nous appellerons le graphe ainsi obtenu, un CP-Net document par analogie au CP-Net requête à partir duquel il est calqué. Le CP-Net document correspondant est donné en figure 3.7.

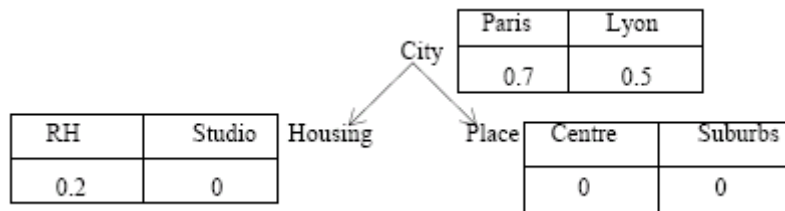
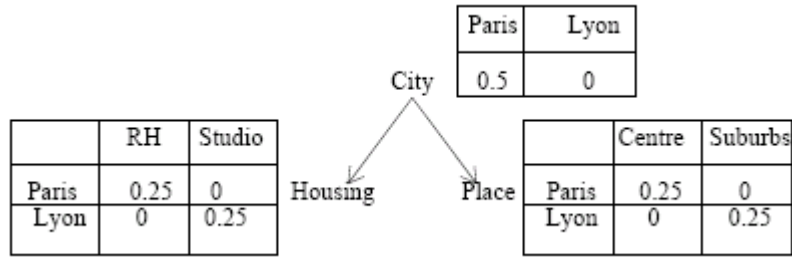


FIGURE 3.7 : D_1 vu comme un UCP-Net

CHAPITRE 3. MODELE DE RI FLEXIBLE BASE SUR LES CP-NETS

En considérant la requête UCP-Net introduite en figure 3.6 comme suit :



Cette requête ainsi que le document UCP-Net introduit en figure 3.7 sont interprétés respectivement en utilisant les formules (3.8) et (3.9), comme suit :

$$Q = ((Paris,0.5) \wedge (RH,0.25) \wedge (Center,0.25)) \vee ((Paris,0.5) \wedge (RH,0.25) \wedge (Suburbs,0)) \vee ((Paris,0.5) \wedge (Studio,0) \wedge (Center,0.25)) \vee ((Paris,0.5) \wedge (Studio,0) \wedge (Suburbs,0)) \vee ((Lyon,0) \wedge (RH,0) \wedge (Center,0)) \vee ((Lyon,0) \wedge (RH,0) \wedge (Suburbs,0.25)) \vee ((Lyon,0) \wedge (Studio,0.25) \wedge (Center,0)) \vee ((Lyon,0) \wedge (Studio,0.25) \wedge (Suburb,0.25)) .$$

$$D_1 = ((Paris, 0.7) \wedge (RH, 0.2) \wedge (center, 0)) \vee ((Paris, 0.7) \wedge (RH, 0.2) \wedge (suburbs,0)) \vee ((Paris, 0.7) \wedge (Studio, 0) \wedge (center, 0)) \vee ((Paris, 0.7) \wedge (Studio, 0) \wedge (suburbs,0)) \vee ((Lyon, 0.5) \wedge (RH, 0.2) \wedge (center, 0)) \vee ((Lyon, 0.5) \wedge (RH, 0.2) \wedge (suburbs,0)) \vee ((Lyon, 0.5) \wedge (Studio, 0) \wedge (center, 0)) \vee ((Lyon, 0.5) \wedge (Studio, 0) \wedge (suburbs,0)) .$$

Pour chaque conjonction de la requête, le poids global est calculé comme la somme des facteurs d'utilités individuels de chacun de ses termes selon la formule (3.12). Ainsi par exemple, le poids de la conjonction $T_1 = (Paris,0.5) \wedge (RH,0.25) \wedge (Center,0.25)$ égale $0.5+0.25+0.25$ soit 1.

Pour chaque conjonction correspondante dans le document, le poids global est calculé selon la formule (3.14), comme somme des poids individuels des termes de la conjonction pondérés par le niveau d'importance associé au nœud correspondant dans le CP-Net document, tel que défini dans la formule (3.13). Ainsi pour calculer le poids d'une conjonction T_k , on calcule d'abord les poids d'importance associés aux différents nœuds du graphe CP-Net document, soit selon la formule (3.13) :

$$G_{Housing} = G_{Place} = 1 \quad \text{et} \quad G_{City} = 1+1 = 2$$

Ainsi par exemple, le poids de la conjonction $T_1 = (Paris,0.7) \wedge (RH,0.2) \wedge (Center,0)$ égale :

$$\frac{G_{City} * 0.7 + G_{Housing} * 0.2 + G_{Place} * 0}{G_{City} + G_{Housing} + G_{Place}} = \frac{1.4 + 0.2 + 0}{2 + 1 + 1} = \frac{1.6}{4} = 0.4$$

En effectuant ces calculs sur l'ensemble des conjonctions tant dans la requête UCP-Net Q que dans le document UCP-Net D_1 , on aboutit aux représentations disjonctives de la requête et du document respectivement définies comme suit:

$$Q = (T_1, 1) \vee (T_2, 0.75) \vee (T_3, 0.75) \vee (T_4, 0.5) \vee (T_5, 0) \vee (T_6, 0.25) \vee (T_7, 0.25) \vee (T_8, 0.5).$$

$$D_1 = (T_1, 0.4) \vee (T_2, 0.4) \vee (T_3, 0.35) \vee (T_4, 0.35) \vee (T_5, 0.3) \vee (T_6, 0.3) \vee (T_7, 0.25) \vee (T_8, 0.25).$$

Où T_i , $1 \leq i \leq 8$ est donné dans le tableau 3.2, Les poids des T_i dans la requête Q (respectivement dans le document D_1) sont calculés en utilisant la formule (3.12) (respectivement (3.14)).

$$\begin{aligned} T_1 &= (Paris \wedge RH \wedge Center) & T_2 &= (Paris \wedge RH \wedge Suburbs) \\ T_3 &= (Paris \wedge Studio \wedge Center) & T_4 &= (Paris \wedge Studio \wedge Suburbs) \\ T_5 &= (Lyon \wedge RH \wedge Center) & T_6 &= (Lyon \wedge RH \wedge Suburbs) \\ T_7 &= (Lyon \wedge Studio \wedge Center) & T_8 &= (Lyon \wedge Studio \wedge Suburbs) \end{aligned}$$

TABLEAU 3.2 : Sous-requêtes conjonctives

3.4.3.2 Evaluation de la requête

Soit Q un CP-Net requête exprimée sous forme d'une expression booléenne en forme normale disjonctive telle que présentée dans la formule (3.11), et D le document retourné est exprimé sous forme booléenne tel que défini dans la formule (3.10). Pour évaluer la pertinence du document D pour la requête pondérée Q , soit $RSV(D, Q)$, différentes formules de calcul peuvent être utilisés. En particulier, les opérateurs d'agrégation somme, moyenne, moyenne pondérée, moyenne pondérée ordonnée (OWA) (introduits en section 1.3.2.2), ou des opérateurs de tri tels que définis dans l'approche de [Loiseau et al., 07] (section 1.3.3) peuvent être utilisés. Nous proposons pour notre cas d'adapter et d'utiliser l'opérateur du minimum pondéré [Dubois et al., 86 ; Yager, 87] comme suit :

Soit U_k le poids d'importance de T_k dans Q , $F(D, T_k) = S_k$ le poids de T_k dans le document D , on note $RSV_{T_k}(F(D, T_k), U_k)$ la fonction d'évaluation de T_k pour le document D . Les

CHAPITRE 3. MODELE DE RI FLEXIBLE BASE SUR LES CP-NETS

différentes conjonctions pondérées (T_k, U_k) étant liées par une disjonction, ce qui donne:

$$RSV_{T_k}(F(D, T_k), U_k) = \min(S_k, U_k) \quad (3.15)$$

$RSV(D, Q)$ est alors obtenue par aggrégation de l'ensemble des poids de pertinence calculés dans (3.15) comme suit :

$$RSV(D, Q) = \max_k(\min(S_k, U_k)) \quad (3.16)$$

Illustration

En considérant le document et la requête de l'illustration précédente (section 3.4.3.1), soit :

$$Q = (T_1, 1) \vee (T_2, 0.75) \vee (T_3, 0.75) \vee (T_4, 0.5) \vee (T_5, 0) \vee (T_6, 0.25) \vee (T_7, 0.25) \vee (T_8, 0.5) .$$

$$D_I = (T_1, 0.4) \vee (T_2, 0.4) \vee (T_3, 0.35) \vee (T_4, 0.35) \vee (T_5, 0.3) \vee (T_6, 0.3) \vee (T_7, 0.25) \vee (T_8, 0.25) .$$

et en utilisant les égalités (3.15) et (3.16), on calcule les pertinences partielles du document D_I pour chaque sous-requête T_k comme suit :

$$\begin{aligned} RSV_{T_1}(F(D, T_1), U_1) &= \min(S_1, U_1) = \min(0.4; 1) = 0.4 \\ RSV_{T_2}(F(D, T_2), U_2) &= \min(S_2, U_2) = \min(0.4; 0.75) = 0.4 \\ RSV_{T_3}(F(D, T_3), U_3) &= \min(S_3, U_3) = \min(0.35; 0.75) = 0.35 \\ RSV_{T_4}(F(D, T_4), U_4) &= \min(S_4, U_4) = \min(0.35; 0.5) = 0.4 \\ RSV_{T_5}(F(D, T_5), U_5) &= \min(S_5, U_5) = \min(0.3; 0) = 0 \\ RSV_{T_6}(F(D, T_6), U_6) &= \min(S_6, U_6) = \min(0.3; 0.25) = 0.25 \\ RSV_{T_7}(F(D, T_7), U_7) &= \min(S_7, U_7) = \min(0.25; 0.25) = 0.25 \\ RSV_{T_8}(F(D, T_8), U_8) &= \min(S_8, U_8) = \min(0.25; 0.5) = 0.25 \end{aligned}$$

La pertinence globale du document pour la requête disjonctive Q est alors calculé comme le maximum des pertinences partielles ainsi calculés. L'ensemble des résultats obtenus est résumé dans le Tableau 3.3. Le document D_I peut ainsi être ordonné soit partiellement selon sa pertinence partielle pour chaque sous-requête T_k , ou globalement selon sa pertinence globale pour la requête $Q = \vee T_k$.

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	$GRSV^{12}$
D_1	0.4	0.4	0.35	0.35	0	0.25	0.25	0.25	0.4

TABLEAU 3.3 : Pertinences partielles et totale du document D_1

3.5 Conclusion

Nous avons présenté dans ce chapitre notre approche de RI flexible basée sur les CP-Nets. Cette approche est fondée d'une part sur l'expression de requêtes flexibles traduisant les préférences d'un utilisateur, utilisant les CP-Nets. Le formalisme utilisé est graphique et qualitatif ce qui permet une formulation naturelle et intuitive et une représentation simple et compacte des préférences. Le formalisme qualitatif possède une puissance d'expression élevée mais décline en puissance de calcul. Nous avons proposé de l'allier aux utilités. Les utilités, représentant des poids d'importance conditionnelle des termes de la requête, sont calculées automatiquement. L'utilisateur est ainsi déchargé de cette tâche fastidieuse et non moins improbable de pondération, et les poids générés sont corrects puisque basés sur des fondements théoriques établis (validité d'un UCP-Net). D'autre part, cette approche est fondée sur l'évaluation flexible, des requêtes dans la sémantique CP-Net basée sur l'utilisation d'un opérateur d'agrégation flexible, en l'occurrence le minimum pondéré, que nous avons adapté à la sémantique CP-Net. Notons cependant que le graphe CP-Net introduit par l'utilisateur peut être incorrect (inconsistant). Des outils d'aide à la formulation et des outils de correction automatique seraient nécessaires pour garantir la validité des descriptions fournies par l'utilisateur.

¹² Global RSV

Chapitre 4

Approche de RI sémantique

4.1 Introduction

Nous avons présenté dans le chapitre précédent, notre modèle de RI flexible basé sur les CP-Nets. Les représentations des documents et requêtes d'abord formalisées en graphes CP-Nets, sont traduites sous forme booléenne classique. L'appariement document-requête est basé sur les mots clés que leurs représentations respectives appariant. Or, les techniques basées sur les mots clés (dite technique en sacs de mots) mènent aux problèmes cruciaux de disparité des termes (term mismatch) et d'ambiguïté RI. Les approches d'indexation sémantique tentent de pallier ces problèmes en autorisant la représentation du contenu informationnel des documents et requêtes par les sens des mots plutôt que par les mots qu'ils contiennent. Notre contribution présentée dans ce chapitre est liée à ce contexte et porte sur une nouvelle approche de RI sémantique. En particulier, nous proposons une approche d'indexation sémantique des documents et une approche d'évaluation des requêtes.

1. *Approche d'indexation sémantique des documents* : notre approche d'indexation sémantique est basée sur l'utilisation conjointe de WordNet pour la détection des concepts représentatifs du document, et des règles d'association pour la découverte des relations entre ces concepts. Toute approche d'indexation sémantique étant intrinsèquement liée à la désambiguïsation, nous proposons pour notre cas, une nouvelle technique de désambiguïsation basée sur le calcul de scores dépendants d'une part de la distance sémantique des concepts dans l'ontologie et d'autre part de l'importance des termes correspondants dans le document. Nous proposons en outre de découvrir une certaine sémantique latente du texte portée par les associations implicites entre les termes du document. Une telle sémantique est découverte par une technique de fouille de textes [Ahonen et al., 97], en l'occurrence les règles d'association. Le résultat de l'indexation est un ensemble de concepts représentatifs du document, et de relations entre concepts.

2. *Approche d'évaluation des requêtes* : notre approche d'évaluation est basée sur une mesure de similarité des graphes. Pour cela, la requête est exprimée sous forme d'un CP-Net suivant notre approche décrite dans le chapitre 3. Nous proposons alors une technique de construction du CP-Net document à partir de l'ensemble des concepts et relations associées découverts à l'issue de l'étape d'indexation sémantique. La correspondance requête_ document est alors évaluée comme mesure de similarité des graphes CP-Nets correspondants.

Le chapitre est structuré comme suit : En section 4.2, nous présentons les motivations qui ont été à l'origine de nos propositions. En section 4.3, nous présentons les outils sur lesquels est basée notre approche d'indexation sémantique, à savoir WordNet et les règles d'association. Les fondements théoriques de notre approche d'indexation sémantique, un exemple illustratif ainsi que quelques résultats expérimentaux sont donnés en section 4.4. Enfin, la section 4.5 présente notre approche d'évaluation des requêtes CP-Nets.

4.2 Motivations

Les approches d'indexation sémantique permettent de pallier les inconvénients de l'indexation classique basée mots-clés en offrant le moyen de lever l'ambiguïté des mots et les disparités grâce à l'utilisation des sens des mots plutôt que des mots eux-mêmes en tant qu'entités lexicales. L'indexation LSI permet en outre de retrouver une dimension sémantique latente plus abstraite que les sens donnés par un dictionnaire, et portée par une certaine corrélation entre les termes du document. Néanmoins, il reste à notre sens deux points problématiques dans ces approches, sur lesquels nous souhaitons nous pencher :

La désambiguïsation sémantique s'appuie sur le degré de corrélation mutuelle des sens des mots du document. Un score est ainsi associé à chaque sens sur la base de sa distance sémantique cumulée par rapport à tous les sens des autres mots dans le document. Cette distance sémantique ne tient compte que des relations sémantiques des sens, telles que définies dans le dictionnaire ou l'ontologie. Ceci est à notre sens problématique car un sens associé à un terme peut être choisi alors que son score est fortement influencé par sa corrélation avec un ou plusieurs sens associés à des termes de moindre importance dans le document, tandis que si son meilleur sens est moyennement corrélé avec tous les termes importants dans le document, il sera ignoré.

Pour résoudre ce problème, nous proposons une autre approche de calcul du score de désambiguïsation, basée sur la distance sémantique des concepts associés dans l'ontologie et tenant compte du degré d'importance des termes correspondants dans le document. Le degré d'importance des termes est calculé par une mesure de $tf*idf$

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

adaptée aux termes composés. Les termes simples et composés sont découverts dans le document par projection sur l'ontologie.

Les sens des mots fournis par un dictionnaire sont statiques, et ne dépendent pas du contexte local d'utilisation du mot dans le document. L'approche d'indexation par la sémantique latente tente de résoudre ce problème par un *clustering* des mots similaires via une technique de réduction de la dimensionnalité de la matrice termes-documents. L'indexation conceptuelle tente à partir d'une taxonomie conceptuelle extraite du texte, de construire sa sémantique. Les liens entre les différents concepts d'une telle taxonomie sont des liens fonctionnels entre entités lexicales.

Il nous paraît qu'une autre dimension sémantique, celle définissant les liens de dépendance conditionnelle entre les sens des termes, définissant par là même l'orientation sujet, le *topic* du document pourrait être une voie intéressante pour l'exploration de la sémantique du document. Les liens implicites entre les sens des mots (ou concepts) pourraient être exploités pour retrouver non seulement les documents qui traitent des termes ayant le même sens que la requête mais aussi des documents qui traitent de sujets connexes. Nous proposons alors de découvrir la sémantique implicite du texte par le biais d'une technique de découverte de connaissances, à savoir des règles d'association. Plus particulièrement, nous décrivons dans ce chapitre, une approche d'indexation sémantique de documents basée sur les CP-Nets. Les nœuds du CP-Net sont des concepts. Les relations du CP-Net traduisent des dépendances contextuelles entre concepts. En résumé, nous définissons :

1. une approche d'extraction des termes du document,
2. une formule de pondération des termes tenant compte de leur sémantique,
3. une méthode de désambiguïsation des termes basée sur l'utilisation de l'ontologie WordNet,
4. une approche de découverte des relations contextuelles entre concepts via une extension des règles d'association
5. une approche pour combiner les concepts et les relations correspondantes dans une représentation graphique compacte, à savoir le CP-Net. Le formalisme CP-Net est utilisé comme langage d'indexation, pour deux raisons. D'abord, les CP-Nets supportent naturellement l'indexation conceptuelle et offrent un cadre unifié pour organiser de manière relativement compacte et intuitive les concepts et les relations qui les lient. En second lieu, les CP-Nets permettent une représentation plus riche et plus précise des documents puisqu'ils supportent les relations contextuelles et sémantiques entre concepts. Les relations contextuelles sont susceptibles d'améliorer les performances du processus de recherche d'information. Les concepts et les relations sémantiques associées sont susceptibles de résoudre les problèmes de disparité et d'ambiguïté des termes.

Par ailleurs, dans l'objectif de s'orienter vers un modèle « tout CP-Net », et afin d'éviter la traduction des représentations CP-Nets des documents et requêtes dans le paradigme booléen, tel que proposé dans notre première contribution en chapitre 3, nous avons défini une approche d'évaluation de la pertinence des documents pour des requêtes, basée sur une mesure de similarité des graphes CP-Nets correspondants.

Avant de décrire notre approche de RI sémantique, nous présentons tout d'abord les outils sur lesquels elle se base, soit WordNet et les règles d'association.

4.3 Les outils d'aide à l'indexation sémantique

Notre approche d'indexation se base sur l'utilisation de deux principaux outils: WordNet et les règles d'association. WordNet est utilisée pour la détection de concepts et leur désambiguïsation. Les règles d'association permettent la découverte de relations entre ces concepts. Cette section est dédiée à la présentation de ces outils. La section 4.3.1 est dédiée à la présentation de l'ontologie linguistique WordNet. La section 4.3.2 présente les fondements de la découverte des règles d'association.

4.3.1 WordNet

WordNet est un réseau lexical électronique [Fellbaum, 98] développé depuis 1985 à l'université de Princeton par une équipe de psycholinguistes et de linguistes du laboratoire des sciences cognitives de l'université de Princeton, sous la direction de Georges A. Miller. L'avantage de WordNet réside dans la diversité des informations qu'elle contient (grande couverture de la langue anglaise, définition de chacun des sens, ensembles de synonymes, diverses relations sémantiques). En outre, WordNet est librement et gratuitement utilisable. Nous décrivons cette ontologie linguistique que nous utilisons dans la suite de nos travaux sur l'indexation sémantique.

Structure de WordNet

WordNet couvre la majorité des noms, verbes, adjectifs et adverbes de la langue Anglaise, qu'elle structure en un réseau de noeuds et de liens. Les noeuds sont constitués par des ensembles de termes synonymes (appelés *synsets*). Un terme peut être un mot simple ou une collocation (i.e. deux mots ou plusieurs mots reliés par des soulignés pour constituer le mot composé correspondant dans la langue). Un exemple de hiérarchie de synsets correspondant au mot « dog » est donné dans le tableau 4.1.

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

Le tableau 4.2 donne des statistiques¹³ sur le nombre de mots et de concepts dans WordNet dans sa version 3.0.

<p>Noun</p> <p><u>S</u>: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) <i>"the dog barked all night"</i></p> <p><u>S</u>: (n) frump, dog (a dull unattractive unpleasant girl or woman) <i>"she got a reputation as a frump"; "she's a real dog"</i></p> <p><u>S</u>: (n) dog (informal term for a man) <i>"you lucky dog"</i></p> <p><u>S</u>: (n) cad, bounder, blackguard, dog, hound, heel (someone who is morally reprehensible) <i>"you dirty dog"</i></p> <p><u>S</u>: (n) frank, frankfurter, hotdog, hot dog, dog, wiener, wienerwurst, weenie (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)</p> <p><u>S</u>: (n) pawl, detent, click, dog (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)</p> <p><u>S</u>: (n) andiron, firedog, dog, dog-iron (metal supports for logs in a fireplace) <i>"the andirons were too hot to touch"</i></p> <p>Verb</p> <p><u>S</u>: (v) chase, chase after, trail, tail, tag, give chase, dog, go after, track (go after with the intent to catch) <i>"The policeman chased the mugger down the alley"; "the dog chased the rabbit"</i></p>

TABEAU 4.1 : Les concepts de WordNet correspondants au mot dog

Catégorie	Mots	Synsets	Total Paires Mot-Sens
Nom	117 798	82 115	146 312
Verbe	11 529	13 767	25 047
Adjectif	21 479	18 156	30 002
Adverbe	4 481	3 621	5 580
Total	155 287	117 659	206 941

TABEAU 4.2 : Le nombre de mots et de synsets dans WordNet 3.0.

¹³ Statistiques extraites du site web de WordNet : <http://wordnet.princeton.edu/man/wnstats.7WN>

Les concepts de WordNet sont reliés par des relations sémantiques. La relation de base entre les termes d'un même synset est la synonymie. Les différents synsets sont autrement liés par diverses relations sémantiques telles que la subsomption ou relation d'hyponymie-hyperonymie, et la relation de composition méronymie-holonymie. Ces relations sont formellement définies comme suit :

1. la relation **taxonomique** (ou relation de **subsomption**), dite relation d'**Hyperonymie/Hyponymie**:
 X est un *hyponyme* de Y si X est un type de (*kind of / is-a*) Y . Y est alors dit *hyperonyme* de X .

Exemple : {*canine*} a pour hyponymes {*wolf, wild dog, dog*} (selon figure 4.1).

2. la relation d' **Holonymie** et son inverse la **Méronymie** :

X est un *méronyme* de Y si X est une partie constituante (*part of*), substance de (*substance of*) ou membre (*member of*) de Y . Y est alors dit un *holonyme* de X .

Exemple : {*car*} a pour *méronymes* {*wheel, engine, ...*}.

La figure 4.1 donne un exemple de sous-hiérarchie de WordNet correspondant au concept dog (RFIEC¹⁴).

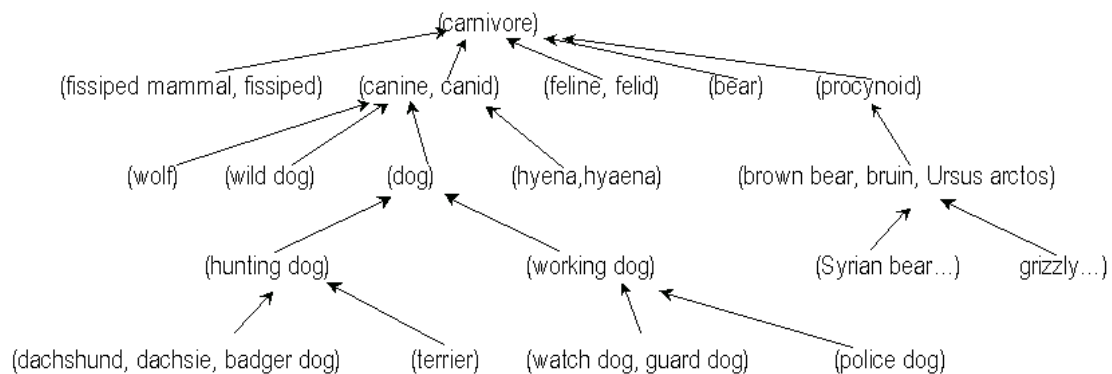


FIGURE 4.1: Sous hiérarchie de WordNet correspondant au concept "dog"

Vu sa large couverture de la langue anglaise, sa disponibilité et sa gratuité, WordNet a été largement utilisée en RI pour l'étiquetage sémantique de corpus (corpus Semcore¹⁵, [Miller et al., 93]), pour l'expansion des requêtes utilisateur par ajout de

¹⁴ <http://www.irit.fr/RFIEC/wordnet/wordnet.htm>

¹⁵ Le corpus SemCore (*Semantic Concordance Corpus*) est un sous-ensemble du corpus BROWN d'environ 100 documents, totalisant quelques 700 000 mots dont plus de 230 000 sont lemmatisés et étiquetés avec les sens de WordNet.

synonymes ou de toutes autres relations sémantiques [Smeaton et al., 95 ;Voorhees, 94 ;[Voorhees, 98], mais aussi dans les travaux d'indexation sémantique [Gonzalo et al., 98 ; Mihalcea et al., 00 ; Stairmand et al., 96 ;Baziz et al., 04 ;05], en particulier pour la désambiguïsation sémantique des mots [Nastase et al., 01 ; Banerjee et al., 02 ;Voorhees, 93 ; Baziz et al., 04 ;05]. Ce sont ces mêmes raisons qui ont motivés notre choix pour cette ressource linguistique.

4.3.2 Les règles d'association

Le concept de règles d'association fut introduit à l'origine dans [Agrawal et al., 93] pour l'analyse des bases de données transactionnelles composées des transactions de ventes dans les grands magasins, afin de comprendre les habitudes de consommation de la clientèle et ainsi mieux gérer les ventes, les stocks, les rayons du magasin, dans l'objectif d'une meilleure planification commerciale. L'extraction des règles d'association permet de retrouver des relations entre les articles qui « vont souvent ensemble ». Ainsi, si dans une base de données transactionnelle de ventes D , où chaque transaction représente l'ensemble des articles achetés par un client, un ensemble d'articles X est souvent accompagné d'un ensemble d'articles Y dans les transactions de la base, on en déduit la règle d'association $X \rightarrow Y$.

Par exemple, les produits $\{pain, confiture\}$ sont présents dans les transactions qui contiennent $\{beurre\}$, impliquant une règle d'association $\{beurre\} \rightarrow \{pain, confiture\}$ qui stipule que «les clients qui achètent du beurre achètent aussi du pain et de la confiture ».

4.3.2.1 Découverte des règles d'association

Le problème de découverte des règles d'association consiste à extraire un ensemble de règles d'association « intéressantes » entre ensembles d'articles définis dans une base de données transactionnelle. Nous abordons ce problème dans la présente section à travers les points suivants :

1. la définition du contexte de découverte des règles d'association,
2. la définition de l'intérêt d'une règle d'association,
3. les algorithmes de découverte des règles d'association.

4.3.2.1.1 Contexte de découverte des règles d'association

Formellement, les règles d'association sont découvertes dans les bases de données transactionnelles, entre ensembles d'items. Le contexte ainsi défini est caractérisé par les propriétés suivantes [Agrawal et al., 93] :

- $I = \{i_1, i_2, \dots, i_n\}$ un ensemble d'items (ou articles),
- $D = \{T_1, T_2, \dots, T_m\}$ un ensemble de transactions T_j telles que $T_j \subseteq I$,
- à chaque transaction est associé un identificateur appelé TID (*Transaction Identification*),
- la quantité d'un item dans une transaction n'est pas considérée. Chaque item est une variable binaire représentant le fait que l'item est concerné par la transaction ou non,
- un ensemble d'items est appelé itemset,
- la taille d'un itemset est le nombre d'items qu'il contient,
- un itemset de taille k est appelé k -itemset,
- étant donné un itemset X , une transaction T ($T \subseteq I$) contient X si $X \subseteq T$,
- on appelle support d'un itemset X , le pourcentage des transactions de D qui contiennent X .

$$\text{support}(X) = \frac{|\{T_i \in D / T_i \supseteq X\}|}{|D|} \quad (4.1)$$

Définition d'une règle d'association : Une règle d'association est une implication de la forme $X \rightarrow Y$ telle que $X \subset I, Y \subset I$ et $X \cap Y = \emptyset$.

X est dit prémisses de la règle $X \rightarrow Y$, Y sa conclusion (ou son conséquent).

4.3.2.1.2 Définition de l'intérêt d'une règle d'association

L'intérêt d'une règle d'association se mesure à travers deux valeurs : son support et sa confiance.

Le support de la règle d'association $R : X \rightarrow Y$ définit le pourcentage de transactions qui contiennent X et Y . Le support indique la fréquence des itemsets de la règle [Chen et al., 96]. Le support est formellement défini par :

$$\text{support}(R) = \frac{|\{T_i \in D / T_i \supseteq (X \cup Y)\}|}{|D|} \quad (4.2)$$

La confiance d'une règle d'association $R : X \rightarrow Y$ définit le pourcentage de transactions contenant X qui contiennent aussi Y . Elle dénote la force de l'implication [Chen et al., 96]. La confiance est formellement définie par :

$$\text{confiance}(R) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \quad (4.3)$$

Etant donné un ensemble de transactions D , le problème de découverte de règles d'association est de générer toutes les règles d'association qui ont un support et une confiance supérieurs à un support minimum $minsup$ et une confiance minimum $minconf$ respectivement fixés par l'utilisateur. De telles règles sont des règles dites fortes [Agrawal et al., 93 ; Piatetsky-Shapiro, 91].

4.3.2.1.3 Algorithmes de découverte des règles d'association

Les algorithmes de découverte des règles d'association se basent sur deux étapes [Agrawal et al., 94] :

- générer toutes les combinaisons d'items (ie. tous les itemsets) dont le support est supérieur à $minsup$. De tels itemsets sont dits fréquents.
- Pour chaque itemset fréquent $Y = i_1 i_2 \dots i_k$, générer la règle d'association $X \rightarrow Y - X$, pour tout $X \subset Y$.

La performance globale d'un algorithme de découverte de règles d'association est déterminée par la première étape. Après avoir déterminé les itemsets fréquents, les règles d'association correspondantes sont extraites de manière triviale. Plusieurs algorithmes sont proposés dans la littérature dont l'un des premiers et qui reste le plus utilisé est l'algorithme Apriori. Nous le détaillons ci-après.

L'algorithme A-priori

L'algorithme Apriori comprend deux étapes. La première étape permet d'extraire l'ensemble des itemsets fréquents de la base de données transactionnelles. La seconde est l'étape de génération des règles d'association entre les itemsets fréquents découverts lors de la première étape. Nous détaillons le fonctionnement de chacune de ces étapes dans ce qui suit.

(1) Génération des itemsets fréquents

Pour générer les itemsets fréquents dans une base de données transactionnelle D , Apriori réalise plusieurs passes sur D .

1. Lors de la première passe, l'algorithme calcule le nombre d'occurrences des différents items de la base, afin de déterminer l'ensemble F_1 des 1-itemsets fréquents.
2. Chaque nouvelle itération k , consiste en deux phases :
3. d'abord, l'ensemble F_{k-1} des $(k-1)$ -itemsets fréquents calculés à l'étape précédente, est utilisé pour générer, par auto-jointure, l'ensemble C_k des k -itemsets candidats,

4. ensuite, le support de chaque k -itemset candidat est calculé. Les seuls candidats fréquents, c'est-à-dire de support supérieur ou égal au seuil minimal de support $minsup$ sont insérés dans l'ensemble F_k . L'algorithme s'arrête lorsqu'il n'y plus de nouveaux itemsets candidats à générer.
5. L'ensemble $F = \cup F_k$ de tous les itemsets fréquents est alors retourné.

Le tableau 4.3 présente le pseudo-code de l'algorithme Apriori.

Algorithme Apriori : Génération des itemsets fréquents
Entrée : Base de transactions D ; Support minimum $minsup$
Sortie : Ensemble des itemsets fréquents F
Algorithme
1. $F_1 \leftarrow \{1\text{-itemsets fréquents}\}$;
2. pour ($k \leftarrow 2$; $F_{k-1} \neq \emptyset$; $k++$) faire
3. Construire (C_k) à partir de (F_{k-1});
4. pour toute transaction $T \in D$ faire
5. $C_T \leftarrow \{c \in C_k, c \subseteq T\}$
6. pour chaque candidat $c \in C_T$ faire support (c) ++ fin pour ;
7. fin pour
8. $F_k \leftarrow \{c \in C_k / \text{support}(c) \geq minsup\}$;
9. fin pour
10. Retourner $F = \cup F_k$

TABLEAU 4.3 : pseudo-code de l'algorithme Apriori

(2) Génération des règles d'association

Pour générer les règles d'association, on considère l'ensemble F des itemsets fréquents calculés lors de la phase précédente. Pour chaque itemset fréquent X de taille supérieure à 1, on considère tous les sous ensembles non vides de X . Pour chaque sous ensemble Y de X , déduire la règle d'association $Y \rightarrow X - Y$ si sa confiance est supérieure ou égale à $minconf$.

Nous illustrons le fonctionnement de l'algorithme Apriori dans l'exemple suivant.

Exemple

Etant donnée la base de transactions D donnée en Tableau 4.4, et un support minimum $minsup=0.4$.

D	
TID	Transaction
T1	acd
T2	bce
T3	abce
T4	be

TABLEAU 4.4 : Base transactionnelle D , avec 4 des transactions T_i

(1) extraction des itemsets fréquents :

Les étapes de l'extraction des itemsets fréquents sur la base de l'algorithme Apriori sont schématisées à travers les tableaux de la figure 4.2.

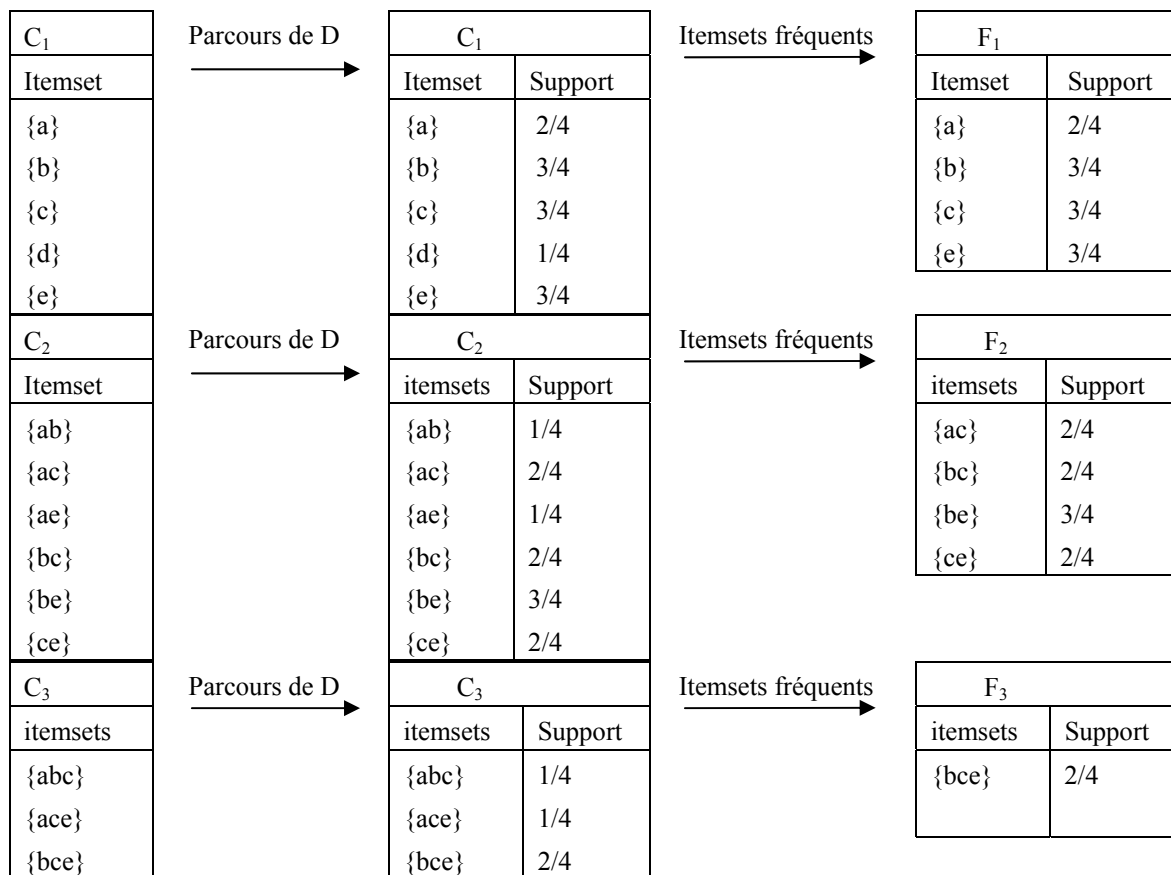


FIGURE 4.2 : Extraction des itemsets fréquents

(2) Génération des règles d'association :

En considérant l'ensemble des k-itemsets fréquents ($k > 1$) générés dans l'exemple précédent, soit $F = F_2 \cup F_3$ où $F_2 = \{\{ac\}, \{bc\}, \{be\}, \{ce\}\}$ et $F_3 = \{\{bce\}\}$, on extrait les règles d'association correspondantes en considérant d'abord les itemsets fréquents de taille 2, puis ceux de taille 3, etc. Les itemsets fréquents de F_2 ont permis de générer les règles d'association du tableau 4.5. Les itemsets fréquents de F_3 , à savoir l'unique itemset $\{bce\}$, ont permis de générer les règles d'association, d'abord avec un conséquent à un item figurant dans le tableau 4.6. Puis les règles avec un conséquent de taille 2 en tableau 4.7. L'exemple montre les règles d'association générées pour une confiance minimum $minconf = 70\%$. Les règles dont le support et la confiance sont supérieurs ou égaux à $minsup$ et $minconf$ respectivement sont des règles d'association dites fortes.

itemset	N° règle	règle	confiance	support	forte ?
ac	1	$a \rightarrow c$	1	2/4	Oui
	2	$c \rightarrow a$	1	2/4	Oui
bc	3	$b \rightarrow c$	2/3	2/4	Non
	4	$c \rightarrow b$	2/3	2/4	Non
be	5	$b \rightarrow e$	1	3/4	Oui
	6	$e \rightarrow b$	1	3/4	Oui
ce	7	$c \rightarrow e$	2/3	2/4	Non
	8	$e \rightarrow c$	2/3	2/4	Non

TABLEAU 4.5 : Règles d'association à 1 item en conséquent.

Itemset	N° règle	règle	confiance	support	forte ?
Bce	11	$bc \rightarrow e$	1	2/4	Oui
	12	$be \rightarrow c$	2/3	2/4	Non
	13	$ce \rightarrow b$	1	2/4	Oui

TABLEAU 4.6 : Règles d'association à 1 item en conséquent.

itemset	N° règle	règle	confiance	support	forte ?
bce	14	$b \rightarrow ce$	2/3	2/4	Non
	15	$c \rightarrow be$	2/3	2/4	Non
	16	$e \rightarrow bc$	2/3	2/4	Non

TABLEAU 4.7 : Règles d'association à 2 items en conséquent

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

Dans l'algorithme Apriori, l'extraction des itemsets fréquents est exponentielle en taille des itemsets. En particulier, dans le cas d'un ensemble d'items I de taille m , le nombre d'itemsets fréquents potentiels est 2^m . En outre, la génération de règles d'association bien que moins coûteuse que la génération des itemsets fréquents, reste néanmoins exponentielle en taille des itemsets fréquents (le nombre de règles pouvant être générées à partir d'un k -itemset, $k > 1$, égale à $2^k - 1$ [Salleb, 03]). Sachant que les bases de données transactionnelles stockent généralement des millions de transactions sur des milliers d'items, on comprend que l'algorithme Apriori ne soit pas optimal. Cependant, il demeure l'un des algorithmes les plus simples à mettre en œuvre.

4.3.2.2 Les règles d'association en RI

L'utilisation des règles d'association en RI vise principalement la découverte de relations non taxonomiques entre les termes (mots clés ou concepts) descripteurs des documents d'une base documentaire. Les relations non taxonomiques sont des relations contextuelles entre termes. Elles sont spécifiques à l'usage particulier des termes dans les documents de la collection considérée. Il s'agit plus particulièrement de relations latentes, enfouies dans les textes, portées par la sémantique même de la cooccurrence des termes dans le document ou dans la base documentaire. Les objectifs à travers la découverte des règles d'association en RI sont multiples et variés comme en témoigne la multitude d'applications existantes :

1. La classification de textes en vue de la réduction de l'espace de recherche [Lin et al., 98],
2. La génération automatique d'associations de termes pour l'aide à l'expansion de requête [Liu et al., 98; Haddad, 02; Delgado et al., 02; Bautista et al., 04; Song et al., 07],
3. L'indexation [Pôssas et al., 05; Kim et al., 04],
4. Le regroupement (clustering) de textes fournit des vues d'ensemble thématiques des collections des textes [Lin et al., 98; Liu et al., 05],
5. ...etc.

Dans le contexte de la RI, une transaction dans la problématique des règles d'association est une entité textuelle, généralement un document, et les items les termes d'indexation de ce document. Les principes d'extraction des règles d'association en RI restent identiques à ceux définis en section 4.3.2.1.

4.4 Approche d'indexation sémantique

L'indexation sémantique a pour objectif la représentation des documents et requêtes par les sens des mots (ou les concepts) plutôt que par les mots d'indexation eux même. L'intérêt d'une telle approche est de lever l'ambiguïté des mots et de résoudre le problème de disparité des termes. C'est dans l'objectif d'améliorer notre approche proposée en chapitre 3, que nous nous orientons vers l'indexation sémantique. Nous définissons, dans cette section, notre approche d'indexation sémantique de documents basée sur les CP-Nets. Documents et requêtes sont alors indexés par des concepts. Les concepts sont extraits de WordNet, puis désambiguïsés. Des relations contextuelles sont ensuite découvertes entre concepts. Et enfin, concepts et relations associées sont organisées en un graphe CP-Net constituant l'index sémantique (ou conceptuel) du document. Notre approche s'articule autour des trois caractéristiques suivantes :

1. *une approche d'identification des concepts représentatifs du document.* L'approche est basée sur la projection des termes d'index sur l'ontologie WordNet et intègre une technique de désambiguïsation des concepts ambigus,
2. *une approche d'identification des relations entre concepts.* Cette approche est basée sur l'utilisation des règles d'association,
3. *une approche qui combine les concepts et les relations* correspondantes au sein d'un formalisme unifié, le CP-Net.

4.4.1 Aperçu général

Nous utilisons l'ontologie WordNet et les règles d'association afin de construire le graphe CP-Net représentatif du document. Le processus d'indexation du document est effectué en trois étapes principales (Figure 4.3) : l'identification des concepts représentatifs du document la découverte des relations entre concepts, et la construction de l'index conceptuel du document.

1. *L'identification des concepts représentatifs du document* : les concepts sont extraits à partir des termes représentatifs du contenu sémantique du document, par projection sur l'ontologie WordNet. Lors de cette projection, si plusieurs concepts correspondent à un terme donné, le terme est désambiguïsé. Les sous étapes de cette première étape sont :
 - 1.1. L'identification des termes : le but de cette étape est d'identifier des mono ou multi termes dans le document. Ces termes correspondent à des entrées dans l'ontologie,
 - 1.2. La pondération des termes : dans cette étape, on propose une variante de $tf*idf$, s'appliquant aux mono et aux multi-termes. Le but est d'éliminer les termes les

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

moins fréquents dans le document et de maintenir seulement les termes les plus représentatifs,

1.3. La désambiguïsation : les termes d'index sont associés à des sens (synsets) correspondants dans l'ontologie. Chaque terme extrait pouvant avoir plusieurs sens possibles, le but de cette étape est de sélectionner le meilleur sens du terme dans le document.

11. *La découverte de relations entre concepts* : les relations contextuelles entre les concepts extraits sont découvertes en utilisant une approche que nous proposons, basée sur la technique des règles d'association,

12. *La construction de l'index conceptuel du document* : les concepts et les relations correspondantes sont organisés en un graphe conceptuel, à savoir le graphe. Les nœuds du CP-Net sont les concepts représentatifs du document. Les arcs du CP-Net traduisent les relations entre concepts.

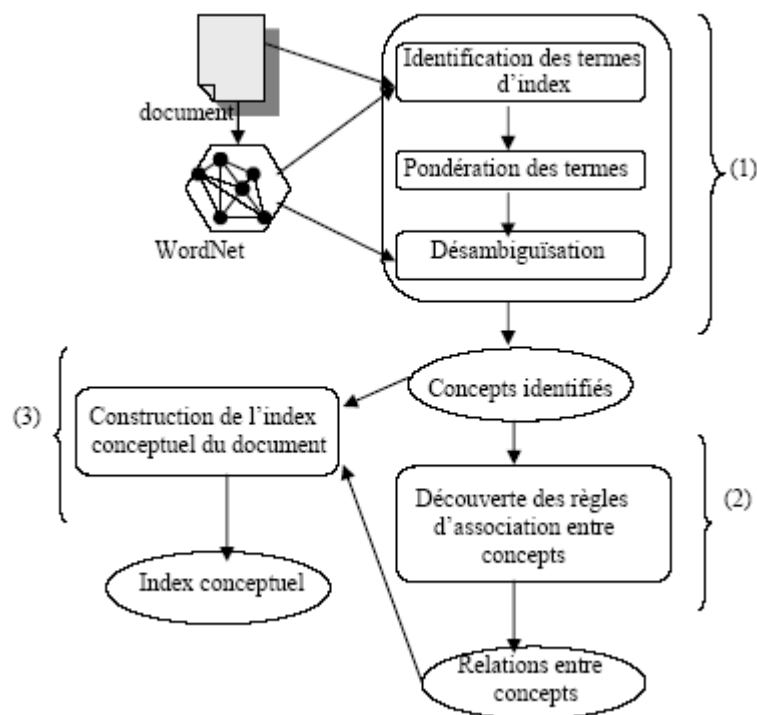


FIGURE 4.3 : Les étapes de l'indexation conceptuelle basée CP-Nets

4.4.2 Identification de concepts représentatifs du document

L'objectif de cette étape est de retrouver, pour chaque terme d'indexation d'un document (ou d'une requête), le concept correspondant dans WordNet. Une première sous étape de projection de l'index du document (ou de la requête) sur WordNet permet d'identifier les concepts correspondants aux termes d'indexation. Une seconde sous étape de désambiguïsation, permet de retrouver pour chacun de ces termes le seul sens correct correspondant dans le document (ou la requête). Rappelons que les sens des mots dans WordNet sont regroupés dans des synsets (correspondant à des concepts). Enfin, la sous étape de pondération permet d'associer aux différents concepts ainsi identifiés, leur degré de représentativité dans le document (ou la requête). Avant de décrire ces étapes, nous présentons d'abord quelques définitions préliminaires.

4.4.2.1 Notions préliminaires

a. Le terme vu comme une liste de mots

Le but du processus d'indexation est d'identifier et d'extraire les termes sensés représenter au mieux le contenu sémantique du document. Les termes sont généralement représentés comme des listes de mots. Un mot étant une entité lexicale représentée par une chaîne de caractères. La longueur d'un terme t , notée $|t|$, est alors définie comme le nombre de mots dans t . Un terme mono mot consiste en une liste à un seul mot. Un terme multi mots consiste en une liste à plusieurs mots.

Soit t un terme représenté comme une liste de mots, $t = (w_1, w_2, \dots, w_i, \dots, w_l)$. Les éléments dans t peuvent être identiques, représentant différentes occurrences d'un même mot de t . On note w_i le $i^{\text{ème}}$ mot dans t . On définit récursivement la position d'un mot w_i dans la liste t comme suit :

$$pos_t(w_1) = 1; \quad pos_t(w_i) = pos_t(w_{i-1}) + 1, \forall i = 1..l$$

b) Sous-terme vs Sur-terme

Etant données deux listes de mots $L_1 = (w_1, w_2, \dots, w_m)$ et $L_2 = (y_1, y_2, \dots, y_l)$.

Définition d'une sous liste : L_2 est une sous liste de L_1 si l'ensemble de la séquence de mots dans L_1 apparaît aussi dans L_2 . Formellement :

$$L_2 = sub(L_1, p, l) \text{ si } \exists p = pos(w_i), w_i \in L_1 / \forall j \ 0 \leq j \leq (l-1), w_{p+j} = y_{j+1} \quad (4.4)$$

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

Etant donnés deux termes t_1 et t_2 , représentés respectivement par les listes de mots associées respectivement L_1 et L_2 , on définit les notions de sous terme et de sur terme comme suit :

$$\begin{aligned} t_2 \text{ est un sous-terme de } t_1 &\Leftrightarrow L_2 \text{ est une sous-liste de } L_1 . \\ t_1 \text{ est un sur-terme de } t_2 &\Leftrightarrow t_2 \text{ est un sous-terme de } t_1 . \end{aligned}$$

4.4.2.2 Identification des termes d'index

Avant tout traitement de document et en particulier avant l'élimination des mots vides, un processus important pour les étapes suivantes consiste en l'extraction des termes mono et multi mots (on parle aussi de multi termes) qui correspondent à des entrées de WordNet. La technique que nous proposons est basée sur une analyse mot par mot du document. Elle est formellement décrite dans ce qui suit.

Soit w_i le prochain mot (supposé non vide), à analyser dans le document d . On extrait à partir de WordNet, l'ensemble S des termes contenant w_i . S peut être vide dans le cas où aucun terme de WordNet ne correspond à w_i . Dans ce cas, le mot w_i est sélectionné comme terme d'indexation. Le prochain mot (supposé non vide), à analyser dans le document d est le mot w_{i+1} .

Dans le cas où S est non vide, les termes retrouvés appartiennent à des synsets de WordNet. Soit donc $S = \{C_1^i, C_2^i, \dots, C_n^i\}$. S est composé de mono et de multi termes. On ordonne alors S par ordre décroissant de tailles de ses éléments comme suit :

$$S = \{C_{(1)}^i, C_{(2)}^i, \dots, C_{(n)}^i\}$$

où $(j)=1..n$ est une permutation d'indices telle que $|C_{(1)}^i| \geq |C_{(2)}^i| \geq \dots \geq |C_{(n)}^i|$. Les termes de taille identique sont placés indifféremment l'un à côté de l'autre. Pour chaque élément $|C_{(j)}^i|$ dans S , on note :

- $Pos_{C_{(j)}^i}(w_i)$ la position de w_i dans la liste de mots $C_{(j)}^i$. Il y a ainsi $(pos_{C_{(j)}^i}(w_i) - 1)$ mots à gauche de w_i dans $C_{(j)}^i$, ($pos_{C_{(j)}^i}(w_i) > 0$),
- $Pos_d(w_i)$ la position de w_i dans la phrase analysée du document d , vue aussi comme liste de mots.

Définition du contexte relatif : Le contexte relatif d'une occurrence de w_i dans un document d tant donné le terme $C_{(j)}^i$, est la liste de mots CH_j^i définie par :

$$CH_j^i = sub(d, p, l) \text{ tq. } l = |C_{(j)}^i| \text{ et } p = pos_d(w_i) - (pos_{C_{(j)}^i}(w_i) - 1) \quad (4.5)$$

On extrait alors le contexte relatif de w_i dans d , soit $CH_j^i = sub(d, p, l)$ (c.f. Fig. 4.4), puis on compare les listes de mots CH_j^i et $C_{(j)}^i$. Si $CH_j^i \neq C_{(j)}^i$, le terme $C_{(j+1)}^i$ de S est analysé, sinon le terme $t_k = C_{(j)}^i$ est identifié.

Le mot suivant à analyser dans d est le mot w_j tel que $pos_d(w_j) = p + l$.

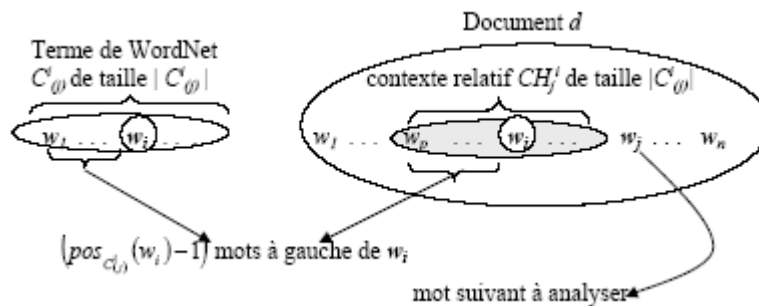


FIGURE 4.4 : Identification du contexte relatif d'un mot dans d

Durant le processus d'identification, trois cas, illustrés dans la figure 4.5, peuvent se présenter:

- Cas a) Le terme courant identifié t_k est complètement disjoint de t_{k-1} . Il peut être identique mais nous ne traitons pas les identités à ce niveau. Ce sera donc un nouveau terme qui sera retenu dans la description du document.
- Cas b) Le terme t_k recouvre partiellement le terme t_{k-1} . Les deux termes sont donc différents et tous deux seront retenus comme descripteurs du document même s'ils ont des mots en commun.
- Cas c) Le terme t_k recouvre complètement un ou plusieurs termes adjacents le précédant, soit $t_{k-1} \dots t_j$, $j \leq k-1$. Dans ce cas, pour permettre une désambiguïsation efficace, nous retenons le terme le plus long soit t_k , comme descripteur du document et éliminons de l'index, les termes adjacents qu'il recouvre ($t_{k-1} \dots t_j$, $j \leq k-1$).

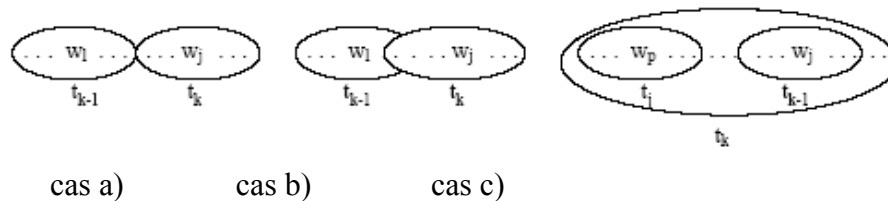


FIGURE 4.5 : Identification des termes

A l'issue de cette première étape, nous aurons identifié l'ensemble $T(d)$ des mono et multi termes qui décrivent le document d , soit donc $T(d) = \{t_1, t_2, \dots, t_n\}$. Nous calculons finalement la fréquence de chaque terme dans d et éliminons les termes redondants (doublons), ce qui conduit à l'index suivant :

$T'(d) = \{(t_1, Occ_1), (t_2, Occ_2), \dots, (t_m, Occ_m) \mid t_i \in d, Occ_i = count(t_i), 1 \leq i \leq m \text{ et } m \leq n\}$
 tel que $count(t_i)$ est le nombre d'occurrences de t_i dans d .

4.4.2.3 Pondération des termes

La pondération des termes assigne à chaque terme un poids qui reflète son importance dans le document. Dans le cas des mono termes, des variantes de la formule $tf * idf$ sont utilisées. Le poids d'un terme t dans un document d est alors exprimé par $W_{t,d} = tf * idf(t)$ tel que tf est la fréquence d'occurrences du terme, idf est sa fréquence documentaire inverse, définie comme suit :

$$idf(t) = \ln\left(\frac{N}{n_t}\right) \quad (4.6)$$

N étant le nombre de documents dans le corpus et n_t le nombre de documents du corpus qui contiennent le terme t .

Dans le cas des multi termes, les approches de pondération des termes utilisent généralement une analyse statistique et/ou syntaxique. Globalement, ils additionnent les fréquences des mots simples ou multiplient le nombre d'occurrences du terme par le nombre de mots simples appartenant à ce terme. Dans le cas de [Baziz et al., 05], la fréquence d'un multi terme t (qui correspond à un concept de WordNet) dans un document égale la somme du nombre des occurrences du terme lui même et du nombre d'occurrences de tous ses sous termes. Formellement :

$$tf(t) = count(t) + \sum_{t_i \in sub(t)} \frac{|t_i|}{|t|} count(t_i) \quad (4.7)$$

Où $sub(t_i)$ est l'ensemble de tous les sous termes possibles t_i qui peuvent être dérivés de t , $|t_i|$ représente le nombre de mots dans t_i et $count(t)$ le nombre d'occurrences de t dans d .

Dans notre proposition, nous définissons une nouvelle approche de pondération qui combine une analyse statistique et une analyse sémantique. Pour cela, nous avons défini une formule de pondération comme variante de $tf*idf$ qui prend en compte les caractéristiques suivantes :

1. une mesure statistique des occurrences du terme lui même,
2. une mesure statistique des occurrences du terme dans ses sur termes,
3. une mesure probabiliste des occurrences du terme dans les différents sens associés à ses sous termes (rappelons que le sens d'un terme dans WordNet est un synset qui définit un ensemble de termes synonymes).

L'idée est que la fréquence globale d'un terme est quantifiée sur la base de sa propre fréquence d'occurrences d'une part, et de sa fréquence d'occurrences dans chacun de ses sur termes ainsi que sa fréquence d'occurrence probable dans les sens de chacun de ses sous termes. La fréquence d'occurrences d'un terme est calculée de manière triviale comme le nombre d'occurrences de ce terme dans le document, soit $Occ(t)$. La

fréquence d'occurrence d'un terme dans ses sur termes est aussi trivialement calculée comme le total des occurrences de tous ses sur termes, soit donc:

$$f = \sum Occ(Sur_i(t)).$$

La fréquence probable d'un terme dans les sens de son sous terme se mesure par le produit du nombre d'occurrences du sous terme considéré par la probabilité que le terme soit un sens possible de son sous terme. Généralisée à l'ensemble des sous termes, la fréquence probable se mesure comme la somme des fréquences probables du terme dans les sens de tous ses sous termes. Formellement, la fréquence probable est définie par:

$$f_p = \sum_j [P(t \in S(Sub_j(t))) * Occ(Sub_j(t))]$$

La probabilité $P(t \in S(Sub_j(t)))$ que le terme t soit un sens possible de son sous terme $Sub_j(t)$ se mesure comme le rapport entre le nombre de sens (synsets) du sous terme qui contiennent le terme t , sur le nombre total de sens du sous terme considéré. Formellement:

$$P(t \in S(Sub_j(t))) = \frac{|\{C \in S(Sub_j(t)) / t \in C\}|}{|S(Sub_j(t))|} \quad (4.8)$$

La formule résultante $W_{t,d} = tf * idf(t)$ est formellement définie par :

$$W_{t,d} = \left(Occ(t) + \sum_i Occ(Sur_i(t)) + \sum_j [P(t \in S(Sub_j(t))) * Occ(Sub_j(t))] \right) * \ln \left(\frac{N}{n_t} \right) \quad (4.9)$$

Où :

- $W_{t,d}$: poids associé au terme t dans le document,
- $Sub_j(t) \in T^n(d)$: $j^{\text{ème}}$ sous-terme de t ,
- $Sur_i(t) \in T^n(d)$: $i^{\text{ème}}$ sur-terme de t ,
- N : nombre total de documents dans le corpus,
- n_t : fréquence documentaire de t ,
- $S(t)$: ensemble des synsets (sens) associés au terme t ,
- $P(t \in S(Sub_j(t)))$ définit la probabilité que t soit un sens possible de $Sub_j(t)$.

L'index du document, $Index(d)$, est alors construit en ne gardant que les seuls termes dont le poids est supérieur à un seuil minimal fixé.

4.4.2.4 Désambiguïisation des termes

L'objectif de la désambiguïisation est de retrouver le sens correct d'un terme dans son contexte d'énonciation. Nous définissons dans le présent paragraphe, notre approche de désambiguïisation basée sur WordNet. L'approche consiste à retrouver pour chaque terme dans $Index(d)$, tous les sens qui lui sont associés dans WordNet, puis à le désambiguïiser si nécessaire.

Ainsi, chaque terme $t_i \in Index(d)$ peut avoir un certain nombre de sens correspondant à des synsets de WordNet. Soit $S_i = \{C_1^i, C_2^i, \dots, C_n^i\}$ l'ensemble de tous les synsets associés au terme t_i . Ainsi, t_i possède $|S_i| = n$ sens. Nous posons l'hypothèse que chaque terme d'index contribue à la définition de la sémantique du document d avec seulement un seul sens (même si cela est quelque peu erroné, puisqu'un terme peut avoir différents sens dans un même document, mais nous considérons ici le seul sens dominant). De là, nous devons choisir pour chaque terme $t_i \in Index(d)$, son meilleur sens dans d . C'est le principe même de la désambiguïisation.

Parmi les différentes approches de désambiguïisation proposées dans la littérature, nous nous sommes particulièrement intéressés à l'approche proposée dans [Baziz et al.,04; 05] pour sa simplicité. Cette approche est basée sur le calcul d'un score (C_Score) pour chaque concept (sens) associé à un terme d'index. Ainsi, pour chaque terme t_i , le score de son j ème sens, noté C_j^i , est calculé par :

$$C_Score(C_j^i) = \sum_{\substack{l \in [1..m] \\ l \neq i}} \sum_{k \in [1..n_l]} Dist(C_j^i, C_k^l)$$

Où m est le nombre de termes dans $Index(d)$, n_l représente le nombre de sens de WordNet propres à chaque terme t_l et $Dist(C_j^i, C_k^l)$ est une mesure de proximité sémantique entre les concepts C_j^i et C_k^l telle que définie dans [Resnik, 99; Leacock et al., 98; Lin, 98]. Le concept-sens qui maximise le score est alors retenu comme le meilleur sens du terme t_i .

Notre approche diffère de celle de Baziz et al. [04; 05] dans la formule de calcul du score. En effet, nous pensons que l'exploitation de la seule proximité sémantique entre concepts est insuffisante pour déterminer le meilleur sens d'un terme car cette mesure est indépendante du contexte (elle ne tient pas compte de la représentativité des termes dans le contexte du document). Nous pensons que le meilleur sens pour un terme t_i dans le document d doit être fortement corrélé aux sens associés aux autres termes importants du document d . Dans cette optique, nous avons d'abord défini le poids d'un concept (sens) $C_j^i \in S_i$ comme le poids du terme correspondant t_i :

$$\forall C_j^i \in S_i, W_{C_j^i, d} = W_{t_i, d} \quad (4.10)$$

Puis nous proposons le score suivant :

$$Score(C_j^i) = \sum_{\substack{l \in [1..m] \\ l \neq i}} \sum_{1 \leq k \leq n_l} (W_{C_j^i, d} * W_{C_k^l, d} * Dist(C_j^i, C_k^l)) \quad (4.11)$$

L'ensemble $N(d)$ des sens (concepts) sélectionnés représente le noyau sémantique du document d .

4.4.3 Découverte des relations entre concepts

Dans l'objectif de construire le graphe conceptuel d'un document conformément à la sémantique CP-Net, nous devons retrouver les relations conditionnelles (i.e. de causalité) existantes entre les concepts du noyau sémantique. Ces relations sont implicites et se traduisent par des liens de co-occurrence entre les termes. De telles relations conditionnelles implicites sont naturellement prises en charge par la technique des règles d'association.

Nous proposons donc d'utiliser la technique des règles d'association pour découvrir les relations contextuelles latentes entre les concepts-nœuds du CP-Net. Un concept-nœud du CP-Net est un concept issu de l'index sémantique $N(d)$ restructuré en $(X, Dom(X))$, tel que X est le représentant du concept et $Dom(X)$, l'ensemble de ses valeurs possibles (correspondant à un ensemble de synsets Y_i de $N(d)$ tels que X subsume Y_i). Les concepts-nœuds sont des entités sémantiques. Le formalisme existant des règles d'association permet de découvrir des relations entre entités lexicales, à savoir les termes, nous proposons alors de l'étendre pour supporter des associations entre entités sémantiques (les concepts-nœuds).

Soit donc $\eta(d) = \{(X, Dom(X))\}$ l'ensemble des concept-nœuds du document CP-Net, et soient $X, Y \in \eta(d)$.

Définition d'une règle d'association sémantique : Une règle d'association sémantique entre X et Y , on note $X \rightarrow_{sem} Y$, est définie comme suit :

$$X \rightarrow_{sem} Y \Leftrightarrow \exists X_i \in Dom(X), \exists Y_j \in Dom(Y) / X_i \rightarrow Y_j \quad (4.12)$$

tel que $X_i \rightarrow Y_j$ est une association entre les termes (1-itemsets) X_i et Y_j .

Le sens intuitif de la règle $X \rightarrow_{sem} Y$ est que, si un document est autour (*is about*) du concept X , il tend aussi à être autour du concept Y . L'*aboutness* du document exprime l'orientation du sujet (topic focus) de son contenu. Cette interprétation s'applique aussi à la règle $X_i \rightarrow Y_j$. Ainsi, la règle $R: X_i \rightarrow Y_j$ exprime la probabilité que la sémantique du contenu du document porte sur Y_j sachant qu'elle porte sur X_i . Relativement à cette sémantique, la confiance associée à la règle R est basée sur le degré d'importance de Y_j dans le document d , sachant le degré d'importance de X_i dans d . Elle est formellement définie dans ce qui suit.

Définition de la confiance d'une règle d'association classique: Soit la règle $R: X_i \rightarrow Y_j$, on définit :

$$Confiance(R) = \frac{Support(X_i \text{ and } Y_j)}{Support(X_i)} = \frac{\min(W_{X_i,d}, W_{Y_j,d})}{W_{X_i,d}} \quad (4.13)$$

Définition de la confiance d'une règle d'association sémantique : La confiance de la règle d'association sémantique $R_{sem}: X \rightarrow_{sem} Y$ est définie par :

$$Confiance(X \rightarrow_{sem} Y) = \max_{i,j} \left(Confiance \left(R: X_i \rightarrow Y_j \right) \right)_{X_i \in Dom(X), Y_j \in Dom(Y)} \quad (4.14)$$

Remarque. $Confiance(X \rightarrow_{sem} Y)$ est toujours égale à 1.

Dans notre contexte, le support d'une règle d'association sémantique $X \rightarrow_{sem} Y$ est basé sur le nombre de règles d'associations individuelles $X_i \rightarrow Y_j$ ($X_i \in Dom(X)$ et $Y_j \in Dom(Y)$), ayant une confiance supérieure ou égale à un seuil de confiance minimal fixé $minconf=1$. Le support est formellement défini dans ce qui suit.

Définition du support d'une règle d'association sémantique : Etant donnée la règle $R: X \rightarrow_{sem} Y$, on définit :

$$Support(R) = \frac{|\{X_i \rightarrow Y_j / Confiance(X_i \rightarrow Y_j) \geq minconf\}|}{|\{X_i \rightarrow Y_j, (X_i, Y_j) \in Dom(X) \times Dom(Y)\}|} \quad (4.15)$$

Nous proposons de découvrir les relations entre les concepts de $\eta(d)$ au moyen des règles d'association sémantiques. Les règles d'association sémantiques sont basées dans notre contexte, sur les principes suivants :

1. une transaction est un document,
2. les items sont les valeurs des concept-nœuds du CP-Net,
3. un itemset est un ensemble d'items appartenant à un même concept-nœud du CP-Net,
4. une règle d'association sémantique $X \rightarrow_{sem} Y$ définit dans le CP-Net, un arc orienté du concept-noeud X vers le concept-noeud Y . X est le noeud parent de Y dans le graphe.

En utilisant les règles d'association, nous visons la construction d'une structure conditionnelle hiérarchique du *topic focus* du contenu du document. Nous visons ainsi à structurer les concepts décrivant le document, en une hiérarchie conditionnelle naturellement supportée par la sémantique des règles d'association extraites.

Pour découvrir les règles d'association entre concepts est, nous appliquons l'algorithme Apriori [Agrawal et al., 94]. Ce qui consiste d'abord à identifier tous les 1-itemsets fréquents, correspondant aux concepts individuels. Un concept fréquent est, dans notre contexte, un concept dont le poids est supérieur ou égal à un seuil minimal fixé. En second lieu, les règles d'association sont découvertes entre les 1-itemsets fréquents (concepts). L'objectif de la découverte des règles d'association est de retenir uniquement les règles fortes, dont le support et la confiance sont supérieurs ou égaux à un seuil minimal de support *minsup* et de confiance *minconf* respectivement.

Quelques problèmes peuvent cependant survenir lors de la découverte des règles tels que les redondances et les cycles. Les règles redondantes découlent généralement de la propriété de transitivité: $X \rightarrow_{sem} Y$, $Y \rightarrow_{sem} Z$ et $X \rightarrow_{sem} Z$. Pour éliminer les redondances, nous proposons de construire la couverture minimale de l'ensemble des règles extraites (c'est-à-dire le sous-ensemble minimal de règles non transitives).

Par ailleurs, l'existence de cycles est généralement due à la découverte simultanée de règles d'association $X \rightarrow_{sem} Y$ et $Y \rightarrow_{sem} X$, ou de règles d'association telles que $X \rightarrow_{sem} Y$, $Y \rightarrow_{sem} Z$ et $Z \rightarrow_{sem} X$. Pour résoudre ce problème, nous éliminons la règle la plus faible (i.e. ayant le support le plus faible) parmi les règles ayant conduit au cycle. Si toutes les règles ont un même support, nous éliminons aléatoirement une règle du cycle.

4.4.4 Construction de l'index conceptuel du document

Le but de cette étape est de construire l'index conceptuel CP-Net. Nous proposons d'utiliser le formalisme CP-Net comme langage d'indexation pour deux raisons. D'abord, les CP-Nets supportent naturellement les associations contextuelles conditionnelles. Ensuite, les CP-Nets permettent une représentation compacte tant des relations sémantiques que des relations contextuelles entre concepts, dans un formalisme unifié, à savoir le graphe CP-Net. Dans ce qui suit, nous décrivons le processus de construction des nœuds du CP-Nets et des relations entre eux.

(1) Les nœuds du CP-Net

Soit $N(d) = \{C_1, C_2, \dots, C_j, \dots\}$ le noyau sémantique du document d . Notre approche pour la construction des nœuds du CP-Net nodes est basée sur les principes suivants :

1. Les nœuds du CP-Net sont des variables attachées aux concepts C_i du noyau sémantique du document d . Dans ce qui suit, nous désignerons chaque variable par le concept correspondant,
2. chaque variable C_i prend ses valeurs dans l'ensemble $Dom(C_i) = \{C_1^i, C_2^i, C_3^i, \dots\}$,
3. chaque valeur dans $Dom(C_i)$, est un concept $C_j^i \in N(d)$ tel que C_j^i is - a C_i .

A l'issue de cette étape, nous aurons construit l'ensemble $\eta(d) = \{(C_i, Dom(C_i))\}$, on notera plus simplement $\eta(d) = \{(X, Dom(X))\}$, des concepts noeuds du CP-Net.

(2) Les relations du CP-Net

Les noeuds du CP-Net sont liés par des relations conditionnelles définies par les règles d'association correspondantes découvertes à l'issue de l'étape précédente.

Une fois le CP-Net construit, chaque noeud X dans le graphe est annoté par une table de valeurs, nommée $CPT(X)$ (par analogie aux tables CPT dans un CP-Net) telle que :

$$\forall X_i \in Dom(X), CPT(X_i) = W_{X_i, d} \quad (4.16)$$

4.4.5 Illustration

L'approche d'indexation présentée ci-dessus est illustrée à travers l'exemple suivant.

Soit d ((*Paris*, 0.5), (*Toulouse*, 0.9), (*Center*, 0.1), (*Studio*, 0.4), (*Suburbs*, 0.7), ...) un document décrit par un ensemble donné de concepts pondérés qui constituent alors les noeuds du CP-Net document. En considérant la relation taxonomique *is-a* de WordNet et en supposant que nous ayons *Toulouse is-a City* et *Paris is-a City*, alors *Paris* et *Toulouse* appartiennent au domaine de valeurs du concept noeud *City*. De manière similaire, *Center* et *Suburbs* appartiennent au domaine de valeurs du concept noeud *Place*, tandis que *Studio* est associé au concept noeud *Housing*. Ainsi, nous avons:

$$\begin{aligned} \eta(d) &= \{(City, Dom(City)), (Place, Dom(Place)), (Housing, Dom(Housing))\} \\ Dom(City) &= \{Toulouse, Paris\}, Dom(Place) = \{Suburbs, Center\} \text{ et } Dom(Housing) = \\ &\quad \{Studio\}. \end{aligned}$$

Nous souhaitons découvrir les associations entre les noeuds *City*, *Housing* et *Place*. L'application de l'algorithme Apriori conduit à :

- (1) extraire tous les 1-itemsets fréquents,
- (2) générer les règles d'association entre les 1- itemsets fréquents.

Les relations qui nous intéressent étant définies entre concepts individuels plutôt qu'entre ensembles de concepts, nous calculons alors les seuls k -itemsets pour $k=1, 2$. Les 1-itemsets correspondent aux valeurs respectives des différents noeuds de $\eta(d)$. Soit donc *Toulouse*, *Paris*, *Center*, *Suburbs*, *Studio*. Le support d'un 1-itemset correspond à son poids dans le document considéré. Ainsi, on a: Support (*Toulouse*) = 0.9 ; Support(*Paris*) = 0.5 ; Support (*Center*) = 0.1 ... etc. En supposant un seuil minimal de support $minsup = 0.1$, on a $Support(Center) < minsup$: le 1-itemset *Center* n'est pas fréquent, il est alors éliminé. Les 2-itemsets candidats sont ensuite construits à partir des seuls 1-itemsets fréquents. Ce qui donne les ensembles d'items suivants : {*Toulouse*, *Studio*}; {*Toulouse*, *Suburbs*}; {*Paris*, *Studio*}; {*Paris*, *Suburbs*} et {*Studio*, *Suburbs*}.

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

Les supports des 2-itemsets sont alors calculés selon la formule (4.13), comme dans l'exemple suivant:

$$Support(\{Toulouse, Studio\}) = \min(Support(Toulouse), Support(Studio)) = \min(0.9; 0.4).$$

Les k -itemsets ($k = 1, 2$) fréquents extraits, et leurs supports associés sont donnés en Tableau 4.8. Les règles d'association extraites à partir des 2-itemsets fréquents sont données en Tableau 4.9.

1-Itemsets		Itemset	Support
		Toulouse	0.9
		Paris	0.5
		Center	0.1
		Suburbs	0.7
		Studio	0.4
Frequent itemsets	2-	Toulouse, Studio	0.4
		Toulouse, Suburbs	0.7
		Paris, Studio	0.4
		Paris, Suburbs	0.5
		Studio, Suburbs	0.4

TABLEAU 4.8 : Génération des k -itemsets fréquents

R_1 : Toulouse \rightarrow Studio	R_2 : Studio \rightarrow Toulouse
R_3 : Toulouse \rightarrow Suburbs	R_4 : Suburbs \rightarrow Toulouse
R_5 : Paris \rightarrow Studio	R_6 : Studio \rightarrow Paris
R_7 : Paris \rightarrow Suburbs	R_8 : Suburbs \rightarrow Paris
R_9 : Studio \rightarrow Suburbs	R_{10} : Suburbs \rightarrow Studio

TABLEAU 4.9 : Règles d'association générées

En appliquant la formule (4.13), on calcule les confiances des règles d'association ainsi obtenues comme dans l'exemple suivant:

$$\begin{aligned}
 Confiance(R) &= \frac{\min(Support(Toulouse), Support(Studio))}{Support(Toulouse)} \\
 &= \frac{\min(0.9; 0.4)}{0.9} = 0.44
 \end{aligned}$$

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

Les confiances calculées pour l'ensemble des règles d'association générées sont données en Tableau 4.10.

R_i	R_1	R_3	R_5	R_7	R_9
$Confiance(R_i)$	0.44	0.77	0.8	1	1
	R_2	R_4	R_6	R_8	R_{10}
	1	1	1	0.71	0.57

TABLEAU 4.10 : Confiances des règles

En supposant un seuil minimum de confiance $minconf = 1$, on retient les seules règles dont les confiances sont supérieures ou égales à $minconf$. Les règles sélectionnées sont présentées en Tableau 4.11.

- R_2 : Studio \rightarrow Toulouse
- R_4 : Suburbs \rightarrow Toulouse
- R_6 : Studio \rightarrow Paris
- R_7 : Paris \rightarrow Suburbs
- R_9 : Studio \rightarrow Suburbs

TABLEAU 4.11 : Règles d'association sélectionnées

Ces règles sont tout d'abord utilisées pour construire les règles d'association sémantiques, qui correspondent aux relations entre les concepts-nœuds du CP-Net. Ainsi, nous déduisons:

- de R_2 : Studio \rightarrow Toulouse et R_6 : Studio \rightarrow Paris, la règle : $Housing \rightarrow_{sem} City$
- de R_4 : Suburbs \rightarrow Toulouse, la règle : $Place \rightarrow_{sem} City$
- de R_7 : Paris \rightarrow Suburbs, la règle : $City \rightarrow_{sem} Place$
- de R_9 : Studio \rightarrow Suburbs, la règle : $Housing \rightarrow_{sem} Place$

Nous calculons ensuite le support de chacune des règles sémantiques obtenues. Le support d'une règle d'association sémantique est calculé selon la formule (4.15) comme dans l'exemple suivant:

$$Support(Housing \rightarrow_{sem} City) = \frac{|\{X_i \rightarrow Y_j / Confiance(X_i \rightarrow Y_j) \geq minconf\}|}{|\{X_i \rightarrow Y_j, (X_i, Y_j) \in Dom(Housing) \times Dom(City)\}|}$$

Les seules règles d'associations dérivées de $Housing \rightarrow_{sem} City$ sont R_2 et R_6 . D'où :

$$|\{X_i \rightarrow Y_j, (X_i, Y_j) \in Dom(Housing) \times Dom(City)\}| = 2$$

Par ailleurs, les confiances de ces deux règles sont égales à 1 (*minconf*), d'où :

$$|\{X_i \rightarrow Y_j / Confiance(X_i \rightarrow Y_j) \geq minconf; (X, Y) \in Dom(Housing) \times Dom(City)\}| = 2$$

Finalement, le support de la règle d'association sémantique $Housing \rightarrow_{sem} City$ égale 1, soit :

$$Support(Housing \rightarrow_{sem} City) = \frac{2}{2} = 1$$

Les résultats obtenus pour l'ensemble des règles d'association sémantiques sont présentés en Tableau 4.12.

$Housing \rightarrow_{sem} City$	1
$Place \rightarrow_{sem} City$	0.5
$City \rightarrow_{sem} Place$	0.5
$Housing \rightarrow_{sem} Place$	1

TABLEAU 4.12 : Supports des règles d'association sémantiques

Nous retenons évidemment les règles dont le support égale 1. Deux associations existent entre les concepts *City* et *Place*, avec un même support. Nous retenons aléatoirement l'une d'elles. Supposons que c'est la règle $Place \rightarrow_{sem} City$ qui est retenue.

L'ensemble des règles d'association sémantiques est donné en Tableau 4.12. Il est clair que, retenir les trois règles mènerait à un cycle dans le graphe CP-Net. Pour éviter ce problème, nous éliminons la règle la plus faible (i.e. la règle ayant le support le plus bas) $Place \rightarrow_{sem} City$. Finalement, les seules règles sémantiques sélectionnées sont :

$$Housing \rightarrow_{sem} City \text{ et } Housing \rightarrow_{sem} Place.$$

Enfin, les tables CPT sont associées aux concepts-noeuds *Housing*, *City* and *Place* respectivement, sur la base de la formule (4.16), ce qui conduit au CP-Net document donné en figure 4.6.

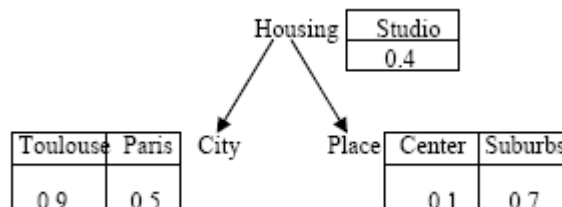


FIGURE 4.6 : Le CP-Net document

4.5 Evaluation des requêtes basée CP-Nets

Cette étape a pour objectif d'évaluer la pertinence d'un document pour une requête. Comparativement à notre proposition présentée dans le chapitre 3, plutôt que d'interpréter les CP-Nets document et requête en expressions booléennes pour évaluer leur degré de correspondance, nous proposons ici une approche d'évaluation des requêtes basée sur l'appariement des graphes CP-Net. En particulier, en combinant les résultats de nos deux contributions précédentes portant respectivement sur la définition de requêtes CP-Nets pondérées (chapitre 3) et de documents CP-Nets (chapitre 4, section 4.4), nous définissons pour un document CP-Net d et une requête CP-Net pondérée Q , un mécanisme d'évaluation de la pertinence de d pour Q basé sur la similarité des graphes CP-Nets correspondants.

4.5.1 Définition formelle

L'évaluation des requêtes consiste en la sélection de documents (D) supposés pertinents pour une requête utilisateur (Q). Pour cela, les documents sont ordonnés selon leurs valeurs de pertinence pour la requête ($RSV(D,Q)$) calculées dans notre approche en utilisant une mesure de similarité entre les graphes CP-Nets du document et de la requête respectivement. Formellement, cette valeur de similarité est exprimée par [Boubekeur et al., 07] :

$$RSV(D, Q) = SIM(G_D, G_Q) \quad (4.17)$$

Où G_D et G_Q sont les graphes CP-Nets correspondant respectivement au document D et à la requête Q . Cette similarité est calculée comme agrégation des similarités partielles des deux graphes à travers leurs concepts communs, comme suit :

$$SIM(G_D, G_Q) = \frac{|\eta(D) \cap \eta(Q)|}{|\eta(D) \cup \eta(Q)|} * \max_{X \in \eta(D) \cap \eta(Q)} (Sim^X(D, Q)) \quad (4.18)$$

Où :

$\eta(G_D)$ et $\eta(G_Q)$ sont les concepts-nœuds respectivement du CP-Net document G_D et du CP-Net requête G_Q .

$Sim^X(D, Q)$ est la similarité partielle entre D et Q au niveau du concept X . En se basant sur la topologie des graphes CP-Nets, cette mesure est calculée comme combinaison de la similarité structurelle et de la similarité relationnelle comme suit:

$$Sim^X(D, Q) = \alpha * Sim_{struct}^X(D, Q) + (1 - \alpha) * Sim_{relat}^X(D, Q) \quad (4.19)$$

Où $0 \leq \alpha \leq 1$ est une valeur donnée qui spécifie l'importance de la similarité structurelle par rapport à la similarité relationnelle.

La similarité structurelle Sim_{struct}^X définit la proportion de valeurs (instances) de X communes dans D et Q . Une instance commune dans D et Q est un terme de la requête Q qui appartient au document D .

La similarité relationnelle Sim_{relat}^X indique le degré de représentativité de X correspondant à son importance aussi bien dans le document que dans la requête. Sim_{relat}^X est mesurée en fonction de la profondeur associée au concept X dans la hiérarchie correspondant au graphe CP-Net. Les définitions formelles des similarités relationnelle et structurelle sont données dans ce qui suit.

Mesure de similitude structurelle

Soit $\eta(D) \cap \eta(Q)$ l'ensemble des concept-nœuds communs aux CP-Nets document et requête G_D et G_Q respectivement.

$\forall X \in \eta(D) \cap \eta(Q)$, considérons les domaines $Dom_{X,D}$ et $Dom_{X,Q}$ des instances (valeurs) associées au concept-noeud X respectivement dans G_D et G_Q . La similarité structurelle de D à Q au niveau du concept X est définie par :

$$Sim_{struct}^X(D, Q) = \frac{|Dom_{X,D} \cap Dom_{X,Q}|}{|Dom_{X,D} \cup Dom_{X,Q}|} \quad (4.20)$$

Mesure de similitude relationnelle

Pour un concept X , on définit $Deg_D(X)$, $Deg_Q(X)$ comme le niveau d'importance de X respectivement dans le document D et dans la requête Q . Le niveau d'importance du concept-noeud X est inversement proportionnel à la profondeur du noeud correspondant dans le graphe. Ainsi, pour un graphe de profondeur maximale n , la racine du graphe est de niveau 1 et d'importance 1. Ses descendants directs sont de niveau 2 et d'importance $1/2$...etc. Les éléments de niveau n ont une importance de $1/n$.

Soit $W_{X,D}$, $W_{X,Q}$ les poids associés aux valeurs X_j de X respectivement dans Q et D . $W_{X,D}$ est le poids *inconditionnel* de X_j dans D . Tandis que le poids de X_j dans Q est un poids conditionnel défini par $CPT(X_j / U_k)$ étant donnée une valeur U_k de ses parents dans le graphe CP-Net requête. Dans un premier temps, nous définissons simplement :

$$W_{X_j,Q} = Average_k (CPT(X_j / U_k)) \quad (4.21)$$

La similarité relationnelle de D à Q au niveau du concept X est d'abord définie au niveau de chaque valeur X_j de X comme le minimum entre son degré d'importance dans le document et dans la requête, normalisé par la somme de ses degrés d'importance dans le document et la requête CP-Net. Le degré d'importance de la valeur X_j de X est défini comme le produit de son poids de représentativité dans le

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

document ou dans la requête, et de son degré d'importance (défini par rapport à la position du noeud correspondant dans le graphe CP-Net) dans le CP-Net correspondant. Formellement:

$$Sim_{relat}^{X_j}(D, Q) = \frac{\min(W_{X_j, D} * Deg_D(X), W_{X_j, Q} * Deg_Q(X))}{(W_{X_j, D} * Deg_D(X) + W_{X_j, Q} * Deg_Q(X))} \quad (4.22)$$

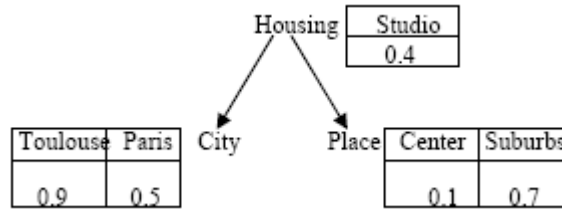
Pour toute valeur X_j de X dont les poids respectifs dans D et Q sont nuls, $Sim_{relat}^{X_j}(D, Q) = 0$.

La similarité relationnelle de D à Q au niveau du concept X est alors calculée comme la somme des similarités structurales de D et Q au niveau de toute instance X_j de X . Formellement :

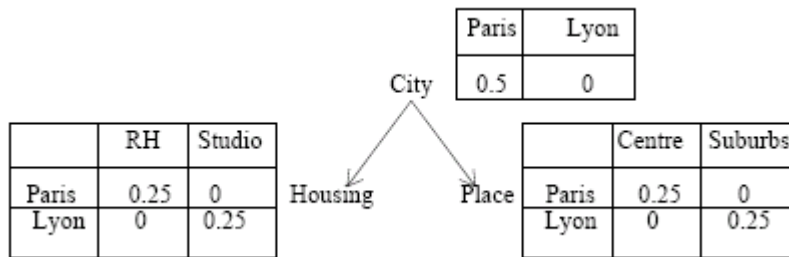
$$Sim_{relat}^X(D, Q) = \sum_j Sim_{relat}^{X_j}(D, Q) \quad (4.23)$$

4.5.2 Illustration

Considérons le CP-Net document présenté en Figure 4.6 et l'UCP-Net requête donné en Figure 3.6, repris respectivement ci-après:



CP-Net document



UCP-Net requête

On a:

$$\eta(G_D) = \eta(G_Q) = \{City, Housing, Place\},$$

$$Dom_{City, Q} = \{Paris, Lyon\},$$

$$Dom_{City, D} = \{Paris, Toulouse\},$$

$$Dom_{Housing, Q} = \{RH, Studio\},$$

$$Dom_{Housing, D} = \{Studio\},$$

$$Dom_{Place,Q} = Dom_{Place,D} = \{Center, Suburbs\}.$$

On calcule la pertinence du document D pour la requête Q suivant les étapes décrites en section 4.5.1.

(1) Calcul de la similarité structurale (selon la formule (4.20))

Au niveau du noeud $City$, on a:

$$|Dom_{City,D} \cap Dom_{City,Q}| = |\{Paris, Lyon\} \cap \{Paris, Toulouse\}| = 1$$

$$|Dom_{City,D} \cup Dom_{City,Q}| = |\{Paris, Lyon\} \cup \{Paris, Toulouse\}| = 3$$

$$Sim_{struct}^{City}(D, Q) = \frac{|Dom_{City,D} \cap Dom_{City,Q}|}{|Dom_{City,D} \cup Dom_{City,Q}|} = \frac{1}{3}$$

Au niveau du noeud $Housing$:

$$Sim_{struct}^{Housing}(D, Q) = \frac{|Dom_{Housing,D} \cap Dom_{Housing,Q}|}{|Dom_{Housing,D} \cup Dom_{Housing,Q}|} = \frac{1}{2}$$

Au niveau du noeud $Place$:

$$Sim_{struct}^{Place}(D, Q) = \frac{|Dom_{Place,D} \cap Dom_{Place,Q}|}{|Dom_{Place,D} \cup Dom_{Place,Q}|} = \frac{2}{2}$$

(2) Calcul de la similarité relationnelle:

Au niveau du noeud $City$, on a:

$$Sim_{relat}^{City}(D, Q) = \sum_j Sim_{relat}^{X_j}(D, Q), X_j \in Dom(City)$$

Cette similarité étant fonction du degré d'importance de chaque noeud tant dans le document que dans la requête CP-Net, nous calculons d'abord ces degrés d'importances associés aux noeuds comme suit:

(1) Dans le document:

$$Deg_D(Housing) = 1 \quad ; \quad Deg_D(Place) = Deg_D(City) = \frac{1}{2}$$

(2) Dans la requête:

$$Deg_Q(City) = 1 \quad ; \quad Deg_Q(Place) = Deg_Q(Housing) = \frac{1}{2}$$

On calcule alors la similarité relationnelle par rapport à chaque instance de $City$ selon la formule (4.22), ce qui donne:

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

$$\begin{aligned}
 Sim_{relat}^{Paris}(D, Q) &= \frac{\min(W_{Paris^j, D} * Deg_D(City), W_{Paris^j, Q} * Deg_Q(City))}{(W_{Paris^j, D} * Deg_D(City) + W_{Paris^j, Q} * Deg_Q(City))} \\
 &= \frac{\min(0.5 * 0.5; 0.5 * 1)}{0.5 * 0.5 + 0.5 * 1} \\
 &= \frac{0.25}{0.75} \\
 &= 0.33
 \end{aligned}$$

$$Sim_{relat}^{Lyon}(D, Q) = 0$$

$$\begin{aligned}
 Sim_{relat}^{Toulouse}(D, Q) &= \frac{\min(W_{Toulouse^j, D} * Deg_D(City), W_{Toulouse^j, Q} * Deg_Q(City))}{(W_{Toulouse^j, D} * Deg_D(City) + W_{Toulouse^j, Q} * Deg_Q(City))} \\
 &= \frac{\min(0.9 * 0.5; 0)}{0.9 * 0.5 + 0} = 0
 \end{aligned}$$

D'où, d'après la formule (4.23):

$$\begin{aligned}
 Sim_{relat}^{City}(D, Q) &= \sum_j Sim_{relat}^{X_j}(D, Q); X_j \in Dom(City) \\
 &= Sim_{relat}^{Paris}(D, Q) + Sim_{relat}^{Lyon}(D, Q) + Sim_{relat}^{Toulouse}(D, Q) \\
 &= 0.33 + 0 + 0 \\
 &= 0.33
 \end{aligned}$$

On considère dans l'exemple que les similarités structurelle et relationnelle ont la même importance dans l'évaluation de la pertinence du document pour la requête, ainsi nous fixons le paramètre α à la valeur 0.5, ce qui donne, selon la formule (4.19):

$$\begin{aligned}
 Sim^{City}(D, Q) &= 0.5 * Sim_{struct}^{City}(D, Q) + 0.5 * Sim_{relat}^{City}(D, Q) \\
 &= 0.5 * \frac{1}{3} + 0.5 * 0.33 \\
 &= 0.33
 \end{aligned}$$

Ces calculs sont reproduits pour chacun des concepts des CP-Nets document et requête. Les résultats correspondants sont donnés en tableau 4.13.

Ainsi, nous avons:

$$Sim^{Housing}(D, Q) = 0.317 \text{ et } Sim^{Plcae}(D, Q) = 0.795$$

D'où la similarité globale du document pour la requête calculée selon la formule (4.18) comme suit :

$$\begin{aligned}
 SIM(G_D, G_Q) &= \frac{|\eta(D) \cap \eta(Q)|}{|\eta(D) \cup \eta(Q)|} * \max_{X \in \eta(D) \cap \eta(Q)} (Sim^X(D, Q)) \\
 &= \frac{3}{3} * \max \left(Sim^{City}(D, Q); Sim^{Housing}(D, Q); Sim^{Place}(D, Q) \right) \\
 &= 1 * \max(0.33; 0.317; 0.795) \\
 &= 0.795
 \end{aligned}$$

X	City			Housing		Place	
$Deg_D(X)$	1/2			1		1/2	
$Deg_Q(X)$	1			1/2		1/2	
	Paris	Lyon	Toulouse	RH	Studio	Center	Suburbs
$W_{X,Q}$	0.5	0		0.125	0.125	0.125	0.125
$W_{X,D}$	0.5		0.9		0.4	0.1	0.7
$Sim_{reim}^X(D, Q)$	0.33			0.134		0.59	
$Sim_{struct}^X(D, Q)$	1/3			1/2		2/2	
$Sim^X(D, Q)$	0.33			0.317		0.795	
$SIM(G_D, G_Q)$	0.795						

TABLEAU 4.13 : Calcul de similarité entre les CP-Nets document et requête

4.6 Évaluation expérimentale

L'objectif de ces expérimentations est de mesurer les performances et la viabilité de notre approche de RI sémantique. L'évaluation complète de notre approche consisterait à (1) tester le modèle de RI flexible basé sur les CP-Nets (2) tester l'approche d'indexation sémantique proposée et enfin (3) tester l'approche d'évaluation des requêtes CP-Nets. Compte tenu de l'absence d'un cadre d'évaluation adéquat pour mener les évaluations (1) et (3), nous nous sommes focalisés dans le cadre de ce travail, sur l'évaluation de la seule approche d'indexation sémantique proposée. Nous présentons dans ce qui suit le cadre d'évaluation (collection de test et protocole d'évaluation) ainsi que les résultats expérimentaux préliminaires.

4.6.1 Collection Muchmore

Vu la complexité des calculs induits par les méthodes d'identification de concepts, de pondération et de désambiguïsation inhérentes à notre approche, nous avons opté pour

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

une collection de test de taille moyenne, la collection Muchmore¹⁶ en l'occurrence [Buitelaar et al., 04]. Le corpus MuchMore est un corpus parallèle de résumés médicaux scientifiques anglais-allemands obtenus à partir du site web de Springer. Le corpus se compose d'environ 1 million de termes pour chaque langue. Il comporte des résumés issus de 41 revues médicales de diverses spécialités. Il se décline en deux versions dont l'une annotée et l'autre sans annotations. Seule la collection des textes anglais non annotée a été utilisée. Cette dernière collection est composée de 7823 documents et de 25 requêtes, le tout formant 2.8 MB de données. Toutes les requêtes ont été utilisées. Un document est identifié comme suit :

/ Nom-revue suivi par identificateur-revue /

à l'exemple du document Arthroscopie/00130003.eng.abstr.

Des jugements de pertinence sont associés aux requêtes selon le format suivant :

Numéro de requête / 0 (non pertinent) / Nom-revue suivi par identificateur-revue / 1 (non pertinent)

Les documents et les requêtes sont composés de textes simples. En voici quelques exemples.

Exemple de document (Arthroscopie/00130003.eng.abstr)

« *The posterior cruciate ligament (PCL) is the strongest ligament of the human knee joint. Its origin is at the lateral wall of the medial femoral condyle and the insertion is located in the posterior part of the intercondylar area. The posterior cruciate ligament consists of multiple small fiber bundles. From a functional point of view, one can differentiate fiber bundles with an anterior origin and fiber bundles with a posterior origin at the femur. The anterior fibers insert in the anterolateral part of the tibial insertion zone. These fibers become tense when the knee is flexed. The posterior fibers insert in the posteromedial part of the tibial insertion and become tense when the knee is extended. The main part of the posterior cruciate ligament consists of type I collagen positive dense connective tissue. The longitudinal fibrils of type I collagen are divided into small bundles by thin type III collagen positive fibrils. In the center of the middle third region, the structure of the tissue varies from the typical structure of a ligament. In this region, the structure of the tissue resembles fibrocartilage. Oval-shaped cells surrounded by a metachromatic extracellular matrix lie between the longitudinal collagen fibrils. The femoral origin and the tibial insertion have the structure of chondral apophyseal enthesis. Near the anchoring region at the femur and tibia there should be various mechanoreceptors which might have an important function for the kinematics of the knee joint. The blood supply of the PCL arises*

¹⁶ <http://muchmore.dfki.de/>

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

from the middle geniculate artery. The ligament is covered by a synovial fold where the terminal branches of the middle and the inferior geniculate artery form a periligamentous network. From the synovial sheath, the blood vessels penetrate the ligament in a transverse direction and anastomose with a longitudinally orientated intraligamentous network. The distribution of blood vessels within the PCL is not homogeneous. We detected three avascular areas within the ligament: Both fibrocartilagenous entheses of the PCL are devoid of blood vessels. A third avascular zone is located in the central zone of fibrocartilage of the middle third region.”

Exemples de requêtes :

N°	Requête	Texte de la requête
1		Arthroscopic treatment of cruciate ligament injuries.
6		HIV Epidemiology, Risk Assessment.
9		Patient-controlled analgesia indications and limits.
108		Cause of dysphagia.
109		Treatment of sensorineural hearing loss (SNHL).
124		New approach in cruciate ligament surgery.

Exemples de jugements de pertinence :

1	0	Arthroskopie/00130041	1
1	0	DerChirurg/70681093	1
1	0	DerUnfallchirurg/01030662	1
1	0	DerUnfallchirurg/81010491	1
2	0	DerChirurg/90701174	1
2	0	DerOrthopaede/70260267	1
2	0	DerOrthopaede/80270532	1
2	0	DerRadiologe/90390008	1
2	0	DerUnfallchirurg/01030795	1
2	0	DerUnfallchirurg/91020434	1
3	0	Arthroskopie/90120246	1
3	0	Arthroskopie/90120252	1
6	0	Bundesgesundheitsblatt/0043s003	1

4.6.2 Protocole d'évaluation

L'approche est évaluée en utilisant le système Mercure [Boughanem, 92]. L'évaluation est effectuée selon le protocole TREC. Plus précisément, chaque requête est soumise au système de RI avec les paramètres fixés. Le système renvoie les 1000 premiers documents pour chaque requête. Les valeurs de précision P5, P10, P15, P20, P30, P100 et P1000, ainsi que R-Prec (précision réelle ou exacte) et MAP (précision

moyenne) sont calculées. La précision au point 5, P5, est le ratio des documents pertinents parmi les 5 premiers documents restitués. R-Prec et MAP sont les précisions exacte et moyenne respectivement. Nous comparons ensuite les résultats obtenus à partir de notre approche à la baseline constituée de l'ensemble des résultats obtenus en utilisant le modèle de recherche de base, fondé sur les mots clés et une pondération $tf*idf$ classique.

4.6.3 Résultats expérimentaux

Notre évaluation expérimentale vise deux principaux objectifs :

- le premier objectif consiste à mesurer la viabilité de notre approche de détection de concepts. Pour cela, nous avons effectué une série d'expérimentations dont le but est de comparer l'indexation basée concepts par rapport à l'indexation simple basée mots clés. La formule de pondération utilisée est une formule classique $tf*idf$, tant pour les mots-clés simples que pour les concepts,
- le second objectif se rapporte à la viabilité de notre approche de pondération. Dans ce cas, dans l'objectif de mieux comprendre l'impact (négatif ou positif) de notre pondération, nous avons comparé notre approche basée sur les concepts pondérés par TF*idf (TF proposée), à la même approche utilisant une pondération des concepts par $tf*idf$ (classique).

4.6.3.1 Evaluation de la méthode d'identification de concepts

Les premières expérimentations menées concernent l'approche d'indexation par les concepts détectés selon notre approche (décrite en section 4.3.1), sans tenir compte de la pondération proposée au préalable. Le tableau 4.14 présente les résultats obtenus pour l'ensemble des requêtes tests. Les résultats montrent que notre approche est à l'origine d'un accroissement significatif des performances pour 52% des requêtes de test, avec des taux d'accroissement variables. Plus précisément, les taux significatifs (supérieurs à 5%) varient de 25% à 100%, de 33% à 100%, de 20,01% à 200% et de 8,83 à 101,98 pour respectivement la P5, P10, P15 et MAP.

Pour les autres requêtes, on note cependant une diminution des performances qui peut être due au fait que, lors de la projection du document sur l'ontologie, seuls les mots effectivement présents dans l'ontologie sont retenus dans le descripteur sémantique du document. Les termes absents dans l'ontologie sont ainsi ignorés. Même si certains de ces mots peuvent se retrouver dans le contexte relatif d'un mot adjacent, certains autres peuvent être complètement omis.

Requête	Mots-clés_tf_idf_classique				Concepts_tf-idf-classique			
	P5	P10	P15	MAP	P5	P10	P15	MAP
1	0.0000	0.0000	0.0667	0.0540	0.0000	0.0000	0.0667	0.0499
2	0.0000	0.0000	0.2000	0.1112	0.0000	0.1000	0.2000	0.1131
3	0.0000	0.2000	0.1333	0.1307	0.0000	0.2000	0.1333	0.1307
6	0.0000	0.1000	0.0667	0.0621	0.0000	0.1000	0.0667	0.0365
9	0.2000	0.1000	0.1333	0.1313	0.2000	0.2000	0.1333	0.1618
10	0.2000	0.1000	0.0667	0.1513	0.2000	0.2000	0.1333	0.3056
19	0.6000	0.5000	0.3333	0.4746	0.8000	0.5000	0.4000	0.5165
29	0.8000	0.8000	0.7333	0.6739	1.0000	0.8000	0.7333	0.6977
66	0.2000	0.1000	0.0667	0.0600	0.2000	0.1000	0.0667	0.0665
69	0.8000	0.6000	0.5333	0.2801	0.4000	0.5000	0.4000	0.2182
71	1.0000	0.9000	0.7333	0.6018	1.0000	0.8000	0.7333	0.5942
78	0.0000	0.0000	0.0667	0.0468	0.0000	0.0000	0.0000	0.0311
81	0.6000	0.3000	0.2000	0.1325	0.6000	0.4000	0.2667	0.1910
88	1.0000	0.7000	0.6000	0.3487	1.0000	0.7000	0.6000	0.3609
91	0.4000	0.2000	0.1333	0.1126	0.4000	0.3000	0.4000	0.2021
95	0.4000	0.5000	0.4667	0.4210	0.2000	0.1000	0.2667	0.0844
99	0.8000	0.6000	0.4000	0.2780	0.6000	0.5000	0.3333	0.2481
103	0.6000	0.5000	0.4000	0.2907	0.6000	0.4000	0.4000	0.2856
104	0.2000	0.1000	0.2000	0.1784	0.2000	0.2000	0.2000	0.2186
107	0.2000	0.1000	0.0667	0.0594	0.2000	0.1000	0.0667	0.0592
108	0.6000	0.8000	0.6000	0.3869	0.6000	0.8000	0.6000	0.3869
109	0.2000	0.4000	0.4000	0.2764	0.4000	0.2000	0.2000	0.1637
112	0.2000	0.1000	0.2000	0.1059	0.2000	0.1000	0.0667	0.0817
115	0.4000	0.3000	0.2000	0.1397	0.0000	0.0000	0.0000	0.0028
124	0.2000	0.2000	0.1333	0.0889	0.4000	0.2000	0.2000	0.1048

TABLEAU 4.14 : Résultats d'évaluation de la méthode de détection de concepts

4.6.3.2 Evaluation de la méthode de pondération des concepts

La deuxième série d'expérimentations menées concerne l'évaluation de notre approche de pondération des concepts (décrite en section 4.3.1). Le tableau 4.15 présente les résultats obtenus par comparaison à notre méthode d'indexation basée sur les concepts simplement pondérés. Les résultats révèlent que seules sept (7) requêtes ont présenté des taux d'accroissement significatifs relativement à une pondération simple par $tf*idf$. Les autres requêtes ont cependant présenté des résultats non concluants. Ceci peut s'interpréter en partie par le fait que la prise en compte de la fréquence d'occurrences d'un terme d'indexation donné dans l'ensemble des sens de son sous terme, peut produire du bruit. En

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

effet, si le terme est ambigu, il est possible que le sens du sous terme pris en compte ne corresponde pas au sens réel du terme dans le document. Ceci augmente alors sa fréquence sans corrélation avec le sens adéquat, et donc avec sa représentativité réelle dans le document. Ceci étant, il serait intéressant en perspective de ce travail de s'orienter vers une méthode de pondération basée sur la même approche, mais appliquée à des concepts préalablement désambiguïsés.

4.6.4 Conclusion

Nous avons décrit dans ce chapitre, une nouvelle approche de RI flexible basée sur un formalisme unifié, le formalisme CP-Net, à travers les tâches principales entreprises dans un SRI : l'indexation de documents, la formulation des requêtes et l'évaluation des requêtes. L'approche focalise sur deux aspects principaux. Le premier consiste en une indexation conceptuelle basée sur les CP-Nets. Le second concerne une nouvelle approche d'évaluation des requêtes CP-Nets.

L'approche d'indexation sémantique proposée est fondée sur l'utilisation conjointe de (1) l'ontologie WordNet pour identifier les concepts correspondants aux différents termes descriptifs du document, et (2) des règles d'association pour dériver des relations de dépendance contextuelle entre les concepts menant à une représentation plus expressive des documents.

Le principe même de l'approche n'est pas nouveau mais nous avons proposé de nouvelles techniques pour identifier, pondérer et désambiguïser les termes et pour découvrir des relations entre les concepts correspondants au moyen des règles d'association sémantiques proposées. Les règles d'association sémantiques permettent de découvrir des relations contextuelles entre les concepts conduisant à une représentation plus expressive de document.

Requête	Concepts_tf_idf				Concepts_TF_idf			
	P5	P10	P15	MAP	P5	P10	P15	MAP
1	0.0000	0.0000	0.0667	0.0499	0.0000	0.0000	0.0000	0.0413
2	0.0000	0.1000	0.2000	0.1131	0.0000	0.0000	0.0667	0.0867
3	0.0000	0.2000	0.1333	0.1307	0.0000	0.1000	0.1333	0.1083
6	0.0000	0.1000	0.0667	0.0365	0.0000	0.0000	0.0667	0.0739
9	0.2000	0.2000	0.1333	0.1618	0.2000	0.1000	0.0667	0.1269
10	0.2000	0.2000	0.1333	0.3056	0.0000	0.0000	0.0000	0.0258
19	0.8000	0.5000	0.4000	0.5165	1.0000	0.5000	0.4000	0.5972
29	1.0000	0.8000	0.7333	0.6977	0.8000	0.8000	0.8000	0.6877
66	0.2000	0.1000	0.0667	0.0665	0.0000	0.1000	0.0667	0.0222
69	0.4000	0.5000	0.4000	0.2182	0.0000	0.0000	0.1333	0.1392
71	1.0000	0.8000	0.7333	0.5942	0.0000	0.0000	0.0000	0.1491

CHAPITRE 4. APPROCHE DE RI SEMANTIQUE

78	0.0000	0.0000	0.0000	0.0311	0.0000	0.0000	0.0667	0.0293
81	0.6000	0.4000	0.2667	0.1910	0.4000	0.3000	0.2000	0.0924
88	1.0000	0.7000	0.6000	0.3609	1.0000	1.0000	0.8000	0.3957
91	0.4000	0.3000	0.4000	0.2021	0.2000	0.3000	0.2000	0.1391
95	0.2000	0.1000	0.2667	0.0844	0.4000	0.3000	0.3333	0.1717
99	0.6000	0.5000	0.3333	0.2481	0.0000	0.2000	0.2000	0.1813
103	0.6000	0.4000	0.4000	0.2856	0.4000	0.2000	0.2667	0.3068
104	0.2000	0.2000	0.2000	0.2186	0.4000	0.5000	0.4667	0.3068
107	0.2000	0.1000	0.0667	0.0592	0.0000	0.0000	0.0000	0.0190
108	0.6000	0.8000	0.6000	0.3869	0.8000	0.7000	0.5333	0.3427
109	0.4000	0.2000	0.2000	0.1637	0.4000	0.3000	0.2000	0.1504
112	0.2000	0.1000	0.0667	0.0817	0.0000	0.0000	0.0000	0.0513
115	0.0000	0.0000	0.0000	0.0028	0.2000	0.1000	0.0667	0.0566
124	0.4000	0.2000	0.2000	0.1048	0.4000	0.2000	0.1333	0.0706

TABLEAU 4.15 : Résultats d'évaluation de la méthode de pondération de concepts : impact de la méthode d'indexation par les concepts

L'approche d'évaluation proposée vise à évaluer la pertinence d'un document pour une requête donnée sur la base d'une mesure proposée de similarité des graphes CP-Nets correspondants. Nous avons expérimenté notre approche d'indexation sémantique par les concepts de WordNet. Les premiers résultats concernant l'apport de l'indexation par les concepts ont montré des améliorations significatives en précision moyenne et en précision exacte par rapport à l'approche d'indexation par les mots clés.

Conclusion générale

Synthèse

Les travaux présentés dans le cadre de cette thèse s'inscrivent dans trois axes différents mais néanmoins complémentaires d'un SRI :

1. l'amélioration de la représentation des requêtes par la prise en compte des préférences utilisateur,
2. l'amélioration de la représentation des documents par l'indexation sémantique,
3. l'évaluation flexible des requêtes.

Notre première contribution porte sur la prise en compte des préférences utilisateur dans le processus de recherche. Classiquement, les préférences utilisateur sont simplement exprimées par la pondération des critères de recherche. L'attribution des poids numériques de requêtes n'est pas sans problèmes, d'une part car les poids numériques des termes de la requête peuvent être interprétés de différentes manières (vu que différentes sémantiques sont associées aux poids) conduisant à des évaluations incorrectes, et d'autre part, car il n'existe pas de bonnes méthodes pour pondérer correctement les termes d'une requête. Des poids linguistiques, plus naturels et plus intuitifs, ont bien été introduits à travers le concept qualitatif et flou d'importance, néanmoins le problème de définition des poids numériques de requêtes est reporté à la définition du concept flou d'importance. Ces problèmes sont d'autant plus vrais pour les préférences conditionnelles qui, comme nous l'avons montré en introduction, peuvent conduire à des contradictions si quelques précautions ne sont pas prises en compte lors de la pondération. Nous avons alors proposé une approche qui :

1. permet de prendre en compte les préférences conditionnelles,
2. permet d'allier l'expressivité de la pondération intuitive qualitative, à la puissance calculatoire de la pondération quantitative numérique.

L'approche proposée définit le formalisme CP-Net comme langage d'expression des requêtes utilisateur portant sur des préférences qualitatives. L'utilisation des CP-Nets offre plusieurs avantages :

1. les CP-Nets supportent tout naturellement les préférences conditionnelles qualitatives,

2. ils offrent un formalisme graphique simple et intuitif qui permet de structurer ces dernières de manière compacte,
3. les fondements théoriques des UCP-Nets permettent la traduction correcte des valeurs de préférences qualitatives en valeurs numériques correspondantes, offrant ainsi le moyen de pondérer automatiquement des requêtes exprimant les préférences qualitatives de l'utilisateur.

La pondération automatique des requêtes et l'amélioration de la représentation des requêtes ne sont pas nos seules propositions dans le cadre de cette première contribution. Nous avons en effet proposé une approche d'interprétation des documents indexés en CP-Nets par projection dans l'espace des termes de la requête. Puis, nous avons proposé une approche d'évaluation des requêtes CP-Nets. L'approche consiste à :

1. retrouver les documents qui appartiennent aux termes de la requête par une approche de recherche classique,
2. traduire les documents retrouvés dans le formalisme CP-Net. L'approche utilisée ici consiste à projeter le document sur le même espace que la requête, et à le représenter par un CP-Net de même topologie que le CP-Net requête,
3. traduire les CP-Nets document et requête dans le paradigme booléen,
4. calculer la pertinence du document pour la requête booléenne en utilisant un opérateur d'agrégation flexible.

La proposition dans sa globalité définit ainsi un modèle théorique de RI flexible basé sur les CP-Nets. Notons cependant, que faute d'existence d'un cadre d'évaluation adéquat, l'approche n'a pas été testée expérimentalement.

Notre seconde contribution dans le cadre de la présente thèse, a pour objectif d'améliorer le modèle précédent au niveau de la représentation des documents et au niveau de l'évaluation.

Au niveau de la représentation des documents : Les SRI classiques indexent les documents et requêtes par les mots clés qu'ils contiennent. Pour calculer la pertinence document-requête, ces systèmes basent leur comparaison sur le nombre de mots que le document partage avec la requête. Dans une telle approche un document contenant des termes de la requête et pourtant non pertinent est retrouvé, alors que des documents pourtant pertinents ne partageant pas de mots avec la requête sont ignorés. L'indexation sémantique offre un moyen pour pallier ce problème en autorisant l'indexation des documents et requêtes par les sens des mots plutôt que par les mots qu'ils contiennent. Pour identifier le sens correct d'un mot dans un texte, les approches d'indexation sémantique se basent soit sur le contexte local du mot et sur

CONCLUSION GENERALE

des ressources externes, soit sur une certaine dimension sémantique latente entre les mots du texte comme c'est le cas dans la technique LSI.

En rejoignant ces idées, nous avons proposé une approche d'indexation sémantique comme combinaison des approches d'indexation par les sens des mots et par sémantique latente entre les mots du texte. Pour extraire cette dimension sémantique, nous nous sommes basés sur les règles d'association. Les règles d'association permettent la découverte de relations implicites, enfouies dans le texte, entre les termes du document. Notre approche d'indexation sémantique s'appuie sur deux étapes : une première étape d'extraction des sens des mots, et une seconde étape de découverte des relations latentes entre ces sens.

(1) La première étape s'appuie sur le contexte local et sur une ressource linguistique externe (WordNet en l'occurrence) pour déterminer le sens correcte d'un mot dans le document. La nouveauté de notre approche par rapport aux approches exogènes classiques d'indexation sémantique réside dans:

1. une nouvelle technique d'extraction des termes d'indexation par *mapping* sur l'ontologie WordNet,
2. une nouvelle technique de pondération des concepts introduisant le concept d'occurrence probable d'un terme dans les sens possibles de ses sous termes,
3. une approche de désambiguïsation proposée et en particulier le calcul du score de similarité symétrique basé tant sur le degré de corrélation des concepts dans l'ontologie que de l'importance des termes associés dans le document.

(2) La seconde étape concerne la découverte de relations contextuelles implicites entre les concepts issus de l'étape précédente. Le formalisme des règles d'association dans le contexte de la RI est classiquement défini pour découvrir des relations entre entités lexicales, à savoir les termes. Nous avons proposé un nouveau modèle de règles d'association dites règles d'association sémantiques, portant sur les relations d'association entre entités sémantiques du document, à savoir les concepts. Les règles d'association interprétées dans ce modèle, permettent de structurer les concepts représentatifs du document en faisant ressortir son topic. Les règles d'association sémantiques ainsi définies permettent ainsi de découvrir les associations sémantiques latentes entre les concepts représentatifs du document. L'ensemble formé des concepts représentatifs du document d'une part et des associations correspondantes d'autre part, est organisé en un graphe conditionnel, le CP-Net document.

Au niveau de l'évaluation Nous avons proposé une approche d'évaluation des requêtes basée sur l'appariement des graphes. En particulier, pour des requêtes CP-Nets (définies selon l'approche que nous avons proposée dans notre première

contribution), et des documents CP-Nets (définis par l'approche d'indexation sémantique proposée dans notre seconde contribution), la pertinence d'un document pour une requête se traduit par le degré de similarité des graphes CP-Nets correspondants. Cette approche d'évaluation, comparativement à celle proposée dans notre modèle de RI flexible basé sur les CP-Nets (première contribution), s'affranchit du paradigme booléen et offre un mécanisme plus flexible d'évaluation. Nous aurons ainsi posé les bases théoriques d'un modèle de RI flexible entièrement basé sur les CP-Nets.

Dans son état actuel, même si quelques résultats sont disponibles et confortent certaines de nos propositions (en l'occurrence celles concernant l'indexation sémantique), notre contribution reste globalement un modèle théorique de RI.

Perspectives

Les perspectives pour notre travail se déclinent en deux principaux points, le premier concerne la validation expérimentale des approches proposées, le second porte sur les améliorations possibles de ces approches.

Validation expérimentale :

La validation expérimentale d'un modèle de RI a pour objectif de tester et d'évaluer la viabilité du modèle et de le comparer par rapport à d'autres approches et modèles de référence. La validation expérimentale de notre modèle de RI proposé dans notre première contribution, nécessite la construction d'un cadre d'évaluation supportant des requêtes CP-Nets. La construction d'un tel environnement relève d'un travail de recherche à part entière, qui indépendamment du modèle proposé peut servir de base à la prise en compte des préférences conditionnelles dans le processus de RI.

Concernant notre second modèle proposé dans notre deuxième contribution, nous avons obtenu quelques résultats préliminaires qui montrent l'intérêt de notre approche d'indexation sémantique par la combinaison des concepts de WordNet et des règles d'association. Nous n'avons cependant pas expérimenté le modèle dans sa totalité. Il serait intéressant dans le futur de se pencher en particulier, sur la validation expérimentale de notre approche d'indexation sémantique en comparant l'apport de l'association concepts/associations correspondantes à une approche simplement basée sur les concepts. Il serait aussi intéressant de tester l'apport des règles d'association pour le SRI par rapport à des relations sémantiques classiques issues de l'ontologie (telles que celles utilisées dans [\[Baziz et al., 04 ;05\]](#)) . Par ailleurs, comme le modèle entier est aussi basé sur les CP-Nets,

CONCLUSION GENERALE

et en particulier sur les requêtes CP-Nets, on doit aussi construire le cadre d'évaluation adéquat pour expérimenter cette seconde approche.

Améliorations futures

Concernant notre modèle de RI flexible proposé en chapitre 3, l'amélioration principale que l'on pourrait apporter concerne la technique de pondération automatique des requêtes, en particulier la traduction des ordres de préférences qualitatifs en valeurs d'utilités correspondantes. L'approche proposée suggère des ordres de préférences uniformément distribués sur un domaine de valeurs donnés selon l'ordre de préférence qui y est spécifié. Il est alors impossible de prendre en compte des énoncés préférentiels modulés par des opérateurs linguistiques « extrêmes » (à l'exemple de : « je **préfère de loin** le jus d'orange au jus de pomme »). Pour pouvoir moduler ainsi les préférences utilisateur et en tenir compte lors de la pondération, les ordres de préférences qualitatifs devraient être traduits en ensembles flous de valeurs d'utilités. Ceci permettrait de fuzzyfier le langage de requête et fournirait le moyen pour une plus large expressivité des requêtes utilisateur.

Par ailleurs, notre modèle de RI flexible est basé sur la projection des documents sur l'espace de la requête. Un même document aurait alors différentes représentations qui lui sont associées à raison d'une par requête considérée. L'indexation est ainsi un processus coûteux en temps et en espace mémoire.

En outre, l'appariement dans ce modèle est basé mots-clés, et est fondé sur le paradigme booléen pour calculer la pertinence des documents pour la requête.

Ce sont ces raisons et d'autres qui nous ont conduits à la proposition d'un second modèle de RI (à travers notre seconde contribution). Ce dernier est une amélioration de notre première proposition, pour les points suivants :

1. *La représentation des documents* : notre modèle de RI sémantique s'affranchit des limitations des modèles basés mots-clés et propose une indexation par les concepts. Par ailleurs, chaque document possède une seule représentation CP-Net, indépendante de la requête,
2. *l'appariement* s'affranchit du paradigme booléen, en se basant sur une mesure proposée de similarité des graphes.

Finalement, notre second modèle de RI basé sur la sémantique, peut aussi être amélioré particulièrement au niveau de la découverte des règles d'association utilisé. En effet, notre approche de découverte des règles d'association préconise la découverte d'associations entre concepts individuels. Une amélioration possible consisterait à découvrir les relations entre ensembles de concepts, permettant ainsi de créer un réseau sémantique du document, certainement plus complexe mais aussi plus riche que celui que nous construisons à travers notre approche proposée en chapitre 4.

CONCLUSION GENERALE

Cette approche a été à l'origine de notre proposition dans [Boubekeur et al., 07], mais nous l'avons vite abandonnée à cause de la complexité des calculs impliqués. Néanmoins, nous restons convaincus qu'elle mènerait à une meilleure représentation des documents.

Par ailleurs, vu le nombre de concepts qui peuvent indexer les documents, et vu la complexité des calculs impliqués dans la découverte des règles d'association, une autre amélioration possible consisterait à regrouper au sein de clusters, des concepts sémantiquement proches dans l'ontologie, puis à adapter les règles d'association sémantiques à ce nouveau modèle de concepts. La représentation obtenue pourrait alors constituer une sorte d'abstraction orientée topic du document indexé.

Références bibliographiques

- [Agirre et al., 01] E. Agirre and D. Martinez. Knowledge sources for Word Sense Disambiguation. In Proceedings of the Fourth International Conference TSD 2001, Plzen (Pilsen), Czech Republic. Published in the Springer Verlag Lecture Notes in Computer Science series. Václav Matousek, Pavel Mautner, Roman Moucek, Karel Tauser (eds.) Copyright Springer-Verlag.
- [Agrawal et al., 93] Rakesh Agrawal, Tomasz Imielinski And Arun Swami : Mining Association Rules Between Sets Of Items In Large Databases. In Proc. Of The ACM SIGMOD International Conference Management Of Data, pp. 207-216, Washington, D.C., 1993
- [Agrawal et al., 94] R. Agrawal And R. Srikant. Fast Algorithms For Mining Association Rules In Large Databases (Santiago, Chile) In Proceedings Of The 20th Conference On Very Large Data Bases (VLDB'94), Pages 487-499. Morgan Kaufmann, International Conference September 1994.
- [Ahonen et al., 97] H. Ahonen, O. Heinonen, M. Klemettinen, And A. Verkamo. Applying Data Mining Techniques In Text Analysis. Technical Report, Department Of Computer Science, University Of Helsinki, 1997.
- [Albrecht et al., 98] Albrecht, R. and Merkl, D. 1998. Knowledge Discovery In Literature Data Bases. In Library And Information Services In Astronomy III. (ASP Conference Series, Vol. 153.) [Http ://Www.Stsci.Edu/Stsci/Meetings/Lisa3/Albrechtr1.Html](http://www.stsci.edu/stsci/meetings/lisa3/albrechtr1.html).
- [Alvarez et al., 03] Alvarez C., Langlais P., J.Y- Nie. Word Pairs in Language Modeling for Information Retrieval. Rapport interne, RALI. (2003).
- [Ambroziak, 97] J. Ambroziak . Conceptually assisted Web browsing. In the sixth International World Wide Web Conference. Santa Clara, CA.(1997). [http ://www.scope.gmd.de/info/www6/posters/702/guide2.html](http://www.scope.gmd.de/info/www6/posters/702/guide2.html).
- [Apte et al., 94] Apte, C., Damerau, F., And Weiss, S. M., “Automated Learning Of Decision Rules For Text Categorization”, ACM Transactions On Information Systems, Vol. 12. No. 3, July 1994, Pp. 233-251.
- [Audibert, 03] Audibert L., Outils d’exploration de corpus et désambiguïsation lexicale automatique. Thèse de Doctorat en Informatique de l’Université de Provence. Décembre 2003.
- [Audibert, 03] L. Audibert. Outils d’exploration de corpus Et désambiguïsation lexicale automatique. Thèse de Doctorat en Informatique de l’Université de Marseille. Décembre 2003.
- [Aussenac-Gilles et al., 00] Aussenac-Gilles N., Biébow B., Szulman N., Revisiting Ontology Design : a method based on corpus analysis. Knowledge engineering and knowledge management : methods, models and tools, Proc. of the 12th International Conference on Knowledge Engineering and Knowledge Management. Juan-Les-Pins (F). Oct 2000. R

REFERENCES BIBLIOGRAPHIQUES

- Dieng and O. Corby (Eds). *Lecture Notes in Artificial Intelligence Vol 1937*. Berlin : Springer Verlag. pp. 172-188. 2000.
- [Azé et al., 02] Azé J. et Kodratoff Y. (2002), Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. In spécial revue ECA, Actes Colloque EGC 2002, Montpellier, pp. 143-154.
- [Baeza-Yates et al., 99] Ricardo A. Baeza-Yates, Berthier A. Ribeiro-Neto: *Modern Information Retrieval* ACM Press. Addison-Wesley 1999
- [Balpe et al., 95] Balpe, J., Lelu, A., and Saleh, I. *Hypertextes et hypermédiat : réalisations, outils et méthodes*. Paris : Hermès, 1995.
- [Banerjee et al., 02] Banerjee, Satanjeev and Ted Pedersen. "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet" In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-02)*, Mexico City, February, 2002.
- [Banerjee et al., 03] BANERJEE S. & PEDERSEN T. Extended gloss overlaps as a measure of semantic relatedness. In *Eighteenth International Conference on Artificial Intelligence (IJCAI-03)*, p. 805–810, Acapulco, Mexico. (2003).
- [Bartschi, 85] Bartschi, M. (1985). Overview of Information Retrieval Subjects. *IEEE Computer. I&S*, 67-84; 1985.
- [Bautista et al., 04] M.J. Martin-Bautista, D. Sanchez, J. Chamorro-Martinez, J.M. Serrano, M.A. Vila. Mining Web Documents To Find Additional Query Terms Using Fuzzy Association Rules. *Fuzzy Sets And Systems* 148 (2004) 85–104)
- [Bayardo et al., 98] R. J. Bayardo. Efficiently Mining Long Patterns From Databases. In *Proceedings Of The 1998 ACM SIGMOD International Conference On Management Of Data (SIGMOD'98)*, Pages 85-93. ACM Press, June 1998.
- [Baziz et al., 03a] M. Baziz, N. Aussenac-Gilles, M. Boughanem. Désambiguisation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques. Dans : *Revue des Sciences et Technologies de l'Information (RSTI) série ISI*, Hermes, 11, rue Lavoisier, F-75008 Paris, V. 8, N. 4/2003, p. 113-136, décembre 2003.
- [Baziz et al., 03b] M. Baziz, N. Aussenac-Gilles, M. Boughanem. Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information. Dans : *XXIème Congrès INFORSID 2003*, Nancy, France, 3 janvier 6 juin 2003. INFORSID, Inforsid, 20 rue Axel Duboul - 31000 Toulouse, p. 121-134.
- [Baziz et al., 04] Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles. The Use of Ontology for Semantic Representation of Documents. Dans : *The 2nd Semantic Web and Information Retrieval Workshop(SWIR), SIGIR 2004*, Sheffield UK, 29 juillet 2004. Ying Ding, Keith van Rijsbergen, Iad Ounis, Joemon Jose (Eds.), pp. 38-45.
- [Baziz et al., 05a] Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles, Claude Chrisment. Semantic Cores for Representing Documents in IR. Dans : *SAC'2005- 20th ACM Symposium on Applied Computing*, Santa Fe, New Mexico, USA, 13 mars 17 mars 2005. ACM Press, New York, NY, USA, p. 1011 - 1017.
- [Baziz et al., 05b] Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles. A Conceptual Indexing Approach based on Document Content Representation. Dans : *CoLIS5 : Fifth International Conference on Conceptions of Libraries and Information Science*, Glasgow, UK, 4 juin 8 juin 2005. F. Crestani, I. Ruthven (Eds.), *Lecture Notes in Computer*

REFERENCES BIBLIOGRAPHIQUES

- Science LNCS Volume 3507/2005, Springer-Verlag, Berlin Heidelberg, p. 171-186.
- [Baziz et al., 05c] Baziz M., Indexation Conceptuelle Guidée Par Ontologie Pour La Recherche d'Information. Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Décembre 2005.
- [Beale et al., 95] Stephen Beale, Sergei Nirenburg and Kavi Mahesh. 1995. Semantic Analysis in the Mikrokosmos Machine Translation Project. In Proc. of the 2nd Symposium on Natural Language Processing, 297-307. Bangkok, Thailand.
- [Bédard, 07] Y. Bédard. Notes de Cours. (2007). <http://yvanbedard.scg.ulaval.ca/enseigne/SCG66124/DMSpatial.ppt>
- [Belew, 89] Belew R (1989) : Adaptive Information Retrieval : Using a Connectionist Representation to Retrieve and Learn about Documents. In : Belkin and Rijsbergen 1989. pp. 11-20.
- [Belkin et al., 87] BELKIN, N., AND CROFT, W B Retrieval Techniques. Annual Review of Information Science and Technology (ARIST), 22, (1987), 109-145.
- [Belkin et al., 92] Nicholas J. Belkin, Peter Ingwersen, Annelise Mark Pejtersen : Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992 ACM 1992
- [Berger, 99] A. Berger and J. Lafferty, Information Retrieval as Statistical Translation, Research and Development in Information Retrieval, Proc. ACM-SIGIR'99, pp. 222-229, 1999.
- [Berrut, 97] Berrut C., Indexation des Données Multimédia, Utilisation dans le Cadre d'un Système de Recherche d'informations. H.D.R. en Informatique de l'Université Joseph Fourier - Grenoble I. Octobre 1997.
- [Berry et al., 94] M.W. Berry , S. T. Dumais, G. W. O' Brien, Using linear algebra for Intelligent Information Retrieval, 1994.
- [Berzal et al., 02] F. Berzal, I. Blanco, D. Sánchez, M.A. Vila, Measuring The Accuracy And Importance Of Association Rules : A New Framework, *Intell. Data Anal.* 6 (2002) 221–235.
- [Blair et al., 85] David C. Blair, M. E. Maron : An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Commun. ACM* 28(3) : 289-299 (1985)
- [Bodner et al., 96] C. Bodner And Fei Song. Knowledge-Based Approaches To Query Expansion In Information Retrieval Richard (1996) Knowledge-Based Approaches To Query Expansion In Information Retrieval. In Mccalla, G. (Ed.), *Advances In Artificial Intelligence* (Pp. 146-158). New York : Springer.
- [Bookstein, 80] Bookstein, A. Fuzzy requests : an approach to weighted boolean searches. *Journal of the American Society for Information Science*, 31(4), 240-247, 1980.
- [Bordogna et al., 91] Bordogna G., Carrara P., and Pasi G., Query term weights as constraints in fuzzy information retrieval, *Information Processing and Management*, 27[1], 1991, p. 15-26.
- [Bordogna et al., 91b] Bordogna G., and Pasi G. Linguistic aggregation operators of selection criterion fuzzy information retrieval, *International Journal of Intelligent Systems*, 10, 233-248, 1995.
- [Bordogna et al., 93] Bordogna G., Pasi G., A fuzzy linguistic approach generalizing Boolean information retrieval : a model and its evaluation, *Journal of the American Society for*

REFERENCES BIBLIOGRAPHIQUES

- Information Science, 44[2], Mars 1993, p. 70-82.
- [Bordogna et al., 95] Bordogna G., Pasi G., Linguistic aggregation operators of selection criteria in fuzzy information retrieval, *International Journal of Intelligent Systems*, 10, 1995, p. 233-248.
- [Borlund, 03] Pia Borlund. The concept of relevance in IR. In *Journal of the American Society for Information Science and Technology*. Volume 54 , Issue 10 (August 2003).
- [Boughanem et al., 92] Mohand Boughanem, C. Soulé-Dupuy : A Connexionist Model for Information Retrieval. *DEXA 1992* : 260-265.
- [Boughanem, 92] Mohand Boughanem: les Systèmes de Recherche d'Information : d'un modèle classique à un modèle connexionniste. Thèse de Doctorat de l'Université Paul Sabatier, 1992.
- [Boughanem et al., 05] M. Boughanem, Y. Loiseau, and H. Prade. Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. In *Proc. of the 3rd International Workshop on Adaptive Multimedia Retrieval (AMR'05)* , Glasgow, UK, 28/07/05-29/07/05, pages 44–54. *LECTURE NOTES IN COMPUTER SCIENCE*, Springer, juillet 2005.
- [Boughanem et al., 07] Mohand Boughanem, Yannick Loiseau, Henri Prade. Refining Aggregation Functions for Improving Document Ranking in Information Retrieval. Dans : *International Conference on Scalable Uncertainty Management (SUM 2007)*, Washington,DC, USA, 10/10/07-12/10/07, Vol. 4772/2007, Springer-Verlag, p. 255-267, octobre 2007.
- [Bourigault, 96] Bourigault D. (1996) : " Lexter, a Natural Language Processing Tool for Terminology Extraction ". *Proceedings of Euralex'96*, Göteborg University, Department of Swedish, 1996, pp. 771-779.
- [Boutilier et al., 01a] C. Boutilier R. I. Brafman C. Domshlak H. H. Hoos and D. Poole. Preference-Based Constrained Optimization with CP-nets. *Computational Intelligence*, 20(2):137-157, 2001.
- [Boutilier et al., 01b] Craig Boutilier Fahiem Bacchus and Ronen I. Brafman. UCP-Networks: A Directed Graphical Representation of Conditional Utilities.. In *UAI*, pages 56-64, 2001.
- [Boutilier et al., 04b] C. Boutilier R. I. Brafman C. Domshlak H. H. Hoos and D. Poole. CP-Nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements. *Journal of Artificial Intelligence Research (JAIR)*, 21, 2004.
- [Boutilier et al., 97] Craig Boutilier, Ronen Brafman, Chris Geib, David Poole. A Constraint-Based Approach to Preference Elicitation and Decision Making. In *AAAI Spring Symposium on Qualitative Preferences in Deliberation and Practical Reasoning*. 1997.
- [Boutilier et al., 99] C. Boutilier R. I. Brafman H. H. Hoos and D. Poole. Reasoning with Conditional Ceteris Paribus Preference Statements. In *UAI'99*, pages 71-80, 1999.
- [Brafman et al., 02a] R. Brafman and C. Domshlak. Introducing Variable Importance Tradeoffs into CP-Nets. In *Workshop on Planning and Scheduling with Multiple Criteria*, April 2002.
- [Brafman et al., 02b] R. Brafman and C. Domshlak. Introducing Variable Importance Tradeoffs into CP-Nets. In *Proc. 18th Conf. on Uncertainty in AI (UAI'02)*, 2002.
- [Brafman et al., 04] R. Brafman C. Domshlak and E. S. Shimony. Qualitative Decision Making in

REFERENCES BIBLIOGRAPHIQUES

- Adaptive Presentation of Structured Information. *ACM Transaction on Information Systems*, 22(4):503-539, 2004.
- [Brin et al., 97] Brin S., Motwani R. et Silverstein C. (1997a), Beyond market baskets : generalized associations rules to correlations. In *Proceedings of ACM SIGMOD'97*, 1997.
- [Briscoe, 91] Briscoe, T. "Lexical Issues in NLP", en E. Klein & F. Veltman (eds.) *Natural Language and Speech*. The Netherlands: Springer-Verlag. (1991).
- [Bruce et al., 92] Bruce, R., Wilks, Y., Guthrie, L., Slator, B., Dunning, T. : *NounSense - A Disambiguated Noun Taxonomy with a Sense of Humour*. Research Report MCCS-92-246. Computing Research Laboratory, New Mexico State University (1992)
- [Buckley et al., 94] Buckley C., Salton G. and Allan J., The Effect of adding information in a relevance Feedback environment, in the *Proceedings of the ACM SIGIR Conference On Research and Development in Information Retrieval (SIGIR)*, pp 292-300, 1994.
- [Buell et al., 81a] Buell, D.A.; Kraft, D.H. Threshold Values and Boolean Retrieval Systems. *Information Processing & Management*, f7,127-135; 1981.
- [Buell et al., 81b] Buell, D. A. and Kraft, D. H. A model for a weighted retrieval system. *Journal of the American Society for Information Science*, 32(3), May, 211-216, 1981.
- [Buell, 82] Buell, D.A. An Analysis of some Fuzzy Subset Applications to Information Retrieval Systems. *Fuzzy Sets and Systems*, 7, 35-42; 1982.
- [Buitelaar et al., 04] Buitelaar, P., Steffen D., Volk, M., Widdows, D., Sacaleanu, B., Vintar, S., Peters, S., Uszkoreit, H., Evaluation Resources for Concept-based Cross-Lingual IR in the Medical Domain In *Proc. of LREC2004*, Lissabon, Portugal, May 2004.
- [Buitelaar, 98] P. Buitelaar. 1998. *CoreLex : systematic polysemy and underspecification*. Ph.D. thesis, Department of Computer Science, Brandeis University, Boston.
- [Callan et al., 92] J. Callan, B. Croft, and S. Harding. 1992. The INQUERY retrieval system. In *Proceedings of the 3rd Int. Conference on Database and Expert Systems applications*. DEXA 1992 : 78-83.
- [carroll et al., 89] Carroll, J. and C. Grover (1989) 'The derivation of a large Computational lexicon for English from LDOCE' in B. Boguraev and E. J. Briscoe (ed.), *Computational lexicography for natural language processing*, Longman, London, pp. 117-134.
- [Cater et al., 89] Cater, S.C.; Kraft, D.H. A generalization and clarification of the Wailer-&aft Wish List. *Information Processing & Management*, 25(1), 15-25; 1989.
- [Ceglar et al., 06] A. Ceglar and J. F. Roddick. Association Mining. *ACM Computing Surveys*, Vol. 38, No. 2, Article 5, Publication Date : July 2006.
- [Chang et al., 99] Chang, C.H., Hsu, C.C. Enabling Concept-Based Relevance Feedback For Information Retrieval On The WWW. *IEEE Transactions On Knowledge And Data Engineering*, Vol. 11, No.4, 1999.
- [Cheeseman et al., 96] P. Cheeseman And J. Stutz. Bayesian Classification (Autoclass) : Theory And Results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, And R. Uthurusamy, Editors, *Advances In Knowledge Discovery And Data Mining*, Pages 153-180. AAI Press, 1996.
- [Chen et al., 96] M. S. Chen, J. Han, And P. S. Yu, "Data Mining : An Overview From Database Perspective," *IEEE Trans. On Knowledge And Data Engineering*, Vol. 8, No. 6, Dec.

REFERENCES BIBLIOGRAPHIQUES

- 1996, Pp. 866-883.
- [Chen, 01] Chen, H. 2001. Knowledge Management Systems : A Text Mining Perspective. University Of Arizona (Knowledge Computing Corporation), Tucson, Arizona.
- [Cherfi et al., 02] H. Cherfi, Y. Toussaint. Adéquation D'indices Statistiques A L'interprétation De Règles D'association. JADT 2002 : 6es Journées Internationales d'Analysestatistique Des Données Textuelles. 13-15 Mars 2002 . Palais Du Grand Large Saint-Malo . France.
- [Cherfi, 04] Hacène Cherfi. « Etude Et Réalisation D'un Système D'extraction De Connaissances A Partir De Textes». THESE De DOCTORAT DE L'UNIVERSITE HENRI POINCARÉ. Nancy 1. Discipline : Informatique. Novembre 2004.
- [Chevalier, 02] Chevalier M., Interface Adaptative pour l'Aide à la Recherche d'Information sur le Web. Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Décembre 2002.
- [Chevallet, 97] Chevallet J. P., Bruandet M. F., Nie J. Y., Impact De L'utilisation De Multi Termes Sur La Qualité Des Réponses D'un Système De Recherche D'information. in Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information Collection UL3 Lille, Première Journées du Chapitre Français de l'ISKO à Lille. USBN 2-84467-002-4, Lille, France, pp223-238, 16-17 octobre, 1997.
- [Church et al., 90] K. Church and P. Hanks. Word association norms, mutual information and lexicography. In proceedings of the 28th Annual Meeting of the Association for Computational Linguistics. Pages 76-83. 1990.
- [Church, 92] K. W. Church (1992). A stochastic parts program and noun phrase parser for unrestricted text. Second Conference on Applied Natural Language Processing, Austin, Texas. pp. 136-143.
- [Cleverdon, 67] Cleverdon, C.. The cranfield tests on index language devices. In Aslib Proceedings, volume 19, pages 173-193, (1967).
- [Cleverdon, 70] Cleverdon, C. Progress in documentation. evaluation of information retrieval systems. Journal of Documentation 26 (1970), 55–67.
- [Cooper, 71] Cooper, W.S. (1971). A definition of relevance for information retrieval. Information Storage and Retrieval, 7, 19-37.
- [Couturier, 05] Olivier Couturier. « Contribution A La Fouille De Données : Règles D'association Et Interactivité Au Sein D'un Processus D'extraction De Connaissances Dans Les Données ». THESE En Vue De L'obtention Du Doctorat De l'Université d'Artois (Spécialité Informatique). Décembre 2005
- [Crestan et al., 03] CRESTAN E., EL-BÈZE M. & DE LOUPY C. (2003). Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique ? In 10e conférence TALN, p. 85–94, Batz-sur-mer, France.
- [Crestani et al., 98] Crestani, F., Lalmas, M., Campbell, I. and van Risbergen, C.J. Is this document relevant? ...probably. A survey of probabilistic models in information retrieval.ACM Computing Surveys. 1998.
- [Crestani et al., 99] F. Crestani and G. Pasi, "Soft Information Retrieval : Applications of Fuzzy Set Theory and Neural Networks", "Neuro-fuzzy tools and techniques", N.Kasabov Editor, Physica-Verlag , Springer-Verlag Group, 1999, pp. 287-313.
- [Croft et al., 91] W. Bruce Croft, Howard R. Turtle, David D. Lewis : The Use of Phrases and

REFERENCES BIBLIOGRAPHIQUES

- Structured Queries in Information Retrieval. SIGIR 1991 : 32-45
- [David et al., 06] Jérôme David, Fabrice Guillet, Régis Gras, Henri Briand. Aligement De Taxonomies Documentaires : Une Méthode Asymétrique Et Extensionnelle. Sdc 2006 - Semaine De La Connaissance. Nantes - 26 Au 30 Juin 2006
- [David et al., 96] David A. Evans and Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pages 17--24, Santa Cruz, CA, 1996.
- [De Mantaras et al., 90] De Mantaras, R. L., Cortes, U., Manero, J., and Plaza, E. Knowledge engineering for a document retrieval system. Fuzzy Sets and Systems, 38(2), November,1990.
- [Deerwester et al., 90] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas and Richard A. Harshman, 1990. "Indexing by Latent Semantic Analysis". In Journal of the American Society of Information Science, Vol. 41 :6, 391-407.
- [Delgado et al., 07] M. Delgado, M.J. Martín-Bautista, D.Sánchez, J.M. Serrano, M.A. Vila. Fuzzy Association Rules for Query Refinement in Web Retrieval. Book Chapter. In Studies in Fuzziness and Soft Computing book series. PublisherSpringer Berlin / Heidelberg ISSN1434-9922 (Print) 1860-0808 (Online) VolumeVolume 218/2008 Copyright2008 ISBN978-3-540-73184-9 DOI10.1007/978-3-540-73185-6_17 Pages351-362
- [Delgado et al., 02a] M. Delgado, M. J. Martin-Bautista, D. Sanchez Et M.A. Vila. Mining Text Data : Special Features And Patterns. Dans Pattern Detection And Discovery : Proc. Of ESF Exploratory Workshop, Rédacteurs D.J. Hand, N.M. Adams Et R.J. Bolton, Volume 2447 De Lecture Notes In Artificial Intelligence – LNAI, Pages 140–153, London, 2002. Springer-Verlag.
- [Delgado et al., 02b] M. Delgado, M.J. Martín-Bautista, D. Sánchez, J.M. Serrano, M.A. Vila. Association Rule Extraction For Text Mining. In Proceedings Of Flexible Query Answering Systems : 5th International Conference, FQAS 2002. Copenhagen, Denmark, October 27-29, 2002. Pages 154-162.
- [Denoyer, 04] Denoyer L., Apprentissage et Inférence Statistique dans les Bases de Documents Structurés : Application aux Corpus de Documents Textuels. Thèse de Doctorat en Informatique de l'Université de Paris 6. Décembre 2004.
- [Dixon, 97] Mark Dixon, (1997), An Overview Of Document Mining Technology, [Http ://Www.Geocities.Com/Researchtriangle/Thinktank/1997/Mark/Writings/Dixm97_Dm.P](http://www.geocities.com/researchtriangle/thinktank/1997/mark/writings/dixm97_dm.p)
- [Domshlak et al., 00a] C. Domshlak and R. I. Brafman. CP-nets - Reasoning and Consistency Testing. In Proc. 8th Int. Conf. on KR&R, pages 121-132, 2002.
- [Domshlak et al., 00b] Carmel Domshlak, Samir Genaim and Ronen Brafman. Preference-based Configuration of Web Page Content. In Proceedings of 3rd Workshop on Configuration, ECAI-2000, pages 19-22, August 2000.
- [Domshlak et al., 01] C. Domshlak, R. I. Brafman and E. S. Shimony. Preference-Based Configuration of Web Page Content. In Proc. 17th International Joint Conference on AI (IJCAI'01), pages 1451-1456, 2001.
- [Domshlak, 02] C. Domshlak. Modeling and Reasoning about Preferences with CP-Nets. Thesis submitted in Partial Fullfillment of the Requierements of the Degree of “ Doctor of Philosophy” of Ben-Gurion University of the Negev, Israël. 2002.

REFERENCES BIBLIOGRAPHIQUES

- [Doprado, 07] Hércules Antonio Do Prado. *Emerging Technologies Of Text Mining : Techniques And Applications / Hercules Antonio Do Prado & Edilson Ferneda, Editors.* ISBN 978-1-59904-373-9 (Hardcover) -- ISBN 978-1-59904-375-3 (Ebook). 2007
- [Doyle et al., 94] J. Doyle and M.P. Wellman. Representing Preferences as ceteris paribus comparatives. In *Working Notes of the AAAI Symposium on Decision-Theoric Planning.* AAAI, 1994.
- [Edmonds et al., 03] *Journal of Natural Language Engineering* (special issue based on Senseval-2) Editors : Phil Edmonds and Adam Kilgarriff. vol.9 no. 1, Jan. 2003.
- [Efthimiadis, 96] Efthimiadis, R. Query Expansion. *Annual Review Of Information Systems And Technology*, Vol. 31, Pp. 121-187, 1996.
- [El Wakil, 02] Mohamed M. El Wakil. Introducing Text Mining. In *9th Scientific Conference For Information Systems And Information Technology (ISIT02)*, Feb. 2002
- [Ester et al., 95] Martin Ester, Hans-Peter Kriegel, Xiaowei Xu. Knowledge Discovery In Large Spatial Databases : Focusing Techniques For Efficient Class Identification (1995). In *Advances In Spatial Databases, 4th International Symposium, SSD'95.*
- [Fagan, 87] Fagan, Joel L. 1987. Experiments in Automatic Phrase Indexing for Document Retrieval : A Comparison of Syntactic and Non-syntactic methods, PhD thesis, Dept. of Computer Science, Cornell University, Sept. 1987.
- [Fayet et al., 98] A. Fayet, A. Giacometti, D. Laurent, And N. Spyrtos. Découverte De Règles Pertinentes Dans Les Bases De Données. In *Actes Des 14èmes Journées Bases De Données Avancées (BDA'98)*, Pages 197-211, Octobre 1998.
- [Fayyad et al., 96] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From Data Mining To Knowledge Discovery : An Overview, In : U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), *Advances In Knowledge Discovery And Data Mining*, AAAI /MIT Press, California, USA, 1996. Pages 1_35.
- [Fayyad et al., 98] U. Fayyad, G. Piatetsky-Shapiro, And P. Smyth. From Data Mining To Knowledge Discovery : An Overview. *Advances In Knowledge Discovery And Data Mining*, MIT Press, 1 :1-36, 1998.
- [Feldman et al., 07] Ronen Feldman And James Sanger. *THE TEXT MINING HANDBOOK. Advanced Approaches In Analyzing Unstructured Data.* CAMBRIDGE UNIVERSITY PRESS. The Edinburgh Building, Cambridge CB2 8RU, UK. ISBN-10 0-511-33507-5 Ebook (Netlibrary). ISBN-13 978-0-511-33507-5 Ebook (Netlibrary). ISBN-13 978-0-521-83657-9 Hardback ISBN-10 0-521-83657-3 Hardback © Ronen Feldman And James Sanger 2007
- [Feldman et al., 95] R. Feldman Et I. Dagan. Knowledge Discovery In Textual Databases (KDT). Dans *In Proceedings Of The First International Conference On Knowledge Discovery And Data Mining (KDD-95)*, Rédacteurs U. M. Fayyad Et R. Uthurusamy, Pages 112-117, Montréal, Canada, 1995. AAAI Press.
- [Feldman et al., 98] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler Et O. Zamir. Text Mining At The Term Level. Dans *Proc. Of The 2nd Eur. Symp. On Principles Of Data Mining And Knowledge Discovery (PKDD'98)*, J. M. Zytkow Et M. Quafafou Editors, Volume 1510. *Lecture Notes In Artificial Intelligence – LNAI*, Pages 65-73, Nantes, 1998
- [Fellbaum, 98] FELLBAUM, Christiane, ed. (1998). *Wordnet – An Electronic Lexical Database*,

REFERENCES BIBLIOGRAPHIQUES

- The MIT Press, Cambridge, Massachusetts.
- [Fonseca et al., 05] Bruno M. Fonseca . Paulo Golgher. Bruno Pôssas. Berthier Ribeironeto. Nivio Ziviani. Concept Based Interactive Query Expansion.CIKM'05, October 31–November 5, 2005, Bremen, Germany. Copyright 2005 ACM 1595931406/
- [Fox, 92] Fox, C. Lexical analysis and stoplists. *Information Retrieval : Data Structures and Algorithms* (1992), 102–130.
- [Frakes, 92] Frakes, W. Stemming algorithms. *Information Retrieval : Data Structures and algorithms* (1992), 131–160.
- [Gale et al., 92a] Gale, W., Church, K. & Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. *International Conference on Theoretical and Methodological Issues in Machine Translation*, 101–112.
- [Gale et al., 92b] Gale, W., Church, K. & Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. In *Computers and the humanities* (Vol. 26, pp. 415–439). Kluwer Academic Publishers.
- [Ganter et al., 99] B. Ganter et R. Wille. *Formal Concept Analysis*. Edition Springer-Verlag, Heidelberg, 1999.
- [Gauch et al., 93] Gauch, S., Smith, J.B. An Expert System For Automatic Query Reformulation. *Journal Of The American Society For Information Science*, 44(3), 1993. Pp. 124-136.
- [Gaussier et al., 1997] E. Gaussier, G. Grefenstette, et M. Schulze. Traitement du langage naturel et recherche d'informations : quelques expériences sur le français. In *Premières Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*, 1997.
- [Gaussier et al., 2000] E. Gaussier, G. Grefenstette, D. Hull, et C. Roux. Recherche d'information en français et traitement automatique des langues. *revue Traitement Automatique des Langues (TAL)*, 41(2) :473–494, 2000.
- [Gonzalo et al., 98] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of COLING-ACL '98 Workshop on Usage of Word.Net in Natural Language Processing Systems*, Montreal, Canada, August.
- [Gonzalo et al., 99] J. Gonzalo, A. Pefias, and F. Verdejo. Lexical ambiguity and information retrieval revisited. In *Proceedings of EMNLP/VLC*, 1999.
- [Gras et al., 01]. Gras R., Kuntz P., Couturier R. Et Guillet F. (2001), Une Version Entropique De L'intensité D'implication Pour Les Corpus Volumineux, *Revue ECA, Extraction Des Connaissances Et Apprentissage, Hermès*, Vol. 1, 2001, Pp. 69-80.
- [Grishman et al., 94] Ralph Grishman, Catherine Macleod, and Adam Meyers. COMLEX syntax : Building a computational lexicon. In *Proceedings of COLING-94*, Kyoto, Japan. 1994.
- [Grobelnik Et Al., 00] Marko Grobelnik Dunja Mladenic Natasa Milic-Frayling *Text Mining As Integration Of Several Related Research Areas : Report On KDD'2000 Workshop On Text Mining (SIGKDD Explorations. Volume 2, Issue 2)*
- [Grolier] Grolier Multimedia Encyclopedia CD-ROM. Grolier interactive Inc., 90 Sherman Turnpike, Danbury, CT 06816, USA.
- [Gruber, 93] T.R. Gruber, A translation approach to portable ontology specifications, *Knowledge Acquisition*, 5 (2), pp 199-220, 1993.

REFERENCES BIBLIOGRAPHIQUES

- [Gruber, 95] Gruber, T. R., Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43(5-6) :907-928, 1995.
- [Guarino et al., 01] N. Guarino, C. Welty, Identity and Subsumption, In *The Semantics of Relationships : an Interdisciplinary Perspective*, R. Green, C.A. Bean, S. Hyon Myseng (Eds), Kluwer, pp 111-126, 2001.
- [Guthrie et al., 91] J.A. Guthrie, L. Guthrie, Y. Wilks, H. Aidinejad (1991). Subject-dependant co-occurrence and word sense disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkley, CA. :146-152.
- [Haddad et al., 00] H. Haddad, J.P. Chevallet and M. F. Bruandet, "Relations between Terms Discovered by Association Rules," 4th European conference on Principles and Practices of Knowledge Discovery in Databases PKDD'2000, Workshop on Machine Learning and Textual Information Access, France (2000).
- [Haddad, 02] Mohamed Hatem HADDAD. « Extraction Et Impact Des Connaissances Sur Les Performances Des Systèmes De Recherche d'Information ». Thèse De Doctorat De L'université Joseph Fourier. Discipline : Informatique. Septembre 2002.
- [Haddad, 03] Hatem Haddad : French Noun Phrase Indexing And Mining For An Information Retrieval System. *SPIRE 2003* : 277-286
- [Han et al., 95] J. Han and Y. Fu. Discovery Of Multiple-Level Association Rules From Large Databases. In *Proceedings Of The 21st International Conference On Very Large Data Bases (VLDB'95)*, Pages 420-431. Morgan Kaufmann, September 1995.
- [Hansson, 85] S.O Hansson. What is ceteris paribus preference. *Journal of Philosophical Logic*, 25(3): 307-332, 1996.
- [Harman, 92] Donna Harman : Relevance Feedback Revisited, in the *Proceedings of the ACM SIGIR Conference On Research and Development in Information Retrieval (SIGIR)*, pp 1-10, 1992.
- [Hayashi et al., 91] Hayashi, I., Naito, E., Wakami, N., Terano, T., Sugeno, M., Mukaidono, M., and Shigemasu, K. A proposal of fuzzy connective with learning function and its application to fuzzy information retrieval. In *Fuzzy Engineering Toward Human Friendly Systems*, 13-15 November, Yokohama, Japan, Amsterdam, The Netherlands, IOS Press, 446-55, 1991.
- [Hayes, 90] Hayes PJ. Intelligent high volume text processing using shallow, domain specific techniques. *Working Notes, AAAI Spring Symposium on Text-Based Intelligent Systems*, 1990 :134-138.
- [Hearst, 98] Hearst, M. A. (1998). Automated discovery of WordNet relations. In C. FELLBAUM, Ed., *WordNet: an electronic lexical database*, Language, Speech and Communication, chapter 5, pp. 131-151. Cambridge, Massachusetts: The MIT Press.
- [Hernandez, 05] Hernandez N., *Ontologies de Domaine pour la Modélisation du Contexte en Recherche d'Information*. Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Spécialité Informatique. Décembre 2005?.
- [Herrera-Viedma, 00] E. Herrera-Viedma. An Information Retrieval System with Ordinal Linguistic Weighted Queries Based on Two Weighting Semantic. *7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Bases Systems. IPMU'2000*. Madrid, (España), 2000, Vol. I, pp. 454-461.

REFERENCES BIBLIOGRAPHIQUES

- [Herrera-Viedma, 99] E. Herrera-Viedma. Modelling the Query Subsystem of an Information Retrieval System Using Linguistic Variables. IV Congrès ISKO-Espagne EOCONSID IV, 22-24 April 1999, Grenade, Espagne pp. 157-162.
- [Hiemstra, 98] D. Hiemstra, A linguistically motivated probabilistic model of information retrieval, dans Christos N and Stephanides C. (eds), Proc. European Conference of Digital Library (ECDL98), Sept. 1998, Springer Verlag.
- [Hirst, 97] Hirst, G. : Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press. Cambridge, England (1987)
- [Holt et al., 99] John D. Holt and Soon M. Chung . Efficient Mining Of Association Rules In Text Databases. CIKM '99 11/1999 Kansas City, MO, USA D 1999 ACM L-581 13-146-1/99/0010
- [Hornby, 74]. HORNBY, A. S. (ed), Oxford Advanced Learner's Dictionary of Contemporary English, 3e édition, London, OUP.1974,.
- [Hull, 96] Hull D. A., "Stemming Algorithms : A Case Study for Detailed Evaluation" Journal of the American Society for Information Science No 47(1). 1996. pp 70-84.
- [Ide et al., 90] Ide, N., & Véronis, J. Mapping Dictionaries : A Spreading Activation Approach, Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary (pp. 52-64). Waterloo (Canada). (1990).
- [Ide et al., 98] Ide, N. & Véronis, J. (1998). Word sense disambiguation : The state of the art. Computational Linguistics : Special Issue on Word Sense Disambiguation, 24, 1-40.
- [Imafouo, 06] Amélie IMAFOUO et Michel BEIGBEDER. Evaluer le passage à l'échelle dans des environnements à pertinence multivaluée. Dans AZctes de la Troisième Conférence en Recherche d'Information et Applications. Coria 2006. Lyon. 15-17 Mars 2006.
- [Ishioka, 03] Ishioka, T. Evaluation of criteria for information retrieval. Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on Volume , Issue , 13-17 Oct. 2003 Page(s) : 425 - 431.
- [Jacquemin et al., 02] Jacquemin, C., Daille, B., Royanté, J., and Polanco, X. 2002. In vitro evaluation of a program for machine-aided indexing. Inf. Process. Manage. 38, 6 (Nov. 2002), 765-792.
- [Jones et al., 02] Steve Jones, Gordon W. Paynter : Human evaluation of Kea, an automatic keyphrasing system. JCDL 2001 : 148-156.
- [Kamel et al., 90] Kamel, M., Hadfield, B., and Ismail, M. Fuzzy query processing using clustering techniques. Information Processing and Management, 26(2), 279-293, 1990.
- [Kantor, 81] P.B. Kantor, The Logic of Weighted Queries, IEEE Transactions on systems Man and Cybernetics 11 (1981), pp. 816-821.
- [Katz et al., 98] Özlem Uzuner, Boris Katz, Deniz Yuret : Word Sense Disambiguation for Information Retrieval. AAAI/IAAI 1999 : 985
- [Keefer, 94] X.A. Lu and R.B. Keefer. 1994. Query expansion/reduction and its impact on retrieval effectiveness. In The Text Retrieval Conference (TREC-3), pages 231-240.
- [Kelly et al., 75] Kelly, E. F. & Stone, P. J. (1975). Computer recognition of english word senses. North-Holland Publishing. North-Holland, Amsterdam.
- [Khan et al., 03] L. Khan, D. McLeod, E. Hovy, Retrieval effectiveness of an ontology-based

REFERENCES BIBLIOGRAPHIQUES

- model for information selection. Edited by F. Lochovsky. Received : October 7, 2002 / Accepted : May 20, 2003. Published online : September 30, 2003 – c_ Springer-Verlag 2003. The VLDB Journal (2004) 13 : 71–85 / Digital Object Identifier (DOI) 10.1007/s00778-003-0105-1
- [Khan et al., 04] Latifur Khan, Denis Mc Leod, Eduard Hovy. Retrieval effectiveness of an ontology-based model for information selection. The VLDB Journal (2004)13 :71–85.
- [Kilgarrif et al., 99] Adam Kilgarriff and Martha Palmer Computers and the Humanities (special issue based on Senseval-1) Editors : Adam Kilgarriff and Martha Palmer vol.34 no. 1-2, 1999.
- [Kilgarrif, 98] Kilgarrif, A. 1998. SENSEVAL : An Exercise in Evaluating Word Sense Disambiguation Programs. In Proceedings from First International Conference on Language Resources and Evaluation pp.581-588, Granada, Spanien.
- [Kim et al., 04] Hee-Soo Kim Ikkyu Choi Minkoo Kim .Refining Term Weights Of Documents Using Term Dependencies.SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK. ACM 1-58113-881-4/04/0007. 552
- [Kirkpatrick, 87] Kirkpatrick, Betty. (1987). Roget's Thesaurus of English Words and Phrases. Harmondsworth, Middlesex, England : Penguin.
- [Kirkpatrick, 88] Roget's thesaurus of English words and phrases (1988). New edition. Prepared by N. Kirkpatrick.Harmondsworth : Penguin.
- [Knight, 94] Knight K. and S.K. Luk. (1994). Building a Large-Scale Knowledge Base for Machine Translation. Proceedings of the AAAI Conference, 773–778
- [Kodratoff, 98] Yves KODRATOFF. Techniques Et Outils De L'extraction De Connaissances à Partir Des Données. Revue Signaux N°92 - Mars 1998.
- [Kodratoff, 99] Kodratoff Y. (1999), Quelques Contraintes Symboliques Sur Le Numérique En ECD Et En ECT, Ecole Modulad/Sfds-Inria, 1999.
- [Kohonen, 89] T. Kohonen. Self_Organization and Associative Memory. Springer Verlag. ISBN 0387513876, 1989.
- [Kraft et al., 03] D.H. Kraft , M.J. Martin-Bautista , J. Chen , D. Sanchez. Rules And Fuzzy Rules In Text : Concept, Extraction And Usage. In International Journal Of Approximate Reasoning 34 (2003) 145–161
- [Kraft et al., 78] D. E. Kraft, A. Bookstein. Evaluation of Information Retrieval System : A Decision Theory approach, Journal of the American Society for Information Science, 29 : 31–40, 1978.
- [Kraft et al., 83] Kraft, D. H. and Buell, D. A. Fuzzy sets and generalized Boolean retrieval systems. International Journal of Man-Machine Studies, 19(1), July, 45-56, 1983.
- [Kraft et al., 95] Kraft, D. H., Bordogna, G. and Pasi, G. An extended fuzzy linguistic approach to generalize Boolean information retrieval, Journal of Information Sciences -Applications, 2(3), 1995, pp 119-134.
- [KROEZE et al., 03] JAN H. KROEZE, MACHDEL C. MATTHEE AND THEO J.D. BOTHMA. Differentiating Data- And Text-Mining Terminology. Proceedings Of SAICSIT 2003, Pages 93 –101
- [Krovetz et al., 92] R. KROVETZ and W. B. CROFT. Lexical Ambiguity and Information

REFERENCES BIBLIOGRAPHIQUES

- Retrieval. ACM Transactions on Information Systems, Vol. 10, No 2, pp. 115_141. April 1992.
- [Krovetz, 93] Krovetz R, "Viewing Morphology as an Inference Process", in Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 191-202, 1993.
- [Krovetz, 97] R. Krovetz. 1997. Homonymy and polysemy in information retrieval. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (A CL-97}, pages 72-79.
- [Kwok, 89] Kwok K. L. (1989) : A Neural Network for Probabilistic Information Retrieval. In : Belkin and Rijsbergen 1989. pp. 21-30.
- [Lallich et al., 04]. S. Lallich, O. Teytaud. Évaluation Et Validation De L'intérêt Des Règles D'association, Revue RNTI, [Http ://Www.Techno-Science.Net/ ?onglet=Ouvrages & ID=2854286464,2004](http://www.techno-science.net/?onglet=Ouvrages&ID=2854286464,2004). [Http ://Perso.Wanadoo.Fr/Olivier.Teytaud/Publis/Evaluationetvalidationdelinteretdesreglesdasociation.Pdf](http://Perso.Wanadoo.Fr/Olivier.Teytaud/Publis/Evaluationetvalidationdelinteretdesreglesdasociation.Pdf)
- [Lang et al., 05] J. Lang, J. Goldsmith, M. Truszczynski, N. Wilson. The computational complexity of dominance and consistency in CP-nets- in Proceedings of IJCAI-05, 2005.
- [Latiri et al., 03] C. Ch. Latiri And S. Ben Yahia And J.P. Chevallet And A. Jaoua 3, Query Expansion Using Fuzzy Association Rules Between Terms, In JIM'2003 Conference Journées De l'Informatique Messine, Metz , France, September 3-6, 2003.
- [Lauer, 95] Lauer, Mark. 1995. Corpus statistics meet the noun compound : some empirical results. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp. 47-54.
- [Lavrenko01] V. Lavrenko and W.B. Croft, Relevance-based Language Models, Research and Development in Information Retrieval, Proc ACM-SIGIR'2001, pp. 120-127, 2001.
- [Leacock et al., 98] Leacock, C., Miller, G. A., and Chodorow, M. 1998. Using corpus statistics and WordNet relations for sense identification. *Comput. Linguist.* 24, 1 (Mar. 1998), 147-165.
- [Lent et al., 97] B. Lent, R. Agrawal, And R. Srikant. Discovering Trends In Text Databases. In Proceedings Of The 3rd International Conference On Knowledge Discovery And Data Mining (KDD'97), Pages 227-230. AAAI Press, August 1997.
- [Lesk, 86] Lesk M.E., Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from a nice cream cone. In Proceedings of the SIGDOC Conference. Toronto, 1986.
- [Lewis, 91] Lewis DD. Representation and learning in information retrieval. PhD Thesis, COINS Technical Report 91-93 Department of Computer and Information Science, University of Massachusetts, Amherst, MA 01003, 1991
- [Lin et al. 91] Lin, X., Soergei, D., and Marchionini, G. (1991) "A self-organizing semantic map for information retrieval". Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, (Chicago, IL, USA), pp. 262-269.
- [Lin et al., 96] Lin, S. H.. Chen, M. C., Ho, J. M.. And Huang, Y. M., "The Design Of An Automatic Classifier For Internet Resource Discovery", International Symposium On

REFERENCES BIBLIOGRAPHIQUES

- Multitechnology And Information Processing (ISMIP'96), December 1996, Pp. 181-188.
- [Lin et al., 98] S.H. Lin, C.S. Shih, M.C. Chen, J.M. Ho, M.T. Ko, Y.M. Huang, Extracting Classification Knowledge Of Internet Documents With Mining Term Associations : A Semantic Approach, In : Proc. ACM/SIGIR'98, Melbourne, Australia, 1998, Pp. 241–249.
- [Lin, 98] D. Lin. (1998) An information-theoretic definition of similarity. In Proceedings of 15th International Conference On Machine Learning, 1998.
- [Liu et al., 03] Rey-Long Liu, Wan-Jung Lin. Mining For Interactive Identification Of Users' Information Needs. *Information Systems* 28 (2003) 815–833
- [Liu et al., 05] Xiangwei Liu and Pilian He. A Study on Text Clustering Algorithms Based on Frequent Term Sets. In *Advanced Data Mining and Applications. Lecture Notes in Computer Science Book series. Volume 3584/2005. August, 2005.*
- [Liu et al., 98] Ye Liu, Hanxiong Chen, Jeffrey Xu Yu, Nobuo Ohbo : Using Stem Rules To Refine Document Retrieval Queries. *FQAS 1998* : 248-259
- [Loiseau, 04] Loiseau Y., Recherche Flexible d'information par Filtrage qualitatif Flou. Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Décembre 2004.
- [Longman, 88] Longman Dictionary of Contemporary English, New Edition, Longman
- [Lu et Keefer, 94] Lu X. A. and Keefer R. B. (1994). Query expansion/reduction and its impact on Retrieval effectiveness. Overview of the Third Text Retrieval Conference (TREC-3), NIST Special Publication 500-225, edited by D.K. Harman, 231-240.
- [Lucas, 99/00] LUCAS, M. 1999/2000. Mining In Textual Mountains, An Interview With Marti Hearst. *Mappa Mundi Magazine, Trip-M, 005, 1–3.* [Http ://Mappa.Mundi.Net/Trip-M/Hearst/](http://Mappa.Mundi.Net/Trip-M/Hearst/).
- [Luhn, 57] Luhn, H. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 4, 1(1957), 309–317.
- [Luhn, 58] Luhn, H. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 24, 2 (1958), 159–165.
- [Lungsawang et al., 99] A. Rungsawang, A. Tangpong, P. Laohawee, T. Khampachua, Novel Query Expansion Technique Using Apriori Algorithm, In : Proceedings Of The Eighth Text Retrieval Conference (TREC 8), 1999, Pp. 453–456.
- [Maedche et al., 00] Alexander Maedche, Steffen Staab : Discovering Conceptual Relations From Text. *ECAI 2000* : 321-325.
- [Mahgoub et al., 07] Hany Mahgoub, Dietmar Rösner, Nabil Ismail, Fawzy Torkey. A Text Mining Technique Using Association Rules Extraction. *INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE VOLUME 4 NUMBER 1 2007 ISSN 1304-2386*
- [Mandala et al., 72] Mandala, R., Tokunaga, T., and Tanaka, H. Combining multiple evidence from different types of thesaurus for query expansion. *Proc. Of the International ACM-SIGIR Conference, 1 (1972), 191–197.*
- [Mandala et al., 99] Mandala, Rila, Takenobu Tokunaga and Hozumi Tanaka (1999). Complementing WordNet with Roget and Corpus-based Automatically Constructed Thesauri for Information Retrieval Proceedings of the Ninth Conference of the European

REFERENCES BIBLIOGRAPHIQUES

Chapter of the Association for Computational Linguistics, Bergen.

- [Maniez et al., 91] Maniez, J., and de Grolier, E. A decade of research in classification. *International Classification* 18, 2 (1991), 73–77.
- [Mannila et al., 94] H. Mannila, H. Toivonen, And A. I. Verkamo. Efficient Algorithms For Discovering Association Rules. In *AAAI'94 Workshop On Knowledge Discovery In databases*, Pages 181-192. AAAI Press, July 1994.
- [Manning et al., 07] Manning C. D., R. Prabhakar, Schütze H., *An Introduction to Information Retrieval* Cambridge University Press. Cambridge, England. 2007
- [Maron et al., 60] Maron, M., and Kuhns, J. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery* 7 (1960), pages 216–244.
- [Masterman, 57] Masterman, M. (1957). The thesaurus in syntax and semantics. *Mechanical Translation*, 4, 1–2.
- [McGeachie, 02] McGeachie, M. Utility functions for ceteris paribus preferences. Master's Thesis. Massachusetts Institute of Technology, Cambridge, Massachusetts. 2002.
- [McRoy, 92] McRoy, S. : Using Multiple Knowledge Sources for Word Sense Discrimination. *Computational Linguistics*, 18(1) (1992)
- [Medelyan, 06] Medelyan O., Witten I. H., Thesaurus Based Automatic Keyphrase Indexing, in *Proceedings of JCDL'06*, June 11–15, 2006, Chapel Hill, North Carolina, USA.
- [Mihalcea et al., 00] Mihalcea, R. and Moldovan, D. : Semantic indexing using WordNet senses. In *Proceedings of ACL Workshop on IR & NLP*, Hong Kong, October 2000
- [Miller 95.] Miller G. (1995) WordNet : A Lexical database for English.. *Actes de ACM* 38, pp. 39-41.
- [Miller et al., 93] George A MILLer, Claudia Leacock, Randee Teng, and Ross T Bunker 1993 A semantic concordance In *Proceedings of the ARPA Human Language Technology Workshop*, pages 303-308
- [Miller, 90] Miller, G. A. (ed.), *WordNet : An on-line lexical database*. *International Journal of Lexicography* (special issue), 3(4) :235--312, 1990.
- [Mitra et al., 98] Mitra, A. Singhal, M., C. Buckley. Improving automatic query expansion. In the *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 206-214.
- [Miyamoto et al., 86] Miyamoto, S. and Nakayama, K. Fuzzy information retrieval based on a fuzzy pseudthesaurus. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-16(2), March-April 1986
- [Miyamoto, 90] Miyamoto, S. *Fuzzy sets in Information Retrieval and Cluster Analysis*. Kluwer Academic Publishers, 1990.
- [Mobasher et al., 01] Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa . Effective Personalization Based On Association Rule Discovery From Web Usage Data. *WIDM01* , 3rd ACM Workshop On Web Information And Data Management, November 9, 2001, Atlanta, Georgia, USA.
- [Moldovan et al., 00] D. Moldovan and R. Mihalcea. Using WordNet and lexical operators to improve Internet searches. *IEEE Internet Computing*, 4(1) :34-- 43. 2000.
- [Mothe, 94] Mothe J., *Modèle Connexionniste Pour la Recherche d 'Informations*. Expansion

REFERENCES BIBLIOGRAPHIQUES

- Dirigée de Requêtes et Apprentissage. Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Octobre 1994.
- [Mothe, 94a] Mothe J (1994) : Search Mechanisms Using a Neural Network Model. In : Intelligent Multimedia Information Retrieval Systems and Management. Proc. of RIAO '94. New York. pp. 275-294.
- [Mothe, 00] Mothe J., Recherche et Exploration d'Informations. Découverte de Connaissances pour l'Accès à l'Information. HDR en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Décembre 2000.
- [Mutlum, 05] B. Mutlum. Word Sense Disambiguation Based on Sense Similarity and Syntactic Context. Master Thesis of Science in Computer Engineering. Koc University September 2005.
- [Nanas et al., 03] Nikolaos Nanas, Victoria Uren, And Anne De Roeck. Building And Applying A Concept Hierarchy Representation Of A User Profile. In 26th International ACM SIGIR Conference On Research And Development In Information Retrieval, 2003.
- [Nastase et al., 01] Nastase, Vivi and Stan Szpakowicz. "Word sense disambiguation in Roget's thesaurus using WordNet." In: Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh, June 2001.
- [Nasukawa et al., 01] NASUKAWA, T. AND NAGANO, T. 2001. Text Analysis And Knowledge Mining System. IBM Systems Journal 40(4), 967-984.
- [Navigli et al., 03] Roberto Navigli and Paola Velardi. An Analysis of Ontology-based Query Expansion Strategies. In 2003 Workshop on Adaptive Text Extraction and Mining held in conjunction with : 14th European Conference on Machine Learning (ECML). www.dsi.uniroma1.it/~navigli/pubs/ECML_2003_Navigli_Velardi.pdf
- [Negoita, 73] Negoita, C.V. (1973). On the application of the fuzzy sets separation theorem for automatic classification in information retrieval systems. Information Science, 5, 279-286; 1973.
- [Neuwirth et al., 82] Neuwirth, E. and Reisinger, L. Dissimilarity and distance coefficients in automation-supported thesauri. Information Systems, 7(1), 1982.
- [Ng et al., 94] Raymond T. Ng and J. Han. Efficient And Effective Clustering Methods For Spatial Data Mining. In Proceedings Of The 20th VLDB Conference Santiago, Chile, 1994.
- [Paice, 84] Paice, C. D. Soft evaluation of Boolean search queries in information retrieval systems. Information Technology : Research Development Applications, 3(1), January, 33-41, 1984.
- [Park et al., 95] J. S. Park, M.-S. Chen, And P. S. Yu. An Efficient Hash Based Algorithm For Mining Association Rules. In Proceedings Of The 1995 ACM SIGMOD International Conference On Management Of Data (SIGMOD'95), Pages 175-186. ACM Press, May 1995.
- [Pasi, 99] Pasi. G., A logical formulation of the boolean model and of weighted boolean model, Proceedings of the Workshop on Logical and Uncertainty Models for Information Systems, London, UK, 1999, p. 1-11.
- [Pasquier et al., 98] N. Pasquier, Y. Bastide, R. Taouil, And L. Lakhal. Pruning Closed Itemset Lattices For Association Rules. In Actes Des 14èmes Journées Bases De Données Avancées (BDA'98), Pages 177-196, Octobre 1998.

REFERENCES BIBLIOGRAPHIQUES

- [Pasquier et al., 99a] N. Pasquier, Y. Bastide, R. Taouil Et L. Lakhal. Efficient Mining Of Association Rules Using Closed Itemset Lattices. *Information Systems*, 24(1) :25–46, 1999.
- [Pasquier et al., 99b] N. Pasquier, Y. Bastide, R. Taouil, And L. Lakhal. Discovering Frequent Closed Itemsets For Association Rules. In *Proceedings Of The 7th Biennial International Conference On Database Theory (ICDT'99), Lecture Notes In Computer Science*, Vol. 1540, Pages 398-416. Springer-Verlag, January 1999.
- [Pasquier, 00a] Nicolas Pasquier. Mining Association Rules Using Formal Concept Analysis. In : *Proceedings Of The ICCS 2000 International Conference On Conceptual Structures*, Pages 259-264, Springer, 2000.
- [Pasquier, 00b] Nicolas Pasquier. Thèse Docteur D'université Université Clermont-Ferrand II Ecole Doctorale Sciences Pour L'ingénieur de Clermont-Ferrand. Spécialité : INFORMATIQUE. « Data Mining : Algorithmes d'Extraction Et De Réduction Des Règles d'Association Dans Les Bases De Données ». Janvier 2000.
- [Peters et al., 04] Martin Braschler, Giorgio Maria Di Nunzio, Nicola Ferro, Carol Peters : CLEF 2004 : Ad Hoc Track Overview and Results Analysis. *CLEF 2004* : 10-26.
- [Piatetsky-Shapiro et al., 96] G. Piatetsky-Shapiro, U. Fayyad, And P. Smith. From Data Mining To Knowledge Discovery : An Overview. In : U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), *Advances In Knowledge Discovery And Data Mining*, AAAI /MIT Press, California, USA, 1996. Pp1_35.
- [Piatetsky-Shapiro, 91] Piatetsky-Shapiro G. (1991), Discovery, Analysis, And Presentation Of Strong Rules. In G.Piatetsky-Shapiro And W. J. Frawley, Editors, *Knowledge Discovery In Databases*, Pp. 229-248. AAAI Press / The MIT Press, 1991.
- [Poinçot, 99] Poinçot P., Classification et Recherche d'Information Bibliographique par l'Utilisation des Cartes Auto-Organisatrices, Applications en Astronomie. Thèse de Doctorat en Informatique de l'Université de Strasbourg. Décembre 1999.
- [Pons-Porrata et al., 07] Aurora Pons-Porrata, Rafael Berlanga-Llavori , José Ruiz-Shulcloper. Topic Discovery Based On Text Mining Techniques. *Information Processing And Management* 43 (2007) 752–768 . [Www.Elsevier.Com/Locate/Infoproman](http://www.Elsevier.Com/Locate/Infoproman)
- [Ponte et al., 98] Ponte, J. M., and Croft, W. B. A language modeling approach to information retrieval. research and development in information retrieval. In *Proc. of the International ACM-SIGIR Conference (1998)*, Proc. Of the International ACM-SIGIR Conference, pp. 275–281.
- [Porter, 80] M. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3) :130-137, July, 1980.
- [PÔSSAS et al., 05] BRUNO PÔSSAS, NIVIO ZIVIANI, And WAGNER MEIRA, JR. Set-Based Vector Model : An Efficient Approach For Correlation-Based Ranking. *ACM Transactions On Information Systems*, Vol. 23, No. 4, October 2005, Pp 397–429.
- [Prestwich et al., 04] S. Prestwich, F. Rossi, K. B. Venable, T. Walsh. Constrained CP-Nets. Preprint n. 13-2004, Department of Pure and Applied Mathematics, University of Padova, Italy.
- [Prié et al., 00] Prié, Y. " Sur la piste de l'indexation conceptuelle de documents. Une approche par l'annotation ". *Document Numérique*, numéro spécial " L'indexation ", 162 (4), p. 11-35. 2000.

REFERENCES BIBLIOGRAPHIQUES

- [Procter, 78] PROCTER P., I LSON R., Eds., Longman Dictionary of Contemporary English. Longman Harlow, Essex, 1978.
- [Pustejovsky, 95] Pustejovsky, J., Boguraev, B. & Johnston, M. A core lexical engine : The contextual determination of word sense (Tech. Rep.). Department of Computer Science, Brandeis University. (1995).
- [Qin et al., 04] Z. Qin, L. Liu, S. Zhang, Mining Term Association Rules For Heuristic Query Construction, In : PAKDD, 2004, Pp. 145–154.
- [Qing et al., 04] Qing Ma; Enomoto, K.; Murata, M., Self-organizing documentary maps for information retrieval. Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on Volume 1, Issue , Page(s) : - 9. 25-29 July 2004
- [Qui et al., 93] Y. Qui and H. P. Feri, "Concept Based Query Expansion," in Proc. of the Sixteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 160-169, 1993.
- [Quillian, 68] M. Quillian. Semantic Memory. In M. Minsky (Ed.), Semantic information Processing. The MIT Press, Cambridge, MA, 1968. Also PhD Thesis, Carnegie Institute of Technology, 1967.
- [Radecki, 79] Radecki, T. Fuzzy set theoretical approach to document retrieval. Information Processing and Management, 15(5), 247-260, 1979.
- [Rajman et al., 97] M. Rajman And R. Besancon. Text Mining : Natural Language Techniques And Text Mining Applications. In Proc. Of The 7th IFIP 2.6 Working Conference On Database Semantics (DS-7), Chapam And Hall IFIP Proceedings Serie,Leysin, Suisse, Octobre 1997.
- [Ralescu et al., 96] Ralescu A.L., Bouchon-Meunier B., Ralescu D.A. Combining Fuzzy Quantifiers, RR LAFORIA96/08, février 1996.
- [Resnik, 93a] Resnik, P. Selection and information : A class-based approach to lexical relationships. Unpublished doctoral dissertation, University of Pennsylvania. (1993).
- [Resnik, 93b] Resnik, P. Semantic classes and syntactic ambiguity. ARPA Workshop on Human Language Technology, 278–283. (1993).
- [Resnik, 95] Resnik, P. Disambiguating noun groupings with respect to WordNet senses. 3th Workshop on Very Large Corpora, 54–68. (1995).
- [Resnik, 99] Resnik, P., "Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research (JAIR), 11, pp. 95-130, 1999.
- [Ricart, 06] M. B. Ricart. Désambiguïation par propagation d'activation dans un thésaurus. rapport de DEA, Groupe MRIM - CLIPS-IMAG, juin, 2006.
- [Richardson et al., 95] Richardson R. and Smeaton A.F. (1995). Using WordNet in a knowledge-based approach to information retrieval, in Dublin City University Technical Reportn, (CA-0395).
- [Rigau et al., 97] Rigau, G., Atserias, J., Agirre, E. : Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. Proceedings of ACL-EACL, Madrid, Spain. (1997)
- [Robertson et al., 76] Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search

REFERENCES BIBLIOGRAPHIQUES

- terms. *Journal of the American Society for Information Science*, 27, 129–146.
- [Robertson et al., 92] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, Marianna Lau : Okapi at TREC. TREC 1992 : 21-30
- [Robertson et al., 97] S. E. Robertson and S. Walker. On relevance weights with little relevance information. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 16–24. ACM Press, 1997.
- [Robertson, 04] Robertson S., Understanding Document Frequency : On theoretical argument for IDF. In *Journal of Documentation* 60, n° 5, pp 503-520.2004.
- [Robertson, 77] Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33 (4), 294-304.
- [Robertson, 94a] ROBERTSON S., WALKER S., JONES S., GATFORD M. H.-B., « Okapi at 3 », *Proceedings of the 3rd Text REtrieval Conference (-3)*, p. 109-126, 1994.
- [Robertson, 94b] ROBERTSON S. E., WALKER S., « Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval », *Proceedings of SIGIR 1994*, p. 232-241, 1994.
- [Rocchio, 71a] Rocchio, J. J. Relevance feedback in information retrieval. In *The SMART Retrieval System, in Experiments in Automatic Document Processing* G. Salton, editor, Prentice-Hall, Englewood Cliffs, NJ, pp. 313–323, 1971.
- [Rocchio, 71b] J.J. Rocchio, Jr. *The SMART Retrieval System : Experiments in Automatic Document Processing*, chapter Relevance Feedback in Information Retrieval, pages 313–323. Prentice- Hall, 1971.
- [Rosario, 00] Rosario B. Latent Semantic Indexing : An overview. *INFOSYS 240*. Spring 2000.
- [Rossi et al., 04] Francesca Rossi, Kristen Brent Venable, Toby Walsh. mCP Nets: Representing and Reasoning with Preferences of Multiple Agents. In *Proceeding of the National Conference on Artificial Intelligence (AAAI' 04)*. San Jose, CA, USA. pages 729—734. July 2004.
- [Rumelhart et al., 86] Rumelhart, D.E., McClelland, J.L. and PDP Research Group. *Parallel Distributed Processing : exploration in the microstructure of cognition*. MIT Press, Cambridge, 1986.
- [Rungsawang et al., 99] A. Rungsawang, A. Tangpong, P. Laohawee, And T. Khampachua. Novel Query Expansion Technique Using Apriori Algorithm. In *TREC*, Gaithersburg, Maryland, 1999.
- [Sabah et al., 00] Gérard Sabah et Brigitte Grau, *Compréhension automatique de textes*, 2000, chap. 13, pp. 293-307, Ingénierie des langues, sous la direction de J.M.Pierrel, Hermes.
- [Salleb, 03] Ansaf Salleb « Recherche De Motifs Fréquents Pour L'extraction De Règles D'association Et De Caractérisation ». THESE De DOCTORAT DE L'UNIVERSITE d'Orleans. Discipline : Informatique. Décembre 2003.
- [Salton et al, 83a] Salton, G., E.A. Fox, H. Wu. Extended Boolean information retrieval system. *CACM* 26(11), pp. 1022-1036, 1983.
- [Salton et al., 73] Salton, G., and Yang, C. On the specification of term values in automatic indexing. In *Journal of Documentation*, 29 (1973), 351–372.
- [Salton et al., 75] Salton, G., Wong, A. & Yang, C. S. (1975). A vector space for information

REFERENCES BIBLIOGRAPHIQUES

- retrieval. *Communication of the Association for Computing Machinery (ACM)*, 18 (11),613–620.
- [Salton et al., 83a] Salton, G., Fox, E., and Wu, H. Extended Boolean information retrieval. *Communications of the ACM*, 26(12), 1983.
- [Salton et al., 83b] SALTON, G., AND MCGILL, M. *Introduction to Modern Information Retrieval* McGraw-Hill, New York, 1983.
- [Salton et al., 88] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management (IPM)* 24, 5 (1988), 513–523.
- [Salton et al.,90] Salton, G., and Buckley, C. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science (JASIS)* 44, 4 (1990), 288–297.
- [Salton, 00] Salton G., *Automatic text indexing using complex identifiers*. Proceedings of the ACM conference on Document processing systems. Santa Fe, New Mexico, United States. pp. 135 – 144. 2000.
- [Salton, 68] Salton, G. *Automatic Information Organization and Retrieval*. New York : McGraw.Hill Book Company, 1968.
- [Salton, 70] G. Salton, *The SMART retrieval system : Experiments in automatic document processing*. Prentice Hall, 1970.
- [Salton, 71] Salton, G. (1971). A comparison between manual and automatic indexing methods. *Journal of American Documentation*, 20(1) :61 {71.
- [Salton, 88] Salton, G. Syntactic approaches to automatic book indexing. In Proc. of the annual meeting on Association for Computational Linguistics (ACL) (1988), Department of Computer Science, Cornell University, Ithaca, New York, pp. 204–210.
- [Salton, 89] Salton, G., “Automatic Text Processing”, Addison Wesley, 1989.
- [Sanchez, 89] Sanchez, E. Importance in knowledge systems. *Information Systems*, 14(6), 1989.
- [Sanderson, 00] M. Sanderson. 2000. Retrieving with good sense. *Information Retrieval*, 2(1) :49--69.
- [Sanderson, 94] M. Sanderson. 1994. Word sense disambiguation and information retrieval. In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pages 142-151, Springer- Verlag.
- [Sanderson, 97] *Word Sense Disambiguation and Information Retrieval*, M. Sanderson, PhD Thesis, Technical Report (TR-1997-7) of the Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK, 1997.
- [Saracevic, 96] Saracevic, T. (1996). Relevance Reconsidered '96. In P. Ingwersen, & N.O. Pors (Eds.), *Proceedings of CoLIS 2, second international conference on conceptions of library and information science : Integration in perspective*, Copenhagen (pp. 201-218). Copenhagen : Royal School of Librarianship.
- [Sauvagnat, 05] Sauvagnat K., *Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés*. Thèse de Doctorat en Informatique de l'Université Paul Sabatier de Toulouse (Sciences). Juin 2005.
- [Savarese et al., 95] A. Savarese, E. Omiecinski, And S. Navathe. An E-Cient Algorithm For Mining Association Rules In Larges Databases. In Proceedings Of The 21st International

REFERENCES BIBLIOGRAPHIQUES

- Conference On Very Large Data Bases (VLDB'95), Pages 432-444. Morgan Kaufmann, September 1995.
- [Savoy, 05] J.Savoy . Indexation manuelle et automatique : une évaluation comparative basée sur un corpus en langue française. Dans proceeding de la 2ème Conférence Francophone en Recherche d'Information et Applications - CORIA 2005. Grenoble 9_11 Mars 2005.
- [Schütze et al., 95] H. Schütze and J. Pedersen. 1995. Information retrieval based on word senses. In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, pages 161-175.
- [Schütze, 92] Schütze H. (1992). Dimensions of meaning. *Supercomputing-1992*, 787–796.
- [Schütze, 98] Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics : Special Issue on Word Sense Disambiguation*, 24 (1), 97–123.
- [Senseval-2] SENSEVAL-2 : Second International Workshop on Evaluating Word Sense Disambiguation Systems 5-6 July 2001, Toulouse, France
- [Senseval-3] Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text An ACL 2004 workshop, Barcelona, Spain, July 25-26, 2004.
- [Senseval-4] Senseval-4/SemEval-1. ACL 2007. 4th International Workshop on Semantic Evaluations, February 26th to April 1st, 2007.
- [Sheridan et al., 92] Paraic Sheridan, Alan F. Smeaton : The Application of Morpho-Syntactic Language Processing to Effective Phrase Matching. *Inf. Process. Manage.* 28(3) : 349-370 (1992).
- [Shortlidge et al., 75] E. Shortlidge, B. Buchanan, A Model Of Inexact Reasoning In Medicine, *Math. Biosci.* 23 (1975) 351–379.
- [Simon, 00] A. Simon. Outils Classificatoires Par Objets Pour L'extraction De Connaissances Dans Les Bases De Données. Thèse De Doctorat, Université Henri Poincaré - Nancy 1, Nancy, 2000.
- [Simpson et al., 89] Simpson J. and Weiner E. The oxford English dictionary, 2nd edition, Oxford university Press, 1989.
- [Sinclair, 87] SINCLAIR, John. M. Looking up : an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English language dictionary , London : Collins ELT, 1987.
- [Singhal et al., 96] A. Singhal, C. Buckley, M. Mitra. Pivoted document length normalization. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval . Zurich, Switzerland .Pages: 21 - 29 . 1996
- [Sivanandam et al., 06] Sumathi, S., Sivanandam, S.N. Introduction To Data Mining And Its Applications. Series : Studies In Computational Intelligence , Vol. 29. 2006, XXII, 828 P. 108 Illus., Hardcover. ISBN : 978-3-540-34350-9.
- [Small et al., 82] Small, S. & Rieger, C. (1982). Parsing and comprehending with word experts (a theory and its realization). In L. Wendy & R. Martin (Eds.), *Strategies for natural language processing* (pp. 89–147). Hillsdale, New Jersey : Lawrence Erlbaum and Associates.
- [Smeaton et al., 95] Smeaton, Alan F., F. Kelledy and R. O'Donnell. TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish. Working paper CA-2295, School of Computer Applications,

REFERENCES BIBLIOGRAPHIQUES

- Dublin City University, Dublin, 1995.
- [Song et al., 05] Min Song, Il-Yeol Song, Xiaohua Hu, Robert B. Allen : Semantic Query Expansion Combining Association Rules With Ontologies And Information Retrieval Techniques. *Dawak 2005* : 326-335
- [Song et al., 07] Min Song, Il-Yeol Song B, Xiaohua Hu B, Robert B. Allen B. Integration Of Association Rules And Ontologies For Semantic Query Expansion. In *Data & Knowledge Engineering 63 (2007)* 63–75.
- [Song et al., 99] Fei Song, W. Bruce Croft : A General Language Model for Information Retrieval. *CIKM 1999* : 316-321.
- [Soualmia et al., 04] Soualmia LF., Darmoni SJ. Combining Knowledge-Based Methods To Refine And Expand Queries In Medicine. *FQAS, Flexible Query Answering Systems 2004*, 24-26 Juin 2004, Lyon France; Pp 14 (2004)
- [Soulé-Dupuy, 90] Soulé-Dupuy C. Systèmes de recherche d'informations : mécanismes d'indexation et d'interrogation. Thèse de Doctorat de l'Université Paul Sabatier, n°612, Toulouse III, février 1990.
- [Sparck Jones, 64] Sparck Jones, Karen (1964). *Synonymy and Semantic Classification*. Ph.D. thesis, University of Cambridge, Cambridge, England.
- [Sparck Jones, 86] Sparck Jones, K. (1986). *Synonymy and semantic classification*. Edinburgh, England : Edinburgh University Press.
- [Srikant et al., 95] R. Srikant and R. Agrawal. Mining Generalized Association Rules. In *Proceedings Of The 21st Int'l Conference On Very Large Databases (VLDB95)*, Zurich, Switzerland, September 1995.
- [Srinivasan et al., 01] P. Srinivasan, M.E. Ruiz, D.H. Kraft, J. Chen, Vocabulary Mining For Information Retrieval : Rough Sets And Fuzzy Sets, *Inform. Process. Manage.* 37 (2001) 15–38.
- [Stairmand et al., 96] Stairmand, Mark A. and W. J. Black. "Contextual and conceptual indexing using WordNet-derived lexical chains." In: *Proceedings of the 18th BCS-IRSG Colloquium on Information Retrieval Research, 1996*, pp. 47 - 65.
- [Stevenson et al., 01] :]STEVENSON M. & WILKS Y. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3), 321–351. (2001).
- [Studer et al., 98] R. Studer, R. Benjamins, D. Fensel, *Knowledge Engineering : Principles and Methods*, *Data and Knowledge Engineering*, 25(1-2) pp 161-197, 1998.
- [Sullivan, 00] SULLIVAN, D. 2000. The Need For Text Mining In Business Intelligence. *DM Review*, Dec. 2000. [Http ://Www.Dmreview.Com/Master.Cfm](http://www.dmreview.com/master.cfm).
- [Sun et al., 06] Renxu Sun And Chai-Huat Ong And Tat-Seng Chua. Mining Dependency Relations For Query Expansion In Passage Retrieval. *SIGIR 2006*.Pp 382-389.
- [Sussna, 93] Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. *2nd International Conference on Information and Knowledge Management (CIKM-1993)*, 67–74.
- [Tamine et al., 08] L. Tamine-Lechani et S. Calabretto. *RI contextuelle et web*. chapitre d'ouvrage. Editions Hermes. À paraître. 2008.
- [Tan, 99] A.-H. Tan. Text Mining : The State Of The Art And The Challenges. Dans *Proc. Of The*

REFERENCES BIBLIOGRAPHIQUES

- Workshop On Knowledge Discovery From Advanced Databases, Pages 65–70, Beijing, China, 1999. In Conjunction The third Pacific-Asia Conf. On Knowledge Discovery And Data Mining (PAKDD'99).
- [Thuraisingham, 99] Thuraisingham, B. 1999. Data Mining : Technologies, Techniques, Tools, And Trends. CRC Press, Boca Raton, Florida.
- [Toivonen et al., 96] H. Toivonen. Sampling Large Databases For Association Rules. In Proceedings Of The 22nd International Conference On Very Large Data Bases (VLDB'96), Pages 134–145. Morgan Kaufmann, September 1996.
- [Tommasi et al., 00] Marc Tommasi Rémi Gilleron. Découverte De Connaissances A Partir De Données. Cours Maîtrise MIAGE LIFL. Université Lille 3. [2000, Juin]. [Http://www.Grappa.Univ-Lille3.Fr/Polys](http://www.Grappa.Univ-Lille3.Fr/Polys).
- [Turtle et al., 90] Turtle, H. and Croft, W. B. Inference networks for document retrieval. In Proceedings of the 13th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Brussels, Belgium, September 05 - 07, 1990). J. Vidick, Ed. SIGIR '90. ACM Press, New York, NY, 1-24.
- [Turtle, 91] Turtle H. R., Inference Networks for Document Retrieval. PHD Thesis of the University of Massachusetts. February 1991.
- [Uschold et al., 95] M. Uschold, M. King, Towards a Methodology for Building Ontologies. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing at the International Joint Conference on Artificial Intelligence (IJCAI'1995), 1995.
- [Uzuner, 98] Uzuner, O. (1998). "Word Sense Disambiguation Applied to Information Retrieval". Master's Thesis. MIT. May 1998.
- [Van Rijsbergen, 77] C. J. Van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33 : 106-119. 1977.
- [Van Rijsbergen, 79] Van Rijsbergen, C.. Information Retrieval. Butterworths & Co., Ltd, London. (1979)
- [Vasilescu et al., 04] F. Vasilescu et P. Langlais. Désambiguïsation de corpus monolingues par des approches de type Lesk. TALN 2004, Fès, 19–21 avril 2004.
- [Veronis et al., 90] Véronis, J. and Ide, N. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. 13th International Conference on Computational Linguistics (COLING-1990), 2, 389–394. 1990.
- [Virginia disc 90] The Virginia disc one CD-ROM, published by Virginia Polytechnic Institute and State university Press. Editor, Project Director, Principal Investigator Edwadr A. Fox, Dept. of Computer Science 562 McBryde Hall, VPU&SU, VA 24061-0106.
- [Voorhees, 93] Voorhees, E. M. Using WordNet to disambiguate word senses for text retrieval. Association for Computing Machinery Special Interest Group on Information Retrieval. (ACM-SIGIR-1993) : 16th Annual International Conference on Research and Development in Information Retrieval, 171–180. (1993).
- [Voorhees, 94] E. M. Voorhees, " Query expansion using lexical-semantic relations," in Annual International ACM SIGIR conference on research and development in information retrieval (SIGIR'94), (pp pages 63–69). ACM press, Dublin, Ireland.
- [Voorhees, 98] Voorhees, Ellen M.. (1998). Using WordNet for text retrieval. In C. FELLBAUM, Ed., WordNet: an electronic lexical database, Language, Speech and

REFERENCES BIBLIOGRAPHIQUES

- Communication, chapter 12, pp. 285-303. Cambridge, Massachusetts: The MIT Press.
- [Voorhees, 99] E.M. Voorhees. Natural language processing and information retrieval. In Information Extraction : towards scalable, adaptable systems. Lecture notes in Artificial Intelligence, #1714, pages 32-48.
- [Vossen et al., 97] Vossen, P., Dez-Orzas, P., Peters, W. : The Multilingual Design of EuroWordNet. In : P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, Y. Wilks (eds.) Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.
- [Waller et al., 79] Waller, W. G. and Kraft, D. H. A mathematical model of a weighted Boolean retrieval system. Information Processing and Management, 15, 235-245, 1979.
- [Wei et al., 00] J. Wei, S. Bressan, B.C. Ooi, Mining Term Association Rules For Automatic Global Query Expansion : Methodology And Preliminary Results, In : First International Conference On Web Information Systems Engineering (WISE'00)-Volume 1, 2000, P. 366.
- [Weiss et al., 91] S. M. Weiss And C.A. Kulikowsky. Computer Systems That Learn : Classification And Prediction Methods From Statistics, Neural Nets, Machine Learning, And Expert Systems. Morgan Kaufman, 1991.
- [Weiss, 73] Weiss, S. F. (1973). Learning to disambiguate. Information Storage and Retrieval, 9, 33_41.
- [Wilks et al., 90] Y. Wilks, D. Fass, C. Guo, J.E. McDonald, T. Plate, B.M. Slator. Providing Machine Tractable Dictionary Tools. In Machine Translation, 5 :99-154. (1990)
- [Wilks et al., 97] Wilks, Y. & Stevenson, M. (1997). Combining independent knowledge source for word sense disambiguation. Conference « Recent Advances in Natural Language Processing », 1-7.
- [Wong et al., 85] Wong, S., Ziarko, W. et Wong, P. (1985). Generalized vector spaces model in information retrieval. In Proc. of the 8th ACM-SIGIR conference, pages 18-25. Montreal, Quebec.
- [Woods, 00] William A. Woods. 2000. Aggressive morphology for robust lexical coverage. In (these proceedings).
- [Woods, 91] William A. Woods. 1991. Understanding subsumption and taxonomy : A framework for progress. In John Sowa, editor, Principles of Semantic Networks : Explorations in the Representation of Knowledge, pages 45-94. Morgan Kaufmann, San Mateo, CA.
- [Woods, 97] William A. Woods. 1997. Conceptual indexing : A better way to organize knowledge. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, April. www.sun.com/research/techrep/1997/abstract-61.html.
- [Xu et al., 96] Xu J. and W.B. Croft. Query Expansion Using Local and Global Document Analysis. In the Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, pp. 4-11, Zurich, 1996.
- [Yager, 87] Yager, R. R. A note on weighted queries in information retrieval systems. Journal of the American Society for Information Science, 38(1) 1987.
- [Yager, 88] Yager, R. R. On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man and Cybernetics, 18(1), 183- 190, 1988.

REFERENCES BIBLIOGRAPHIQUES

- [Yarowsk, 95] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods. In 33rd Meeting of the Association for Computational Linguistics, p. 189–196. 1995.
- [Yarowsky, 92] Yarowsky, David (1992). "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora" Proceedings of the 14th International Conference on Computational Linguistics (COLING-92). Nantes, France, August, 454 – 460.
- [Yarowsky, 93] D. Yarowsky. 1993. One sense per collocation. In Proceedings of the ARPA Human Language Technology Workshop.
- [Yu et al., 82] Yu C.T., Lam K., Salton G., Term Weighting in Information Retrieval Using the Term Precision Model. In Journal of the Association for Computing Machinery. Vol 29, January 1982, pp 152-170.
- [Zadeh, 65] Zadeh, L. A. Fuzzy sets. Information and control, 8, 338-353, 1965.
- [Zadeh, 75] Zadeh, L. A. The concept of a linguistic variable and its application to approximate reasoning, parts I, II. Information Science, 8, 199-249, 301-357, 1975.
- [Zadeh, 75] Zadeh, L. A. The concept of a linguistic variable and its application to approximate reasoning, parts I, II. Information Science, 8, 199-249, 301-357, 1975.
- [Zaki et al., 97] M. J. Zaki, S. Parthasarathy, M. Ogihara, And W. Li. New Algorithms For Fast Discovery Of Association Rules. In Proceedings Of The 3rd International Conference On Knowledge Discovery And Data Mining (KDD'97), Pages 283–286. AAAI Press, August 1997.
- [Zaki et al., 98] M. J. Zaki. Scalable Data Mining For Rules. Phd Thesis, University Of Rochester, 1998.
- [Zhai, 04] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. in ACM Transactions on Information Systems, Vol. 2, Issue 2. 2004.
- [Zhang et al., 96] T. Zhang, R. Ramakrishnan, And M. Livny. Birch : An Efficient Data Clustering Method For Very Large Databases. In Proceedings Of The 1996 ACM SIGMOD International Conference On Management Of Data (SIGMOD'96), Pages 103-114. ACM Press, June 1996.
- [Zipf, 49] Zipf, H. Human behaviour and the principle of least effort. Addison- Wesley, Cambridge, Massachusetts, 1949.

ANNEXES

A. Evaluation des approches de désambiguïsation

Le succès de tout projet en WSD est clairement lié à l'évaluation du système de désambiguïsation utilisé. Pour ce faire, un corpus de test, pré-désambiguïsé est nécessaire d'une part pour évaluer la précision d'un système de désambiguïsation et d'autre part pour comparer les performances de différents systèmes. Les premiers chercheurs en WSD étaient, du fait de l'absence de corpus pré-désambiguïsé standard, souvent confrontés à la tâche fastidieuse de désambiguïsation manuelle de toutes les occurrences des mots à tester. Pour adresser ce problème, Yarowsky [Yarowsky, 93] a rapporté une nouvelle technique complètement automatique. La méthode comporte l'introduction dans un corpus, de mots ambigus artificiellement créés, appelés pseudo-mots. Un pseudo-mot est créé en remplaçant toutes les occurrences de deux mots, par exemple 'banana' et 'kalashnikov', par un nouveau pseudo-mot ambigu 'banana/kalashnikov'. Le corpus ainsi obtenu est utilisé comme corpus standard de test. L'évaluation des résultats d'un désambiguïseur est alors triviale puisque l'on connaît à l'avance les pseudo-sens corrects de chaque occurrence d'un pseudo-mot. Cependant, comme l'ont rapporté Gonzalo et al [Gonzalo et al., 99], la pseudo ambiguïté a un comportement différent de l'ambiguïté réelle. En effet, à la différence des composants d'un pseudo-mot, les différents sens d'un vrai mot polysémique sont souvent liés. L'ambiguïté réelle s'est avérée de ce fait plus tolérante aux erreurs que la pseudo ambiguïté. On comprend alors pourquoi les corpus à base de pseudo-ambiguïté ne peuvent constituer un standard de test.

Partant de ces constats, Kilgarrif [Kilgarrif, 98] a proposé en 1998, de construire un standard à partir duquel tous les modèles de désambiguïsation puissent se mesurer. C'est la première campagne de *Senseval*. La campagne *Senseval* est à notre connaissance le précurseur et la seule compagne d'évaluation standard des systèmes de WSD. Sa mission est d'organiser et de gérer l'évaluation et les activités relatives pour examiner les forces et les faiblesses des systèmes de désambiguïsation en ce qui concerne différents mots, différents aspects de langue, et différentes langues. *Senseval* est organisée par un petit comité sous les auspices d'ACL-SIGLEX (the *Special Interest Group on the Lexicon of the Association for Computational Linguistics*). Depuis son lancement en 1998, plusieurs campagnes d'évaluation ont eu lieu:

ANNEXE A. EVALUATION DES APPROCHES DE DESAMBIGUISATION

(1) *Senseval-1* a eu lieu en septembre 1998 en Angleterre. Lors de cette première campagne, il n'y a eu qu'une tâche évaluée (« *lexical sample task* »), la désambiguïsation d'un nombre limité de mots, sur trois langues, l'anglais, le français et l'italien. Cette évaluation est faite sur 15 noms, 13 verbes, 8 adjectifs et 5 adverbes.

(2) *Senseval-2* a eu lieu en l'été de 2001, et a été suivi d'un atelier tenu en juillet 2001 à Toulouse. *Senseval-2* a inclut des tâches pour le Chinois, le danois, le Néerlandais, l'anglais, l'italien, le japonais, l'espagnol, et autres suédois.

(3) *Senseval-3* a eu lieu en Mars-Avril 2004, suivie d'un atelier tenu en juillet 2004 à Barcelone. *Senseval-3* a inclut 14 tâches différentes pour la désambiguïsation, mais aussi pour l'identification des rôles sémantiques, les annotations multilingues, l'acquisition de sous-catégories... On retrouve les tâches « *all words* » et « *lexical sample* », qui correspondent respectivement aux modèles capables de désambiguïser automatiquement n'importe quel mot du texte et ceux adaptés à une liste limitée de mots. Les données utilisées pour la tâche « *lexical sample* » sont des exemples extraits du BNC (*British National Corpus*) et annotés sémantiquement à l'aide de WordNet 1.7.1. Il y a 60 mots à désambiguïser (noms, adjectifs, verbes ambigus).

(4) *Semeval-2007 / Senseval-4* a eu lieu en juin 2007 à Prague, incluant 19 taches dont la tâche *SemEval-2007* s'exécutant en collaboration avec CLEF (*the Cross-Language Evaluation Forum*). Il s'agit là d'une première tentative où la WSD est évaluée dans le cadre de la recherche documentaire et recherche documentaire inter linguistique (CLIR). Du point de vue de la WSD, cette tâche évaluera des systèmes de WSD indirectement sur une tâche réelle. Du point de vue de CLIR, cette tâche évaluera si les systèmes et les stratégies de WSD fonctionnent mieux, car comme on le verra dans le paragraphe suivant, l'ambiguïté lexicale est effectivement un problème en recherche d'information.

Des corpus de référence ont été rendus disponibles dans le cadre de chacune des campagnes d'évaluation.

B. Les CP-Nets

B.1 Introduction

La capacité de prendre des décisions et d'assumer des actions potentielles est un point-clé dans la majorité des problèmes d'intelligence artificielle incluant les systèmes experts, les systèmes à la décision, les systèmes recommandeurs, les outils de configuration... etc. [Brafman et al., 02a]. De nombreux outils automatisés d'aide à la décision ont été développés, certains pouvant prendre des décisions et les communiquer à l'utilisateur, d'autres aidant simplement l'utilisateur dans le processus de formulation et de prise de décision [Boutilier et al., 97]. Le but de la prise de décision est d'entreprendre l'action qui implique le meilleur résultat (i.e. le résultat le plus préférable). Les actions et les préférences sont représentées par un ensemble de contraintes sur un ensemble de variables (attributs) décisionnelles. Dans de nombreux domaines d'application, l'ensemble des actions possibles et des décisions potentielles est fixe et dépend d'une dynamique bien établie. Les seuls composants variables dans le processus de décision sont les préférences de l'utilisateur qui doivent être prises en compte lors de la prise de décision [Boutilier et al., 99]. En effet, tandis que des utilisateurs peuvent être confrontés à un problème décisionnel, leurs préférences sur les actions à entreprendre aux différents résultats décisionnels ne sont en général pas identiques.

Les fonctions d'utilité constituent un outil idéal pour la représentation et le raisonnement sur les préférences utilisateur [Brafman et al.,02a]. La représentation des préférences par une fonction d'utilité est primordiale pour le succès de nombreuses applications en intelligence artificielle. Une bonne fonction de préférence doit permettre de capturer des énoncés qui sont naturels, simples et intuitifs pour l'utilisateur. Cependant, les fonctions d'utilité peuvent être très difficiles à formuler et un effort considérable est requis de l'utilisateur [Brafman et al., 04]. Une difficulté majeure rencontrée dans l'extraction, la représentation et le raisonnement sur les préférences et les utilités concerne la taille de l'espace des résultats qui est exponentielle en nombre de variables caractéristiques du problème. De ce fait, l'expression directe de la fonction de préférence (fonction d'utilité) est quasiment infaisable. De ce fait, les systèmes d'aide à la décision ont émis différentes hypothèses sur la structure des préférences. L'hypothèse la plus souvent appliquée est

celle d'indépendance préférentielle (dont l'indépendance préférentielle, l'indépendance préférentielle conditionnelle et l'indépendance d'utilité mutuelle, ...) [Boutilier et al., 01b], permettant de décomposer la fonction d'utilité sur les alternatives en une somme ou un produit de fonctions de valeurs partielles sur les caractéristiques individuelles composant les alternatives. L'hypothèse d'indépendance permet ainsi de réduire le nombre d'alternatives à considérer, et de construire des fonctions d'utilité moins complexes.

Une alternative à cette approche consiste à raisonner en termes d'ordres de préférence qualitatifs plutôt qu'avec des fonctions de préférences numériques. Pour de nombreux domaines, en effet, les ordres qualitatifs sont plus naturels que les ordres quantitatifs [McGeachie, 02]. En se basant sur l'hypothèse d'indépendance préférentielle, les alternatives sont alors décomposées en leurs caractéristiques qualitatives individuelles indépendantes et le raisonnement se faisant sur ces ordres de préférence partiels.

Par ailleurs, Doyle et Wellman [Doyle et al., 94] ont observé que les représentations qualitatives de préférences sont une approximation raisonnable d'au moins un type de préférences humaines : les préférences Ceteris Paribus et ont développé des formalisations mathématiques pour de tels énoncés. Une préférence Ceteris Paribus spécifie des ordres de préférence sur certaines caractéristiques (ou attributs) tout en ignorant les caractéristiques restantes (supposées constantes).

Ainsi, considérons l'énoncé suivant (extrait de [Domshlak, 02]) :

« *I prefer red wine to white wine if served fish soup followed steak* »

Ceci signifie que, étant donnés deux repas qui diffèrent seulement dans le type de vin et qui contiennent tous les deux une soupe de poisson et du steak, je préfère le repas avec du vin rouge au repas avec du vin blanc. Cette préférence est ceteris paribus (toutes autres choses étant égales par ailleurs). Hansson [Hansson, 85] a établi que la plupart des préférences humaines semblent être de ce type. Domshlak dans [Domshlak, 02], a énoncé que les énoncés préférentiels qualitatifs ceteris paribus sont les meilleurs candidats pour la construction de modèles pratiques et utiles de préférences utilisateurs à cause de leur nature intuitive.

Des travaux récents ont exploité la structure d'indépendance préférentielle pour construire des modèles graphiques de représentation des énoncés de préférences Ceteris Paribus. Les premiers travaux dans ce sens ont été entrepris par Boutilier, Brafman, Hoos et Pool dans [Boutilier et al., 99]. Les auteurs ont proposé un graphe de représentation compacte de préférences qualitatives, le graphe CP-Net (*Conditional Preference Network*), qui exploite l'indépendance préférentielle conditionnelle pour la structuration des préférences utilisateur sous l'hypothèse Ceteris Paribus. Boutilier, Bacchus et Brafman [Boutilier et al., 99] proposent le modèle UCP-Net, qui étend le modèle CP-Net en permettant la représentation quantitative d'informations d'utilité

plutôt que de simples ordres qualitatifs de préférence. Prestwich, Venable, Rossi et Walsh [Prestwich et al., 04] ont proposé une nouvelle approche graphique étendant le modèle CP-Net à l'utilisation de contraintes fortes et souples. Brafman et Domshlak dans [Brafman et al., 02a] et [Brafman et al., 02b], étendent le modèle CP-Net à la manipulation de la notion d'importance entre variable conduisant au modèle TCP-Net. Finalement, les mCP-Nets proposés dans [Rossi et al., 04] étendent le formalisme CP-Net pour modéliser et supporter les préférences de multiples agents.

Nous nous intéressons dans ce qui suit aux CP-Nets. Nous introduirons d'abord quelques définitions utiles sur les relations de préférence et l'indépendance préférentielle, puis nous explicitons modèle CP-Net et son extension aux valeurs d'utilité : le UCP-Net.

B.2 Description avancée

Les CP-Nets ont été introduits en 1999 par Boutilier et al. dans [Boutilier et al., 99], comme outil de représentation compacte des relations de préférences qualitatives. Ce modèle graphique exploite l'indépendance préférentielle conditionnelle dans la structuration des préférences utilisateur sous l'hypothèse *ceteris-paribus*.

B.2.1 Un exemple illustratif

Cet exemple (extrait de [Boutilier et al., 04a]) spécifie les préférences utilisateur sur son costume de soirée. Le graphe consiste en 3 variables J , P et S correspondant respectivement à *Jacket* (jaquette), *Pants* (pantalons) et *Shirt* (Tee-Shirt) respectivement. L'auteur préfère inconditionnellement la couleur noire (*black*) à la couleur blanche (*white*) tant pour la jaquette que pour le pantalon, tandis que ses préférences entre les tee-shirt rouge (*red*) et blanc (*white*), est conditionnée par la combinaison (*jaquette, pantalons*). Si les deux sont de même couleur, alors il préfère un tee-shirt rouge. Si les deux sont de couleurs différentes, l'auteur préfère le tee-shirt blanc. Les préférences sont encodées dans le graphe CP-Net suivant (figure B.1):

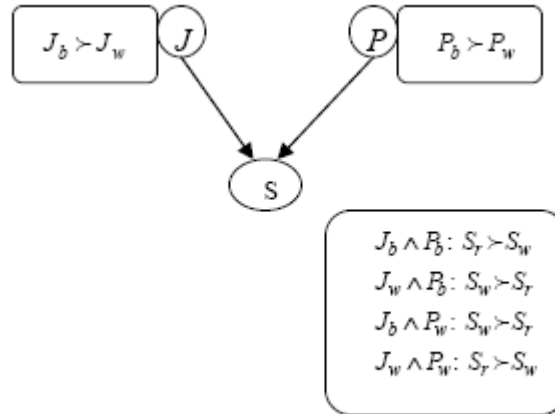


FIGURE B.1 : Le CP-Net

B.2.2 La sémantique du CP-Net

La sémantique de CP-nets est simple, définie en termes d'ensembles d'ordres de préférence qui sont consistants avec l'ensemble des contraintes imposées par les CPT [Boutilier et al., 04b].

Définition Soit N un CP-Net sur des variables V . $X \in V$ une variable, et $U \in V$ l'ensemble des parents de X dans N . Soit $Y = V - (U \cup \{X\})$. Soit \succ_u l'ordre de préférence sur $Dom(X)$ dicté par $CPT(X)$ pour une instantiation u de $Dom(U)$ des parents de X . Et soit \succ un ordre de préférence sur $Dom(V) = Dom(X_1) \times Dom(X_2) \times \dots \times Dom(X_n)$.

Un ordre de préférence \succ satisfait \succ_u ssi $\forall y \in Dom(Y) : x_i \succ_u x_j \Rightarrow yux_i \succ yux_j$

Un ordre de préférence \succ satisfait $CPT(X)$ ssi il satisfait \succ_u pour tout $u \in Dom(U)$

Un ordre de préférence \succ satisfait le CP-Net N ssi il satisfait $CPT(X)$ pour toute variable $X \in V$.

Définition : Un CP-Net N est satisfiable, s'il existe un ordre de préférence \succ qui le satisfait.

Théorème 1 : Tout CP-Net acyclique est satisfiable.

Remarque : Même si rien dans la sémantique des CP-Nets n'impose l'acyclicité, les CP-Nets cycliques peuvent fournir des ordres inconsistants [Domshlak et al., 00a], [Lang et al., 05]. Nous les ignorons volontairement pour ces raisons, et nous considérons dans la suite des CP-Nets acycliques.

ANNEXE B. LES CP-NETS

Généralement, la plupart des CP-Nets acycliques satisfiables sont satisfaits par plus d'un ordre de préférences. Ainsi, dans l'exemple 1 précédent (Figure B.1), il existe quatre ordres de préférences qui satisfont le CP-Net :

$$a_1b_1c_1 \succ \underbrace{a_1b_1c_2 \succ a_1b_2c_1}_{\text{}} \succ a_1b_2c_2 \succ a_2b_2c_2 \succ a_2b_1c_2 \succ a_2b_2c_1 \succ a_2b_1c_1$$

$$a_1b_1c_1 \succ \underbrace{a_1b_2c_1 \succ a_1b_1c_2}_{\text{}} \succ a_1b_2c_2 \succ a_2b_2c_2 \succ a_2b_1c_2 \succ a_2b_2c_1 \succ a_2b_1c_1$$

$$a_1b_1c_1 \succ a_1b_1c_2 \succ a_1b_2c_1 \succ a_1b_2c_2 \succ a_2b_2c_2 \succ \underbrace{a_2b_2c_1 \succ a_2b_1c_2}_{\text{}} \succ a_2b_1c_1$$

$$a_1b_1c_1 \succ a_1b_2c_1 \succ a_1b_1c_2 \succ a_1b_2c_2 \succ a_2b_2c_2 \succ \underbrace{a_2b_2c_1 \succ a_2b_1c_2}_{\text{}} \succ a_2b_1c_1$$

La déduction préférentielle dans les CP-Nets est définie de manière standard.

Définition : Soit N un CP-Net sur un ensemble de variables V . $o, o' \in \text{Dom}(V)$ deux alternatives quelconques.

N induit $o \succ o'$, et on note $N \models o \succ o'$ ssi $o \succ o'$ dans tout ordre qui satisfait N .

Ainsi, dans le cas du CP-Net précédent (Figure B.2), on a :

$$N \models a_1b_2c_2 \succ a_2b_2c_2$$

Mais $N \not\models a_1b_1c_2 \succ a_1b_2c_1$ car il existe un ordre de préférences dans lequel $a_1b_2c_1 \succ a_1b_1c_2$.

La déduction préférentielle pour un CP-Net est transitive :

$$\text{Si } N \models o \succ o' \text{ et } N \models o' \succ o'' \text{ Alors } N \models o \succ o''$$

La sémantique ceteris paribus du CP-Net implique que les préférences sur les parents ont une priorité supérieure à celles de leurs descendants. Ainsi par exemple, dans le CP-Net de l'exemple 1, on a $a_1b_2c_2 \succ a_2b_2c_2$: la plus préférable valeur de A combinée avec les valeurs les moins préférables de B et C , donne une alternative plus préférable que celle combinant la valeur la moins préférable de A avec les valeurs les plus préférables pour B et C étant donnée cette valeur de A .

B.2.3 Raisonner avec les CP-Nets

Comme tout modèle de représentation des préférences, le CP-Net permet deux types de raisonnement sur les préférences [Boutilier et al., 04b]:

3. Le premier concerne la recherche de la meilleure alternative possible: C'est l'optimisation des résultats.
4. Le second consiste à établir une comparaison préférentielle entre deux alternatives données : C'est le test de dominance.

1. Optimisation des résultats

Etant donné un CP-Net acyclique, on peut aisément déterminer la meilleure alternative possible sur les ordres de préférence qui satisfont le CP-Net. Il suffit pour cela de parcourir le graphe des préférences du sommet vers les feuilles, en initialisant chaque variable parcourue à sa plus préférable valeur étant données les instanciations de ses parents. En fait, même si le CP-Net ne détermine pas un ordre de préférence unique, il détermine une meilleure alternative unique.

De façon plus générale, étant donnée une contrainte sur quelques variables $Z \subseteq V$, sous forme d'une instanciation donnée z de Z , déterminer l'alternative la plus préférable consiste à parcourir, comme précédemment, le graphe des préférences de haut en bas, en assignant à chaque variable $X \neq Z$, sa plus préférable valeur étant donnée l'instanciation de ses parents.

2. Le test de dominance

Le problème de dominance dans un CP-Net N , d'une alternative o sur une alternative o' , peut être posé comme suit: $N \models o \succ o'$?

Dans [Boutilier et al., 99], il a été montré que la sémantique *ceteris paribus* du CP-Net, autorisait l'utilisation directe de l'information contenue dans la CPT d'une variable donnée X , pour changer (*flipping*) la valeur de X dans une alternative $o=uxy$, pour obtenir l'alternative $o'=ux'y$ immédiatement plus préférable si $x' \succ x$ (ou immédiatement moins préférable dans le cas où $x \succ x'$). Une séquence de flipping améliorant –*improving flipping sequence*– (respectivement détériorant –*worsening flipping sequence*–) d'une alternative o vers une alternative o' est toute suite d'alternatives o_1, o_2, \dots, o_k telle que $o_1 = o$, $o_k = o'$ et $\forall i = 1..k$, o_{i+1} est un flipping améliorant (respectivement un flipping détériorant) de o_i .

Dans [Boutilier et al., 04b], il a été montré qu'il existe une relation étroite entre l'existence d'une séquence de flipping entre une paire d'alternatives, et la relation de dominance entre elles. Plus précisément, une séquence de flipping améliorant –*improving flipping sequence*– (respectivement détériorant –*worsening flipping sequence*–) d'une alternative o vers une alternative o' fournit la preuve que o' est plus préférable

(respectivement moins préférable) à o dans tous les ordres qui satisfont le CP-Net [Domshlak, 02].

B.2.4 Utilisation des graphes CP-Nets

B.2.4.1 Introduction

Une utilisation intéressante des graphes CP-Nets concerne la présentation adaptative d'informations structurées. Un objectif important est de fournir une personnalisation orientée utilisateur (*viewer*) de l'information visualisée. Les approches proposées dans [Domshlak et al., 00b], [Domshlak et al., 01], [Brafman et al., 04] visent respectivement la présentation adaptative des contenus des pages Web, et plus généralement des documents structurés retournés par un provider de contenus. Contrairement aux approches classiques d'IA dans les hypermédias adaptatifs qui se basent sur l'apprentissage du profil utilisateur *et ne sont effectivement applicables que pour les utilisateurs fréquents*, les approches proposées par les auteurs dans, offrent une présentation dynamique en réponse aux sollicitations d'un utilisateur, sans avoir à apprendre son profil au préalable. Une présentation initiale est configurée selon les préférences de l'auteur, qui constitue l'expert du contenu. Les approches ainsi proposées se basent sur les graphes CP-Nets pour, d'une part structurer les préférences de l'auteur et offrir ainsi la présentation initiale, et d'autre part, pour assurer l'adaptabilité de la présentation en réponse à la sollicitation de l'utilisateur (*viewer*) via des algorithmes spécifiques des CP-Nets, pour la recherche de la configuration optimale (optimisation des alternatives).

Nous présentons ci-après le formalisme donné dans [Domshlak et al., 00b] et généralisé en [Brafman et al., 04], pour la présentation adaptative du contenu des pages Web

B.2.4.2 Le formalisme

Toute page web peut être considérée comme un ensemble de composants C_1, C_2, \dots, C_n . Chaque composant est associé à son contenu. Par exemple, le contenu d'un composant peut être un bloc de texte, une image, ...etc. Chaque composant peut être soit présenté à l'utilisateur *viewer*, soit caché. Ces options de présentation d'un composant C_i constituent ses valeurs possibles, elles sont représentées respectivement par c_i et c_i' . L'ensemble des composants d'une page web constitue ainsi un espace de configuration $\zeta = \{c_1, c_1'\} \times \{c_2, c_2'\} \times \dots \times \{c_n, c_n'\}$. Chaque élément σ dans cet espace constitue une configuration possible du contenu de la page web. En théorie de la décision, l'ensemble des composants de la page web, $V = \{C_1, C_2, \dots, C_n\}$ est un ensemble de variables, et chaque élément $\sigma \in \zeta$ définit une

alternative. Le concepteur de la page web spécifie pour chaque composant C_i de la page, l'ensemble des composants $\Pi(C_i)$ qui influencent ses préférences sur les options de présentation de C_i . Pour chaque configuration $\pi \in \Pi(C_i)$, le concepteur doit spécifier ses ordres de préférences sur les options $\{c_i, c'_i\}$ de C_i étant donnée π . Formellement, si $\overline{C}_i = \{C_1, \dots, C_n\} \setminus \{C_i, \Pi(C_i)\}$, alors C_i et \overline{C}_i sont conditionnellement préférentiellement indépendants étant donné $\Pi(C_i)$. Cette information est utilisée pour construire le graphe CP-Net, qui structure ainsi les préférences de concepteur sur la présentation de la page web. Ce CP-Net définit un ordre de préférence \succeq sur ζ , tel que $\forall \sigma_1, \sigma_2 \in \zeta, \sigma_1 \succeq \sigma_2$ signifie que le concepteur (auteur) de la page web voit σ_1 comme au moins aussi préférable que σ_2 . Une configuration optimale étant donné cet ordre de préférences sur ζ , est une alternative σ telle que $\sigma \succeq \sigma', \forall \sigma' \in \zeta$. Cet ordre de préférences est statique et ne dépend nullement de l'utilisateur (viewer). Pour assurer une présentation adaptative du contenu de la page web, on doit tenir compte des préférences de ce dernier. L'approche consiste dans un premier temps, à présenter la page dans sa meilleure configuration, à l'utilisateur. Ce dernier peut décider de visualiser un composant caché ou au contraire cacher un composant présenté (l'utilisateur peut par exemple interagir ainsi par simple click sur les composants). Ces préférences utilisateur constituent une contrainte que le système utilise pour reconstruire la meilleure alternative possible en respect aux préférences utilisateur.

B.2.4.3 Un exemple

Le processus décrit ci-dessus est illustré à travers l'exemple suivant extrait de [Domshlak et al., 00b]. La page web conçue est constituée de sept composants : quatre articles courts, et trois publicités. Les articles portent sur les élections en cours (*Elections*), un accident routier (*Traffic Accident*), un nouveau airbag de voiture (*New airbag*), et les résultats des récents jeux de NBA (*NBA*). Les publicités pour le magazine New York Times (*NY Times*), les voitures Volvo (*Volvo*), et les chaussures Nike (*Nike*). Après spécification du contenu de la page web, le concepteur exprime ses préférences sur la présentation du contenu :

Par défaut, la présentation de l'article central *Elections* ($C1$) est préférée à son masquage (i.e. $C1$ **on** est préférée à $C1$ **off**). Pour le second article *Traffic Accident*, $C2$ **off** est préférée à **on**.

L'article *New airbag* ($C3$) : $C3$ **on** est préféré seulement si *Traffic Accident* ($C2$) **on** et *Elections* ($C1$) est **off**.

L'article *NBA* ($C4$) on est préféré seulement si *Traffic Accident* n'est pas présenté.

ANNEXE B. LES CP-NETS

La publicité *NY Times* (C_5) est préférée seulement si *Elections* et *Traffic Accident* sont présentés.

La publicité *Volvo* (C_6) *on* est préférée si *New airbag* ou *Traffic Accident* sont présentés.

La publicité *Nike* (C_7) *on* est préférée seulement si *NBA* est *on*.

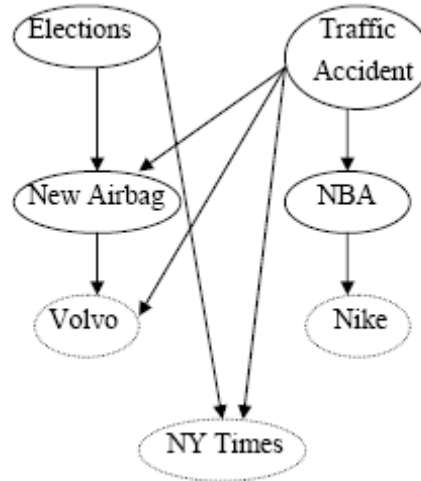


FIGURE B.2 : Exemple de CP-Net

Le graphe CP-Net correspondant est donné en figure B.2 et les tables CPT sont données comme suit :

$$C_1 \Rightarrow c_1 \succ c'_1$$

$$C_2 \Rightarrow c_2 \succ c'_2$$

$$C_3 \Rightarrow (c'_1 \wedge c_2) \rightarrow c_3 \succ c'_3; (c_1 \vee c'_2) \rightarrow c'_3 \succ c_3$$

$$C_4 \Rightarrow c_2 \rightarrow c'_4 \succ c_4; c'_2 \rightarrow c_4 \succ c'_4$$

$$C_5 \Rightarrow (c_1 \wedge c_2) \rightarrow c_5 \succ c'_5; (c'_1 \vee c'_2) \rightarrow c'_5 \succ c_5$$

$$C_6 \Rightarrow (c_2 \vee c_3) \rightarrow c_6 \succ c'_6; (c'_2 \wedge c'_3) \rightarrow c'_6 \succ c_6$$

$$C_7 \Rightarrow c_4 \rightarrow c_7 \succ c'_7; c'_4 \rightarrow c'_7 \succ c_7$$

Au téléchargement de cette page web, la présentation initiale de son contenu, donnée en figure B.3 (a), est déterminée par une procédure de reconfiguration. Dans la figure, les nœuds grisés représentent des composants visibles, les autres des composants invisibles (cachés). En supposant que le viewer clique sur le lien de *Traffic Accident* pour le visualiser, toute la structure du CP-Net original est repensée en fonction de cette nouvelle contrainte. Le résultat de la reconfiguration est donné en figure B.3 (b).

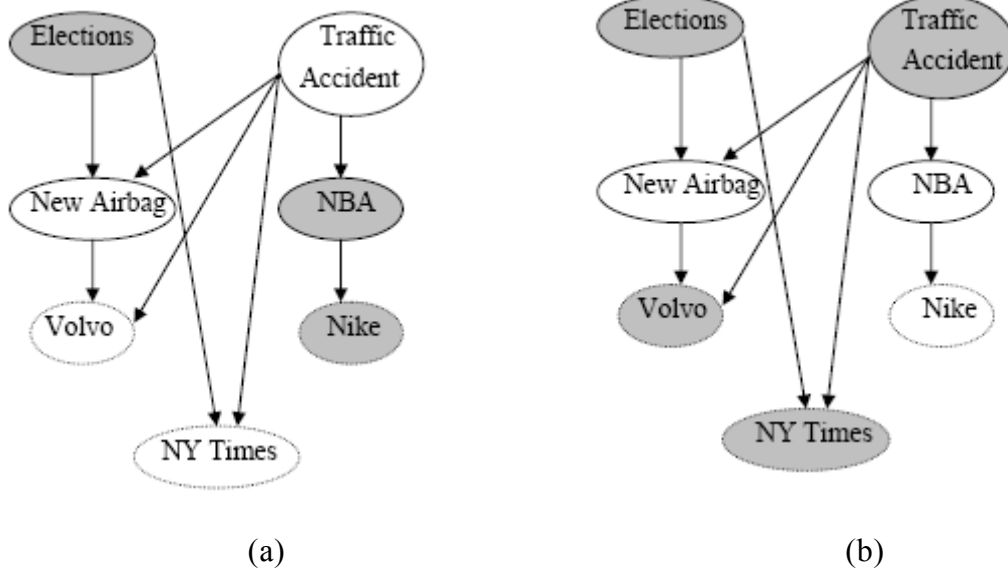


FIGURE B.3 : Exemple de reconfiguration du contenu

Nous avons présenté dans cette annexe les fondements théoriques du modèle CP-Net, et explicité sa sémantique. Puis, nous avons défini son extension à l'utilisation de valeurs d'utilité, conduisant au formalisme UCP-Net. Les CP-Nets ont été utilisés avec succès dans divers problèmes décisionnels.

C. La découverte de connaissances en RI

C.1 Introduction

Dans cette partie, nous présentons le processus d'extraction de connaissances dans les textes ou ECT (en anglais KDT, acronyme de *Knowledge Discovery in Texts*). L'ECT tire ses origines de l'ECBD, acronyme d'Extraction de Connaissances dans les Bases de Données (en anglais, KDD pour *Knowledge Discovery in Databases*), dont il hérite les techniques et méthodes. Nous introduisons dans un premier temps, le concept d'extraction de connaissances dans les bases de données, puis nous détaillons les méthodes d'extraction des connaissances dans les textes et ses applications en RI.

C.2 Extraction de connaissances dans les bases de données (ECBD)

C.2.1 Généralités

L'ECBD désigne le processus de découverte non triviale d'informations implicites, précédemment inconnues et potentiellement utiles à partir de données dans les bases de données [Piatetsky-Shapiro et al., 96]. Par le processus de KDD, les informations intéressantes et les régularités peuvent être extraites à partir d'un ensemble de données pertinentes contenues dans les bases de données et peuvent être analysées de différents points de vue.

L'extraction de connaissances à partir de grandes bases de données a été reconnue par de nombreux chercheurs comme un point clé dans les systèmes de base de données, et par de nombreuses compagnies industrielles comme un domaine important ayant des retombées capitales sur leur gestion. Les champs d'application du KDD sont vastes et variés allant de la gestion d'informations, le traitement des requêtes, la prise de décision, ... et autres analyse documentaire. Ainsi par exemple, l'extraction des connaissances d'une BDD transactionnelle des achats clientèle d'un super marché

permettrait de connaître les habitudes de consommation des clients. Ces connaissances serviront pour l'aide à la décision dans la réorganisation du rayonnage et dans la révision de la politique de marketing en fonction des produits les plus vendus, et ce dans l'objectif d'améliorer les ventes. En analyse documentaire, le KDD permettrait par exemple de regrouper des documents par topics, ou encore de classer des documents similaires.

Un système de KDD s'articule autour de quatre composants [Cherfi, 04 ; Simon, 00] :

1. Une ou plusieurs bases de données et leurs systèmes de gestion respectifs. Un système de KDD doit être capable d'une part, de traiter des masses de données volumineuses et de différents types (données temporelles, données spatiales, données légales, données transactionnelles, données multimédia,...), et d'autre part d'assurer la scalabilité (ou passage à l'échelle) de façon transparente pour l'analyste.
2. Un système à base de connaissances qui permet la gestion des connaissances. En particulier, différents types de connaissances peuvent être découvertes à partir de grandes bases de données. Le système de KDD doit adopter des techniques expressives de représentation des connaissances de sorte que les connaissances découvertes puissent être présentées à l'utilisateur dans une forme compréhensible et directement exploitable.
3. Un système de fouille de données (ou de *Data Mining*) qui permet l'exploration ou l'analyse des données de la base et la découverte de connaissances implicites précédemment enfouies. La fouille de données est le coeur du processus de KDD.
4. Une interface servant à l'interaction entre le système et l'analyste, et à la visualisation des résultats obtenus. L'analyste est chargé de guider les recherches et de valider les connaissances extraites.

C.2.2 Le Data Mining (DM)

L'expression de DM réfère souvent à l'ensemble des outils et méthodes permettant d'accéder aux données et de les analyser afin d'en extraire des modèles implicites, en prévision d'une utilisation future. C'est le processus automatique d'extraction non triviale de connaissances implicites, précédemment inconnues, et potentiellement utiles à partir de grandes bases de données. La définition la plus communément admise est donnée par [Fayyad et al., 98] :

« le Data Mining est un processus non trivial qui consiste à identifier, dans des données, des schémas nouveaux, valides, potentiellement utiles et surtout compréhensibles et utilisables »

Les méthodes et techniques de Data Mining sont multiples et variées. Néanmoins, les méthodes les plus novatrices concernent les règles d'association dans les bases de données relationnelles ou transactionnelles [Agrawal et al., 93 ; Agrawal et al., 94; Srikant et al., 95 ; Savarese et al., 95 ; Mannila et al., 94 ; Park et al., 95 ; Han et al., 95]. Nous examinons cette approche en section suivante.

Le Data Mining a été appliqué dans divers domaines allant de la grande distribution, la vente par correspondance, les opérateurs de télécommunications, les banques et assurances, etc... Le domaine majeur où le Data Mining a prouvé son efficacité est la gestion de la relation client (CRM ou Customer Relationship Management). En effet, le Data Mining permet par une meilleure connaissance de la clientèle d'accroître les ventes. Un domaine d'application plus récent concerne la gestion de connaissances dans les corpus textuels. Le text mining est ainsi né de l'application des techniques du Data mining aux textes. C'est cette application particulière que nous détaillerons en section 4.4.

C.3 Extraction de connaissances dans les bases de données textuelles (ECT)

C.3.1 Introduction

Du fait de l'importance croissante du contenu électronique et des médias électroniques pour le stockage et l'échange de documents textuels, est apparu un intérêt, de plus en plus croissant, pour les outils qui peuvent aider à retrouver l'information enfouie dans les textes de documents. La « découverte de connaissances à partir de bases de données textuelles » (DCT) [Haddad, 02] ou l'extraction de connaissances à partir de textes (ECT) [Cherfi, 04] ou encore le KDT « knowledge discovery in textual databases » [Feldman et al., 95], est une technologie naissante pour analyser de grandes collections de documents non structurés dans le but d'extraire des modèles (ou connaissances) intéressants, non triviaux et potentiellement utiles. Comme la forme la plus triviale de stockage de l'information est le texte, le KDT est censée avoir un plus haut potentiel commercial que l'exploitation de données structurées. Une étude récente a indiqué que 80% de l'information d'une compagnie est contenue dans des documents textuels tels que les emails, les notes, les correspondances de clients et les rapports. Les capacités pour distiller cette source inexploitée d'information, ces documents à textes libres, fournissent des avantages concurrentiels substantiels pour une compagnie pour réussir à l'ère de l'économie basée sur la connaissance.

C.3.2 La fouille de texte

L'expression fouille de textes ou text mining suggère qu'il s'agit de l'exploration de textes dans le but de retrouver l'information utile enfouie dans le texte. Quelques définitions citées ci-après explicitent le concept :

- "La fouille de textes (ou text mining) peut être définie comme l'application de méthodes calculatoires et de techniques sur des données textuelles dans le but de retrouver l'information pertinente, intrinsèque et la connaissance précédemment inconnue" [Doprado, 07].

- "La fouille de textes doit prospecter des pépites de nouvelles connaissances dans les montagnes de textes qui sont devenues accessibles aux recherches sur ordinateur grâce à la révolution de l'information et à l'interconnexion des réseaux " [Lucas, 99/00].

- "La fouille de textes est l'établissement de relations précédemment inconnues et insoupçonnées entre caractéristiques dans les bases de données textuelles..." [Albrecht et al., 98].

- "Nous définissons le texte mining comme étant le Data Mining sur des données textuelles. La fouille de textes est tout ce qui porte sur l'extraction de modèles et d'associations précédemment inconnus à partir de grandes bases de données textuelles" [Thuraisingham 99; Nasukawa et al., 01].

Il ressort de ces définitions que le text mining peut être vu comme un champ d'application du Data Mining aux textes ou du KDD aux textes [Ahonen et al., 97], [El Wakil, 02]. Le text mining réfère ainsi à l'ensemble des techniques et méthodes du Data Mining, en vue de retrouver, dans les textes de documents de grandes bases de données textuelles, l'information pertinente, utile, et précédemment inconnue.

C.3.2.1 Cadre du text mining

La fouille de textes est étroitement liée aux domaines du Data Mining, du traitement de la langue naturelle (NLP), de la gestion de connaissances (*knowledge management*), de l'extraction de l'information (IE), et la recherche d'information. Un système de fouille de données textuelles combine ainsi des techniques du Data Mining avec des techniques de traitement de la langue naturelle et d'extraction d'information et de recherche d'information. L'objectif est d'obtenir des connaissances utiles, précédemment inconnues et enfouies dans les textes [Haddad, 02]. Par différence aux systèmes de recherche d'information, les systèmes de text mining retrouvent l'information latente (cachée) et précédemment inconnue dans le texte alors que les SRI focalisent sur l'information visible, connue, contenue dans le texte. Chen dans [Chen, 01] rajoute que 'le text mining réalise différentes fonctions de recherche, d'analyse linguistique et de catégorisation. Les moteurs de recherche

eux, focalisent sur la recherche du texte, et plus particulièrement orientés sur la recherche par le contenu' [Kroeze et al., 03] .

C.3.2.2 Les étapes du text mining

La fouille de textes implique le prétraitement des collections de document (catégorisation des textes, extraction de l'information, extraction de terme), le stockage des représentations intermédiaires, les techniques pour analyser ces représentations intermédiaires (telles que l'analyse des distributions, le regroupement, l'analyse de tendances, et les règles d'association), et la visualisation des résultats [Feldman et al., 07]. Un système de fouille de textes suit les trois étapes principales suivantes [Cherfi, 04] :

- (1) la modélisation du contenu des textes,
- (2) la fouille de données,
- (3) l'analyse des résultats et validation.

L'étape de modélisation du contenu correspond à l'étape de préparation des données textuelles. La modélisation du contenu des textes permet d'extraire des données à partir de textes et les organiser dans une forme intermédiaire choisie. Elle correspond ainsi à une indexation terminologique des textes.

L'étape de fouille de données peut être lancée sur la base de données constituée. La forme intermédiaire basée-document permet de déduire des modèles et rapports entre documents. Le regroupement (clustering)/visualisation et la catégorisation de documents sont des techniques du text mining fondées sur des représentations intermédiaires basées-document. Les représentations intermédiaires basées-concepts permettent de déduire des modèles entre des concepts. Les travaux de fouille de données, telles que la modélisation prédictive et la découverte d'associations, sont basés sur cette catégorie [Tan, 99]. Dans le cas typique de la découverte d'associations, il s'agit d'extraire les règles d'association entre les termes-index identifiés, de classer ces règles selon des mesures de qualité et de les interpréter.

C.3.2.3 Les techniques du text mining

1. Techniques de prétraitement

Les techniques de prétraitement extraient des représentations structurées à partir de données textuelles non structurées [Feldman et al., 07]. Une grande variété de techniques de prétraitement des textes existe. Toutes tentent d'une certaine manière, de structurer des documents et, par extension, des collections de document. On distingue :

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

Les techniques de traitement du langage naturel (NLP)

Ces techniques utilisent et produisent les caractéristiques linguistiques indépendantes du domaine (dans le sens que leur résultat n'est pas spécifique à problème particulier). Les tâches impliquées peuvent inclure la tokénisation, l'analyse morphologique, l'étiquetage syntaxique (Pos tagging), et l'analyse syntaxique.

La catégorisation des textes

les tâches de catégorisation (ou classification) des textes étiquettent chaque document avec un nombre restreint de concepts ou de mots-clés. L'ensemble de tous les concepts ou mots-clés possibles est le plus souvent préparé manuellement.

Les techniques d'extraction d'information (EI)

L'EI est peut-être la technique la plus utilisée dans des opérations de prétraitement des textes. Sans techniques d'EI, les systèmes de fouille de textes auraient des possibilités plus limitées de découverte de la connaissance. L'EI doit être distingué de la recherche documentaire (ou recherche d'information). La recherche documentaire renvoie les documents qui appartiennent à une requête donnée mais exige toujours de l'utilisateur de lire ces documents pour localiser l'information pertinente. L'EI, vise pour sa part, à indiquer exactement l'information pertinente et à la présenter dans un format structuré.

2. Techniques de fouille de textes

Le noyau d'un processus de fouille de textes se compose de divers mécanismes pour découvrir des modèles d'occurrence de concepts dans une collection de documents ou dans un sous-ensemble donné de cette collection. Les trois types de modèles les plus communs en fouille de textes sont : les distributions (et les proportions), les ensembles fréquents et fréquents proches et les associations.

Les distributions

Les systèmes de fouille des textes permettent d'identifier la proportion de documents d'une collection donnée D , indexés avec un concept particulier K_i d'un ensemble de concepts K . On peut également identifier la proportion de documents indexés avec un concept K_2 qui sont également indexés par K_1 . Cette dernière proportion est connue sous le nom de : proportion conditionnelle de concepts. Généralement, un système de fouille des textes doit analyser la distribution des concepts qui sont des descendants d'un noeud particulier dans une hiérarchie de concepts. Une distribution importante de concept pour des opérations de découverte de la connaissance est la distribution de

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

proportions de concept qui donne la proportion de documents dans une certaine collection qui sont indexés avec chacun des concepts d'un ensemble donné de concepts.

Les ensembles fréquents et fréquents proches

En plus des proportions et des distributions, un autre type de base de modèle qui peut être dérivé d'une collection de document est un ensemble fréquents de concepts. Ceci est défini comme un ensemble de concepts représentés dans la collection de document avec des cooccurrences égales ou supérieures à un seuil minimal de support s (c.-à-d., tous les concepts de l'ensemble fréquent de concepts apparaissent dans au moins dans s documents). Essentiellement, un document peut être vu comme une transaction, et l'ensemble de ses caractéristiques (termes ou concepts représentatifs) comme des items. La découverte des ensembles fréquents peut être utile comme type de recherche des modèles à elle seule et comme étape préparatoire dans la découverte des associations. Les ensembles de concepts fréquents Proches établissent une relation non orientée entre deux ensembles fréquents de concepts. Cette relation peut être mesurée comme degré de recouvrement par exemple, sur la base du nombre de documents incluant tous les concepts des deux ensembles de concepts fréquents proches, ou comme fonction de distance entre les ensembles de concepts. Des relations orientées entre les ensembles de concepts peuvent également être identifiées. On parle alors d'associations.

Les associations

Une description formelle des règles d'association a été présentée pour la première fois dans les recherche sur le problème " du panier du marché ou panier de la ménagère". Elle est spécifiquement basée sur l'identification des ensembles fréquents. Dans la fouille de textes, les règles d'association ont été appliquées afin d'apprendre des relations de corrélations entre des éléments textuels, par exemple les termes constituant les mots-clés d'un texte [[Feldman et al., 98](#); [Kodratoff, 99](#); [Delgado et al., 02](#)].

3. Techniques de visualisation

Les approches de visualisation pour la fouille de textes supportent généralement un ensemble de buts différents de ceux des interfaces classiques. Bien que les deux visent à rendre l'interaction avec les données possible, les outils de visualisation sont des interfaces graphiques plus sophistiquées incluant les hiérarchies de concepts, les graphes d'associations entre concepts, les histogrammes, les courbes, les graphes circulaires, les cartes à auto-organisation ...

Les règles d'association constituent, tant dans le data mining que dans le text mining, la technique la plus utilisée. Nous la décrivons ci-après.

C.3.3 Découverte de règles d'association

La tâche d'association pour la fouille de données consiste à trouver quels attributs d'une base de données (relationnelle ou transactionnelle) "vont ensemble". La tâche d'associations aussi connue sous l'appellation d'analyse d'affinités ou analyse du panier de la ménagère a pour objectif de découvrir des règles pour mesurer le rapport entre deux ou plusieurs articles (items). Les règles d'association sont de la forme "si antécédent alors conséquent". Des mesures de support et de confiance, liées à la règle, sont définies pour déterminer un ensemble de règles fortes respectant un seuil minimal de support et de confiance respectivement.

C.3.3.1 Algorithmes de découverte des règles d'association

Les algorithmes de découverte des règles d'association se basent sur deux étapes [Agrawal et al., 94] :

générer tous les itemsets fréquents pour chaque itemset fréquent $Y = i_1 i_2 \dots i_k$,

générer toutes les règles d'association $X \rightarrow Y - X$, $\forall X \subset Y$.

La performance globale d'un algorithme de découverte de règles d'association est déterminée par la première étape. Après avoir déterminé les itemsets fréquents, les règles d'association correspondantes sont extraites de manière triviale. Plusieurs algorithmes pour la découverte des itemsets fréquents, nous le détaillons ci-après.

Algorithmes de découverte des itemsets fréquents

Etant donné un ensemble I , d'items de taille m . L'ensemble des parties de I (de taille 2^m), muni de la relation d'inclusion (relation d'ordre) définit un treillis d'itemsets de hauteur $(m+1)$. Par exemple, le treillis des parties de l'ensemble d'items $I = \{a, b, c, d, e\}$ est représenté en figure C.1 suivante. L'ensemble I contenant 5 items, ce treillis contient 32 itemsets et sa hauteur est égale à six.

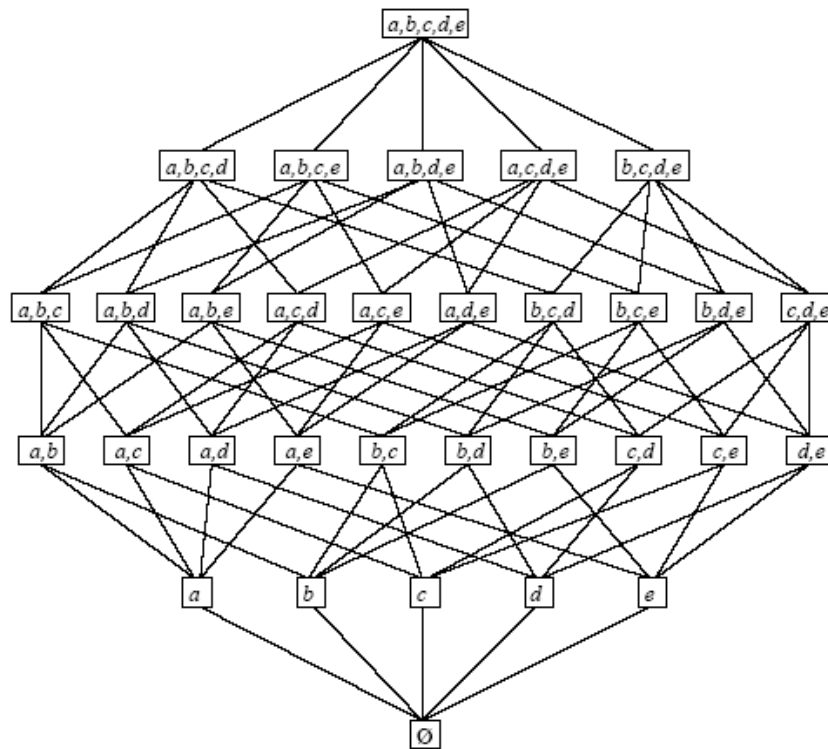


FIGURE C.1 : Treillis des parties associé à I

L'itemset fréquent Φ n'est pas considéré lors de la recherche car les règles d'association générées à partir de ce dernier ne sont pas des règles informatives.

Les différents algorithmes de recherche des itemsets fréquents tentent d'optimiser le nombre de parcours du treillis des itemsets potentiellement fréquents. En pratique, trois approches ont été proposées: Les algorithmes par niveau pour l'extraction d'itemsets fréquents, les algorithmes pour l'extraction d'itemsets fréquents maximaux et les algorithmes pour l'extraction d'itemsets fréquents fermés .

1. Algorithmes par niveaux pour l'extraction d'itemsets fréquents

Ces algorithmes réalisent un balayage du treillis des itemsets, par niveaux, de bas en haut. A chaque itération k , un ensemble de k -itemsets candidats (itemsets fréquents potentiels) est généré à partir des $(k-1)$ -itemsets fréquents découverts lors de l'itération précédente. Plusieurs algorithmes d'extraction des itemsets fréquents, par niveaux, ont été proposés dans la littérature. L'algorithme AIS [Agrawal et al., 93] par exemple crée l'ensemble F_k des k -itemsets fréquents à partir de F_{k-1} en parcourant la base de données transactionnelle D . Pour chaque transaction lue o , et pour chaque itemset $a \in V_{k-1}$, si $a \subseteq o$ alors l'algorithme étend a avec chacun des items de o qui co-occurrent après le dernier item de a étant donné l'ordre lexicographique entre les items de chaque transaction. On génère ainsi un ensemble C_k d'itemsets candidats. Chaque itemset candidat c_k ainsi généré est alors examiné. Si $c_k \in C_k$, alors son

nombre d'occurrences ($count(c_k)$) est incrémenté de 1, sinon c_k est rajouté à C_k avec $count(c_k)=1$. L'ensemble des itemsets fréquents F_k est le sous ensemble de C_k obtenu en ne retenant que les itemsets candidats c_k tels que $count(c_k)$ est supérieur ou égal à un seuil minimal de support $minsup$. L'inconvénient d'une telle approche est la génération de trop nombreux itemsets candidats qui en réalité sont bien moins nombreux. L'algorithme Apriori [Agrawal et al. 94] apporte une solution à un tel problème en offrant une technique de réduction de l'espace de recherche. Apriori est également un algorithme d'extraction des itemsets fréquents, par niveaux. Les k -itemsets candidats sont générés à partir des $(k-1)$ itemsets fréquents extraits lors de la $(k-1)$ ème itération. En pratique, pour limiter le nombre d'itemsets examinés lors de chaque itération, l'algorithme Apriori se base sur 2 propriétés fondamentales :

Propriété 1 : Tous les sur-ensembles d'un itemset non fréquent sont non fréquents. Cette propriété permet d'ignorer, lors de la génération des k -itemsets candidats, les $(k-1)$ -itemsets non fréquents.

Propriété 2 : Tous les sous-ensembles d'un itemset fréquent sont fréquents.

Cette propriété permet de limiter les k -itemsets candidats examinés aux seuls itemsets qui contiennent les $(k-1)$ -itemsets fréquents découverts lors de la précédente itération. Ainsi, les itemsets candidats à k items peuvent être générés par jointure des itemsets fréquents de taille $k-1$, et en supprimant ceux qui contiennent un sous ensemble quelconque non fréquent. Cette procédure implique la génération d'un nombre d'itemsets candidats moins grand que l'algorithme AIS par exemple. L'algorithme AprioriTID [Agrawal et al., 94] étend Apriori en éliminant les parcours multiples de la base de données, à travers la construction d'un ensemble de comptage de base \bar{C}_k , durant la construction de l'ensemble des 1-itemsets fréquents F_1 . L'ensemble \bar{C}_k possède la structure suivante : $\langle TID, \{c_k\} \rangle$ où TID est l'identificateur de la transaction de la base transactionnelle D , et $\{c_k\}$ dénote l'ensemble des k -itemsets candidats contenus dans la transaction identifiée par TID. Le support d'un itemset candidat c_k correspond au nombre d'apparition de ce dernier dans l'ensemble \bar{C}_k . L'ensemble \bar{C}_1 est d'abord généré en transformant chaque item i en l'itemset $\{i\}$. L'ensemble F_1 des 1-itemsets fréquents est déterminé après un parcours de la base D . Puis, à chaque nouvelle itération k ($k \geq 2$), l'ensemble C_k est généré par auto-jointure de F_{k-1} . L'ensemble \bar{C}_k est déterminé à partir de \bar{C}_{k-1} et de C_k tel que chaque élément de \bar{C}_k correspond à un objet o de \bar{C}_{k-1} et contient son TID et la liste des k -itemsets candidats de C_k contenus dans o . L'ensemble F_k est construit en déterminant pour chaque candidat de C_k son nombre d'occurrences dans \bar{C}_k et en ne gardant que les k -

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

itemsets candidats dont le support est supérieur ou égal à un seuil minimal de support *minsup*.

Bien d'autres algorithmes par niveaux existent proposant pour les uns d'autres optimisations de l'algorithme Apriori, tels l'algorithme DHP (*Direct Hashing and Pruning*) proposé par Park et al. [Park et al., 95] qui utilise des tables de hachage afin de diminuer le nombre de candidats générés, la parallélisation du calcul étudiée par Zaki [Zaki et al., 98], ou proposant de nouvelles techniques pour l'extraction des itemsets fréquents tels les algorithmes Partition [Savarese et al., 95] et Sampling [Toivonen et al., 96]. Dans l'algorithme Partition, le contexte est décomposé en partitions qui tiennent en mémoire. Pour chaque partition, tous les itemsets fréquents dans la partition sont déterminés puis fusionnés et un balayage de la totalité du contexte est réalisé pour calculer leurs supports sur l'ensemble du contexte. L'algorithme Sampling utilise les techniques d'échantillonnage pour extraire les itemsets fréquents dans un échantillon du contexte et vérifier leurs supports en réalisant un balayage de l'ensemble du contexte [Pasquier, 00].

2. Algorithmes d'extraction des itemsets fréquents maximaux

L'objectif de ces algorithmes est de réduire l'espace de recherche, et donc le nombre d'itemsets candidats considérés pendant l'extraction. Le principe consiste à extraire les itemsets fréquents maximaux, c'est-à-dire les itemsets fréquents pour lesquels il n'existe pas de sur-ensemble fréquent. L'extraction des itemsets fréquents maximaux est effectuée par un parcours itératif du treillis des itemsets pour :

extraire les itemsets fréquents maximaux dans la base, c'est à dire les itemsets dont le support est supérieur ou égal à *minsup* et dont tous les sur-ensembles sont non fréquents, déterminer les supports de tous les sous-ensembles des itemsets fréquents maximaux en réalisant un balayage de la base,

éliminer les itemsets dont au moins un sous-ensemble est non fréquent.

Plusieurs algorithmes d'extraction des itemsets fréquents maximaux ont été proposés dans la littérature dont les algorithmes *MaxClique* et *Max-Eclat* [Zaki et al., 97], et *Max-Miner* [Bayardo et al., 98]. Ces algorithmes réduisent le nombre d'itérations, et diminuent ainsi le nombre de scans et le nombre d'opérations CPU réalisées en comparaison aux algorithmes à niveaux.

3. Algorithmes pour l'extraction d'itemsets fréquents fermés

Les itemsets fréquents fermés sont définis sur la base de l'opérateur de fermeture de la connexion de Galois [Ganter et al., 99]. Ils forment alors le treillis des itemsets fermés, et des itemsets fermés fréquents. L'ensemble des itemsets fermés fréquents constituent un ensemble générateur, également appelé base, pour l'ensemble des itemsets fréquents. Cela signifie que les itemsets fréquents et leurs supports peuvent être générés à partir des itemsets fermés fréquents et leurs supports sans accéder à la base de données. Le problème de l'extraction de règles d'association consistant alors à

extraire les itemsets fermés fréquents au lieu des itemsets fréquents. Cette décomposition permet d'améliorer les temps de réponse car le nombre d'itemsets fermés fréquents est bien souvent inférieur au nombre d'itemsets fréquents. En utilisant les itemsets fermés fréquents, des bases pour les règles d'association sont aussi définies. Ces bases, qui sont des sous-ensembles de l'ensemble des règles d'association valides, permettent d'améliorer la pertinence et l'utilité de l'ensemble de règles extraites. Les algorithmes Close [Pasquier et al., 98 ; Pasquier et al., 99c] et A-close [Pasquier et al., 99b], réalisent un parcours du treillis des itemsets en largeur d'abord, à la recherche de générateurs (fréquents) des itemsets fréquents fermés par niveaux. Durant l'itération k , l'algorithme Close considère un ensemble de générateurs de candidats de taille k , il détermine leurs supports et leurs fermetures qui constituent les itemsets fermés fréquents, et puis supprime tous les générateurs peu fréquents. Pendant l'itération $(k+1)$, les $(k+1)$ -générateurs de candidats sont construits en joignant deux k -générateurs fréquents si leurs $k-1$ premiers items sont identiques, et les $(k+1)$ -générateurs de candidats obtenus sont éliminés s'ils sont non fréquents ou leur fermeture est déjà calculée. Dans l'algorithme A-Close, les itemsets générateurs sont identifiés selon leurs supports seulement, puisque le support d'un itemset générateur est différent des supports de tous ses sous-ensembles, et un passage supplémentaire de la base de données est ensuite réalisé à la fin de l'algorithme pour calculer les fermetures de tous les générateurs fréquents découverts. Les résultats expérimentaux ont montré que ces algorithmes sont particulièrement efficaces pour l'extraction de règles d'association à partir de données denses ou corrélées qui représentent une partie importante de bases de données réelles. Sur de telles données, close surpasse A-Close, et tous deux surpassent clairement les algorithmes d'extraction des itemsets fréquents par niveaux, tandis que pour les données faiblement corrélées, A-Close surpasse Close [Pasquier, 00a].

C.3.3.2 Mesure de l'intérêt d'une règle d'association

Par delà le grand intérêt des mesures de support et de confiance comme critères d'extraction, on insistera d'abord sur une importante qualité de ces mesures qui est leur grande intelligibilité. Le sens concret des valeurs du support et de la confiance est parfaitement assimilable par l'utilisateur non spécialiste. Toutefois, les algorithmes liés à cette approche engendrent un très grand nombre de règles qui sont difficiles à gérer et dont beaucoup n'ont que peu d'intérêt. La condition de support qui est le moteur même du processus d'extraction écarte les règles ayant un petit support alors que certaines peuvent avoir une très forte confiance et présenter un réel intérêt. Si l'on baisse le seuil de support pour remédier à cet inconvénient, les ensembles fréquents sont trop nombreux et les algorithmes d'extraction sont asphyxiés. Enfin, les seules conditions de support et de confiance ne suffisent pas à assurer le réel intérêt d'une règle. En effet, une règle $A \rightarrow B$

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

dont la confiance est égale à la probabilité de B , soit $P(B/A) = P(B)$ ce qui est la définition de l'indépendance de A et B , n'apporte aucune information. Par exemple, si $P(A) = 80\%$ et $P(B) = 90\%$, la règle $A \rightarrow B$ a un support égal à 72 % et une confiance de 90 % en cas d'indépendance. En résumé, il faut au minimum prendre en compte d'autres mesures d'intérêt des règles que le support et la confiance. Une multitude de mesures d'intérêt sont proposées dans la littérature dont un récapitulatif est donné dans [Lallich et al., 04]. On citera à titre d'exemple :

La mesure de Piatetsky-Shapiro : $nP(A)(P(B/A) - P(B))$ [Piatetsky-Shapiro, 91]

Le lift : $\frac{P(AB)}{P(A)P(B)}$ [Brin et al., 97]

La surprise : $\frac{P(AB) - P(A\bar{B})}{P(B)}$ [Azé et al., 02]

(où n est le nombre total de transactions).

Une mesure doit distinguer les différentes règles associant A et B . En particulier :

1. La mesure doit impérativement permettre de choisir entre $A \rightarrow B$ et $A \rightarrow \bar{B}$, les exemples de l'une étant les contre-exemples de l'autre.
2. On préférera les mesures dissymétriques qui respectent la nature des règles d'association transactionnelles : "si tels articles (A) sont dans le panier, alors le plus souvent tels autres (B) y sont". Les mesures symétriques comme le *support*, la *mesure de Piatetsky-Shapiro* ou le *lift* évaluent de la même façon les règles $A \rightarrow B$ et $B \rightarrow A$, alors que celles-ci ont les mêmes exemples mais pas les mêmes contre-exemples [Lallich et al., 04].

C.3.3.3 Application des règles d'association en RI

L'objectif principal d'un SRI est de rechercher l'information pertinente pour une requête utilisateur à partir d'un ensemble de documents préalablement traités puis stockés dans une base documentaire. Le traitement des documents ou indexation, constitue une étape fondamentale dans tout SRI. De la qualité de l'indexation dépend en effet la qualité des résultats (ou la performance du SRI). L'indexation consiste à construire une représentation intermédiaire du contenu du document. Dans les approches classiques d'indexation, le document est représenté par un ensemble de mots clés. Dans des approches plus évoluées dites d'indexation sémantique, le document est représenté par un ensemble de concepts et de liens entre concepts. Les liens entre concepts sont des relations taxonomiques extraites de thésaurus ou d'ontologies.

L'utilisation des règles d'association en RI vise principalement la découverte de relations non taxonomiques entre les termes (mots clés ou concepts) descripteurs des

documents d'une base documentaire. Les relations non taxonomiques sont des relations contextuelles entre termes. Elles sont spécifiques à l'usage particulier des termes dans les documents du corpus considéré. Il s'agit plus particulièrement de relations latentes, enfouies dans les textes, portées par la sémantique même de la cooccurrence des termes dans le document ou dans la base documentaire. Les objectifs à travers la découverte des règles d'association en RI sont multiples et variés comme en témoigne la multitude d'applications existantes :

1. Le regroupement (clustering) de textes fournit des vues d'ensemble thématiques des collections des textes,
2. La classification de textes en vue de la réduction de l'espace de recherche,
3. La génération automatique d'associations de termes pour l'aide à l'expansion de requête,
4. L'indexation, ...etc.

Dans ce qui suit, nous explicitons quelques travaux en application des règles d'association dans le contexte de la RI.

Application des règles d'association pour la classification des documents

La classification thématique appliquée aux documents, permet de regrouper les textes traitant de la même thématique. Deux textes de documents traitent de thématiques différentes s'ils appartiennent à des classes distinctes. L'objectif est de regrouper autour d'un même thème (i.e. au sein d'une même classe) des documents similaires. Ceci permet de retrouver, pour une requête portant sur un mot clé d'une classe donnée, tous les documents de la classe, mais aussi de les classer ensemble comme documents pertinents.

Le but visé est de pouvoir diminuer efficacement la taille de recherche et d'augmenter la sémantique de classement des documents retournés à l'utilisateur. Dans cette perspective, les auteurs dans [Lin et al., 98], proposent un système (le système ACIRD *_Automatic Classifier for Internet Resource Discovery_*) qui extrait et généralise des termes des documents Internet pour représenter la classification d'une hiérarchie de classes donnée. La mesure de *support* est proposée pour évaluer l'importance d'un terme dans une classe de la hiérarchie de classes. Avec un seuil donné, des termes avec des supports élevés sont filtrés comme mots-clés de la classe, et les termes avec des supports bas sont éliminés. Pour augmenter le taux de rappel de cette approche, la technique d'extraction de règles d'association est appliquée pour découvrir les associations entre termes. Les règles d'association permettent de découvrir les termes exclus mais qui sont cependant représentatifs. Un modèle

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

d'inférence des termes permet alors de les promouvoir au rang de termes représentatifs. Le système, ACIRD, est conçu pour classifier automatiquement les documents collectés par le serveur web Yam¹⁷. Le système a pour but d'améliorer les faibles performances de la classification manuelle actuelle. Le processus de classification de ACIRD se compose de deux phases, une phase d'apprentissage et une phase expérimentale. Dans la première phase, les documents avec leurs classes¹⁸ manuellement assignées dans Yam sont utilisés comme ensemble d'apprentissage pour apprendre la connaissance sur la classification des classes. Puis les documents nouvellement collectés, manuellement classés par catégorie par les experts de Yam, sont appliqués pour vérifier la connaissance de classification apprise en seconde phase. La connaissance sur la classification d'une classe est représentée par un groupe de mots-clés. L'objet correspond à un document Internet (page Web). Le terme est le mot ou l'expression extraite à partir des objets ou généralisée dans des classes par apprentissage. Le support définit le degré d'importance d'un terme qui supporte un certain objet ou classe. La valeur est normalisée sur [0 1]. Pour un seuil minimum de support, les termes sont divisés en termes représentatifs (ou mots-clés), et non représentatifs. Les termes sont appliqués pour découvrir les règles d'association entre termes appartenant aux documents d'une même classe plutôt qu'à tous les documents de la base de données. Les termes correspondent aux items. Les documents dans la classe correspondent aux transactions. La classe correspond à la base de données transactionnelle. La concentration sur des documents d'une classe au lieu de toutes les classes permet de tirer profit de la petite taille de la base de données. Si la taille de la base de données n'est pas très grande, un algorithme d'extraction simple, tel qu'Apriori [Agrawal et al., 94], peut être efficacement appliqué au système. La *confiance* et la *support* sont utilisés respectivement pour promouvoir les termes non représentatifs afin d'affiner la connaissance de classification, et comme seuil pour éliminer des associations de termes bruyantes. Des expérimentations ont été réalisées afin de vérifier que ACIRD peut apprendre et promouvoir des termes représentatifs (mots-clés), qui se rapprochent des concepts des experts humains pour chaque classe. Le rappel et la précision des mots-clés extraits par ACIRD sont comparés aux mots-clés manuellement choisis par les experts. Sur les résultats rapportés, les auteurs concluent que la découverte d'association entre termes de documents est efficace pour l'affinement de la classification.

¹⁷ <http://taiwan.iis.sinica.edu.tw/en/yam/>

¹⁸ (Il existe 12 catégories principales dans la homepage de Yam correspondant chacune à une classe distincte : : . Ce sont : : : : "Arts", "Humanities", "Social Sciences", "Society and Culture", "Natural Sciences", "Computer and Internet", "Health", "News and Information". "Education". "Government and State", "Companies". And "Entertainment and Recreation".

Application des règles d'association pour l'indexation des documents

Dans le domaine de la RI les modèles les plus populaires d'ordonnement des documents d'une collection sont le modèle vectoriel, le modèle probabiliste et le modèle de langue. Les différences entre ces modèles concernent les représentations des documents et requêtes les schémas de pondération et la formule d'évaluation des requêtes (d'ordonnement des documents). La conception de schémas de pondération efficace est une étape critique pour l'amélioration des résultats obtenus. Les meilleurs schémas de pondération sont connus sous le générique de *tf*id*. Ces schémas supposent que les termes sont mutuellement indépendants. Une telle hypothèse est certes erronée. Par ailleurs, il est clairement établi que la prise en compte de relations de co-occurrences entre termes améliore l'efficacité de la recherche dans les SRI. C'est dans cette optique que les auteurs dans [Pôssas et al., 05] proposent un nouveau modèle de RI basé sur les termsets et les règles d'association. La nouveauté concerne deux aspects : D'abord, des modèles de Co-occurrence des termes sont pris en compte lors de l'indexation des documents. Les descripteurs du modèle ne sont plus des termes mais des ensembles de termes d'index (ou termsets), où un termset est un ensemble de termes d'index. Les termsets capturent l'intuition que les termes sémantiquement liés apparaissent près l'un de l'autre dans un document. En second lieu, les poids des termes sont produits en utilisant la technique de découverte des règles d'association. Ceci mène à un nouveau mécanisme d'évaluation appelé le modèle vectoriel basé sur les ensembles. Les résultats expérimentaux montrent que le modèle proposé améliore la précision moyenne pour toutes les collections et types de requêtes évaluées, tout en maintenant des coûts informatiques bas. Pour la collection à 2-gigabyte TREC-8, le modèle a produit un gain en précision moyenne de 14,7% et de 16,4% pour les requêtes disjonctives et conjonctives, respectivement, par rapport au modèle vectoriel standard.

Dans le même contexte d'indexation des documents, une approche pour l'amélioration des poids des termes des documents est proposée dans [Kim et al., 04]. Partant de l'hypothèse que l'utilisation des dépendances de termes est un facteur qui affecte l'exactitude des poids des termes, les auteurs proposent d'utiliser les dépendances pour améliorer la performance du système de recherche. Afin de calculer les dépendances de termes, les auteurs adoptent la méthode des règles d'association proposée dans [Agrawal et al., 93]. La méthode proposée est composée de deux étapes. La première étape consiste à découvrir les dépendances entre termes d'un ensemble de documents en utilisant la découverte des règles d'association. Les documents sont des transactions et les termes des items. En recherchant les associations entre termes, seules les dépendances entre termes individuels sont

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

considérées. Dans la deuxième étape, les poids des termes de tous les documents sont mis à jour par les dépendances entre termes. L'idée est que les termes sont mutuellement affectés par les autres termes dans le document considéré. Un graphe d'association (Figure C.2) des termes est d'abord construit pour chaque document. Les termes du document sont les nœuds du graphe d'associations. Les arcs entre les nœuds du graphe dénotent les associations découvertes entre les termes correspondants. Les arcs sont pondérés par une valeur représentant la confiance de la règle d'association qui lie les termes représentant les deux nœuds de l'arc.

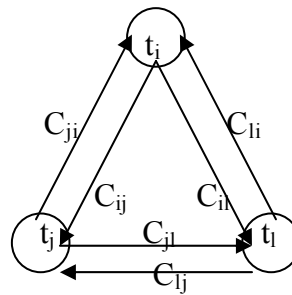


FIGURE C.2 : Graphe d'association

Les poids des termes dans les documents sont alors mis à jour selon la formule suivante:

D : la base documentaire

d_k : le k ième document

t_i : le i ème terme de D

$C_{i,j}$: la valeur de confiance de « $t_i \rightarrow t_j$ »

$w_{i,k}$: le i ème terme de d_k

dl_k : la taille de d_k

$$w'_{i,k} = \frac{\sum_{j/t_j \in d_k} w_{j,k} * c_{j,i}}{dl_k}$$

Chaque poids de terme est transféré à l'autre poids de terme dans une proportion égale à la valeur de la confiance entre les deux termes. En conséquence, W devient le poids du terme influencé par d'autres termes. Ainsi, le poids amélioré *NewWeight* est calculé par combinaison linéaire du poids original du terme considéré et de son poids basé sur la dépendance des termes.

$$newWeight_{t,k} = \alpha * w_{t,k} + \beta * w'_{t,k} \quad (\text{où } \alpha + \beta = 1)$$

Dans leurs expérimentations, les auteurs ont utilisé 224680 documents de TREC et dix topics (65, 66, 68, 82, 83, 96, 102, 111, 134 et 135) de TREC-1 en tant que requêtes. Afin d'évaluer les poids améliorés des termes, les auteurs ont comparé les performances de recherche dans le cas d'utilisation du poids amélioré avec les performances obtenues avec les poids originaux, dans un SRI vectoriel et dans SRI basé sur un modèle de langue, respectivement. Dans le modèle vectoriel et dans le modèle de langue, les poids des termes sont respectivement calculés par le schéma $tf*idf$ et par les probabilités que ces termes se produisent dans un document. Globalement, les résultats obtenus indiquent que l'utilisation des dépendances entre termes rend les poids des termes plus précis.

Application des règles d'association pour l'expansion de requêtes

Un autre aspect de la RI largement concerné par l'utilisation des règles d'association concerne l'expansion de requêtes. La tâche de formulation d'une requête efficace est difficile dans ce sens qu'elle exige de l'utilisateur, n'ayant aucune connaissance sur la collection de documents, de prédire les mots clés qui apparaîtront dans les documents qu'il souhaite avoir.

L'expansion de requête et la réinjection de pertinence ont été proposées en vue de prendre en compte les relations de similarité entre les mots clés. L'expansion de requête est basée sur l'hypothèse qu'un terme d'index est un bon discriminant des documents pertinents et non pertinents, et qu'ainsi tout terme d'index qui lui est symétriquement proche est probablement un aussi bon discriminant. Par ailleurs la réinjection de pertinence est une méthode dans laquelle les requêtes sont étendues en utilisant les mots clés obtenus à partir de l'ensemble des documents résultats si ces derniers sont sémantiquement proches des mots-clés de la requête. Dans [Liu et al., 98], les auteurs proposent d'utiliser les règles d'association afin de découvrir les relations de co-occurrences entre termes. Les règles d'association contrairement à la co-occurrence ne sont pas symétriques. Dans l'approche proposée un document (une requête) est vu comme une liste de mots-clés. Une telle liste de mot-clés a le même rôle qu'une transaction dans les bases de données transactionnelles. Une règle d'association de la forme $X \rightarrow Y$ signifie que le document qui contient tous les mots-clés de X contient aussi tous les mots-clés de Y . Les règles d'association sélectionnées sont celles dont le support et la confiance dépassent respectivement un seuil minimum de support et un seuil minimum de confiance. Lorsqu'une règle d'association $X \rightarrow Y$ est sélectionnée les mots-clés de Y sont rajoutés à la requête si X y apparaît. Une approche similaire est utilisée dans [Haddad, 02] où deux scénarios pour exploiter les associations entre termes dans un SRI sont proposés. (1) L'expansion automatique de la requête à l'aide des règles d'association extraites. Pour chaque terme d'une requête, l'ensemble des associations relatives à ce terme

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

(appelé son profil relationnel) dans le corpus est ajouté dans la requête d'origine. Par exemple, la neuvième requête de INIST¹⁹ contient les termes système de scolarité. Les termes système et scolaire sont associés à d'autres termes avec les règles d'association suivantes découvertes dans la collection INIST :

- système → structure
- système → infrastructure
- scolarité → collègue
- scolarité → lycée

Les termes structure, infrastructure, collègue et lycée sont ajoutés à la requête. La requête enrichie, éloigne des premières réponses le sens système de blanchiment par exemple.

(2) Le second scénario concerne l'expansion interactive des requêtes (*IQE*). Le procédé consiste d'abord à lancer une requête originale. A partir des résultats de la requête, l'utilisateur peut sélectionner des ensembles de termes ou des termes pour les ajouter à sa requête.

Dans [Delgado et al., 02], les auteurs introduisent en plus la notion de généralisation/spécialisation des requêtes. Le principe de reformulation des requêtes est défini comme suit : À partir d'un premier ensemble de documents recherchés pour une requête initiale, la découverte des règles d'association est appliquée afin de retrouver les relations entre les termes de cet ensemble de documents. Les règles les plus précises qui incluent les mots originaux de la requête dans l'antécédent et/ou le conséquent de la règle, sont utilisés pour étendre automatiquement la requête en lui ajoutant ces termes ou, en présentant à l'utilisateur les termes correspondant dans ces règles, afin qu'il puisse choisir les termes à rajouter à la requête originale. Cette suggestion des termes aide l'utilisateur à réduire l'ensemble de documents en dirigeant la recherche à travers la direction désirée. Si un terme de requête apparaît dans l'antécédent d'une règle, et on considère les termes apparaissant dans le conséquent de la règle pour étendre la requête, une généralisation de la requête est effectuée. Par conséquent, une généralisation d'une requête nous donne une requête sur le même sujet (topic) que l'originale, mais recherche des informations plus générales. Cependant, si le terme de la requête apparaît dans le conséquent de la règle, et on reformule la requête en ajoutant les termes apparaissant dans l'antécédent de la règle, alors une spécialisation de la requête sera effectuée, et la précision du système devrait augmenter. La spécialisation d'une requête recherche une information plus spécifique que la requête originale mais dans le même sujet. Afin d'obtenir autant de documents que possible, les termes apparaissant des deux côtés des règles peuvent également être considérés. Une fois la requête étendue, elle est à nouveau soumise au système.

¹⁹ INIST est la collection de l'Institut d'Informatique Scientifique et technique (INIST). Cette collection contient 163308 documents en français dans tous les domaines scientifiques.

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

Les auteurs dans [Bautista et al., 04], proposent d'utiliser les règles d'association floues et un cadre d'évaluation différent des mesures classiques de confiance/support pour l'amélioration des requêtes. Les avantages des règles d'association floues sont :

- les règles floues tiennent compte du degré d'importance des termes dans la représentation des documents.

les mesures utilisées sont plus appropriées pour déterminer quelles règles sont utiles pour cet objectif.

Les auteurs ont par ailleurs montré que la confiance peut donner des résultats fallacieux dans certains cas. Fondamentalement, le problème avec la confiance est qu'elle ne tient pas compte du support de I_2 et par conséquent elle ne peut pas détecter l'indépendance statistique c.-à-d., une valeur élevée de confiance peut être obtenue dans ce cas. Si l'on suppose par exemple un itemset IC , tel que $\text{supp}(IC) = 1$, chaque règle de la forme $IA \rightarrow IC$ sera forte à condition que $\text{supp}(IA) > \text{minsup}$. La mesure de certitude [Shortlidge et al., 75] a été utilisée comme alternative à la confiance. Dans le cadre flou, le document est une transaction représentée par un vecteur de poids $w = \{w_{i1}, \dots, w_{im}\}$. Deux schémas de pondération flous normalisés dans l'intervalle unité sont utilisés. Il s'agit de la pondération par la fréquence et de la pondération par $\text{tf} * \text{idf}$. Une transaction textuelle floue correspond alors à un document d_i dans lequel les termes sont associés aux poids $w = \{w_{i1}, \dots, w_{im}\}$ issus d'un schéma de pondération flou. Avant qu'un enrichissement de la requête puisse être appliqué, un procédé de recherche est préalablement lancé sur une requête originale. Un ensemble des documents pertinents pour cette requête est alors obtenu. Les documents sont ensuite indexés comme en recherche documentaire classique, puis leur représentation sous forme transactionnelle est effectuée. Les transactions sont traitées pour extraire des règles d'association (floues dans ce cas). Une liste de termes de certaines de ces règles est obtenue. Finalement, l'utilisateur choisit dans cette liste, les termes à ajouter à la requête. Le processus d'interrogation du système est ensuite relancé avec la requête étendue. Le choix des termes utiles pour l'amélioration de la requête dépend, comme dans l'approche précédente, de la présence des termes dans l'antécédent et/ou dans le conséquent.

Une approche différente est proposée dans [Song et al., 07]. En effet, dans ce cas, les auteurs proposent une nouvelle technique d'expansion sémantique de requête qui combine des règles d'association avec une ontologie et des techniques de traitement de la langue naturelle. La technique proposée diffère des autres car (1) elle utilise la sémantique explicite aussi bien que d'autres propriétés linguistiques de corpus des textes non structurés, (2) elle utilise les propriétés contextuelles des termes importants, découvertes par les règles d'association, et (3) des entrées d'ontologie sont ajoutées à la question en désambiguïsant les sens des mots.

Pour appliquer l'extraction de règles d'associations à l'expansion de requête, chaque document est vu comme transaction tandis que chaque mot dans le document est vu

ANNEXE C. LA DECOUVERTE DE CONNAISSANCES EN RI

comme item. Un ensemble de mots séparés du document constitue un wordset. Le composant de sélection de caractéristiques (d'indexation) traite les documents en entrée pour sélectionner les termes importants. Les mots non importants tels que des mots fonctionnels et les mots vides sont exclus. Une technique d'extraction des mots-composés est appliquée et les expressions importantes sont extraites. Par ailleurs, un étiquetage de position, par le *Brill Tagger*, est opéré pour éliminer les termes non importants syntaxiquement. L'ensemble des termes et expressions retenus est désambiguïsé en utilisant WordNet. WordNet est accédée pour trouver les entrées appropriées sémantiquement et syntaxiquement. Le processus global de l'expansion de requête basé sur l'utilisation conjointe des règles d'association et de l'ontologie, dite approche *SemanQE*, est défini par les étapes suivantes :

Étape 1: Commenant par un ensemble d'exemples fourni par l'utilisateur, le système retourne un échantillon de documents dans une base de documents indexé, via un moteur de recherche.

Étape 2: Chacun des documents retrouvés est ensuite parsé en phrases puis indexé dans le but d'extraire les phrases et les termes les plus importants dans le document.

Étape 3: Appliquant un algorithme de hybride d'expansion de requête qui combine des règles d'association et des ontologies pour dériver des requêtes capables d'apparier et de retrouver des documents additionnels semblables aux exemples positifs (pertinents).

Étape 4 : Reformulation de la requête sur la base des résultats de l'étape 3 et nouvelle interrogation du moteur de recherche pour retrouver les ensembles de résultats améliorés pertinents pour les requêtes initiales. Les auteurs ont entrepris une série d'expérimentations pour tester l'efficacité de la recherche basée sur l'expansion de requête ainsi proposée, avec des collections de TREC. La technique dite *SemanQE+Ontologie* a été comparée à un certain nombre d'approches dont *Okapi BM25*, et *SemanQE* sans *Ontologie*. Les résultats ont montré que *SemanQE+Ontologie* surpasse les autres techniques de 8,39% à 14,22% en termes de F-mesure. En outre, en termes de $P@20$, la méthode de *SemanQE+Ontologie* est sensiblement meilleure que les autres techniques de 13,41% à 32,39%.

C.4 CONCLUSION

Nous avons présenté dans cette partie les concepts de base de la découverte de connaissances telles qu'introduites initialement dans les bases de données, à travers notamment le concept des règles d'association et des algorithmes de découverte de ces règles. Nous avons discuté ensuite l'application de la découverte des connaissances aux textes (ou text mining). La fouille de textes a été largement utilisée en RI comme l'ont prouvé la multitude de travaux que nous avons présenté. Cette technique a été en effet utilisée, en indexation et en particulier dans la redéfinition des poids d'indexation, et elle s'est avérée particulièrement efficace dans l'expansion de requêtes.

AUTEUR: Fatiha BOUBEKEUR-AMIROUCHE

TITRE: CONTRIBUTION A LA DEFINITION DE MODELES DE RECHERCHE
D'INFORMATION FLEXIBLES BASES SUR LES CP-NETS

DIRECTEUR DE THESE: MOHAND BOUGHANEM / LYNDA TAMINE-LECHANI

LIEU ET DATE DE SOUTENANCE: IRIT. UNIVERSITE TOULOUSE III - PAUL
SABATIER. JUILLET 2008.

RESUME: Ce travail de thèse traite deux principaux problèmes en recherche d'information : la pondération des requêtes et l'indexation sémantique des documents. Notre contribution globale consiste en la définition d'un modèle théorique de RI basé sur les CP-Nets. Le formalisme CP-Net est utilisé d'une part, pour la représentation graphique de requêtes flexibles exprimant des préférences qualitatives, et pour la pondération automatique de telles requêtes. D'autre part, le formalisme CP-Net est utilisé comme langage d'indexation graphique pour représenter les concepts descriptifs d'un document et les relations correspondantes, d'une manière relativement compacte. Les concepts sont identifiés par projection du document sur WordNet. Les relations entre concepts sont découvertes au moyen des règles d'association sémantiques. Un mécanisme d'évaluation des requêtes basé sur l'appariement de graphes CP-Nets est aussi proposé.

TITLE: *CONTRIBUTION TO THE DEFINITION OF FLEXIBLE INFORMATION
RETRIEVAL MODELS BASED ON CP-NETS*

ABSTRACT: *This thesis addresses two main problems in IR: automatic query weighting and document semantic indexing. Our global contribution consists on the definition of a theoretical flexible information retrieval (IR) model based on CP-Nets. The CP-Net formalism is used for the graphical representation of flexible queries expressing qualitative preferences and for automatic weighting of such queries. Furthermore, the CP-Net formalism is used as an indexing language in order to represent document representative concepts and related relations in a roughly compact way. Concepts are identified by projection on WordNet. Concept relations are discovered by means of semantic association rules. A query evaluation mechanism based on CP-Nets graph similarity is also proposed.*

MOTS CLES : *Recherche d'information flexible, pondération des requêtes, indexation sémantique, WordNet, Règles d'association, CP-Nets.*

DISCIPLINE ADMINISTRATIVE: INFORMATIQUE

ADRESSE DU LABORATOIRE: IRIT, Université Paul Sabatier, 118 Route de Narbonne,
F-31062 TOULOUSE CEDEX 9